Reference Manual

ONLINE TRAINING PROGRAMME ON ANALYSIS OF AGRICULTURAL DATA USING STATISTICAL AND DATA MINING TECHNIQUES

11–20 July, 2023

Coordinators

RVSKVV, Gwalior

Dr. S.S. Tomar	Dr. Nisha Singh
Dr. V. B. Singh	Dr. Ankita Sahu
Dr. Shashi Yadav	Dr. Purnima Singh

ICAR-IASRI, New Delhi

Dr. Sudeep Marwaha
Dr. Shashi Dahiya
Dr. Mrinmoy Ray

By

National Agricultural Higher Education Project - Institution Development Plan (NAHEP-IDP), Rajmata Vijayraje Scindia Krishi Vishwa Vidyalaya, Gwalior, Madhya Pradesh

in collaboration with

National Agricultural Higher Education Project - Component 2 (NAHEP - Comp 2) ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Online Training Programme on

Analysis of Agricultural Data Using Statistical and Data Mining Techniques

11 - 20 July,2023

by

National Agricultural Higher Education Project - Institution
Development Plan (NAHEP-IDP), Rajmata Vijayraje Scindia Krishi
Vishwa Vidyalaya, Gwalior, Madhya Pradesh

in collaboration with

National Agricultural Higher Education Project -Component 2 (NAHEP
- Comp 2) ICAR-Indian Agricultural Statistics Research Institute,
New Delhi

Editors

RVSKVV, Gwalior

Dr. S.S. Tomar
Dr. V. B. Singh
Dr. Shashi Yadav
Dr. Nisha Singh
Dr. Ankita Sahu
Dr. Purnima Singh

ICAR-IASRI, New Delhi

Dr. Sudeep Marwaha
Dr. Shashi Dahiya
Dr. Mrinmoy Ray

Disclaimer: The information contained in this reference manual has been taken from various web resources. Respective URL’s are mentioned in the content.

Citation: Singh, V.B., Dahiya, S., Marwaha, S., S., Ray, M., Sahu, A., Singh, P., Singh, N., Yadav, S., Tomar, S.S. (2023). Analysis of Agricultural Data using Statistical And Data Mining Techniques. Reference Manual, RVSKVV, Gwalior, ICAR-IASRI, NewDelhi

CONTENTS

ARTIFICIAL INTELLIGENCE AND KNOWLEDGE MANAGEMENT IN AGRICULTURE

Sudeep Marwaha
ICAR- Indian Agricultural Statistics Research Institute, New Delhi- 110012
sudeep@icar.gov.in

1. Introduction

The Artificial Intelligence is a very old field of study and has a rich history. Modern AI was formalized by John McCarthy, considered as father of AI. It is a branch of computer science, founded around early 1950's. Primarily, the term Artificial Intelligence (or AI) refers to a group of technique that enables a computer or a machine to mimic the behavior of humans in problem solving tasks. Formally, AI is described as "the study of how to make the computers do things at which, at the moment, people are better" (Rich and Knight, 1991; Rich *et al.*, 2009). The main aim of AI is to program the computer for performing certain tasks in humanly manner such as knowledgebase, reasoning, learning, planning, problem solving etc. The Machine Learning (ML) techniques are the subset of AI which makes the computers/machines/programs the capable of learning and performing tasks without being explicitly programmed. The ML techniques are not just the way of mimicking human behaviour but the way of mimicking how humans learn things. The main characteristics of machine learning is 'learning from experience' for solving any kind of problem. The methods of learning can be categorized into three types: (a) supervised learning algorithm is given with labelled data and the desired output whereas (b) unsupervised learning algorithm is given with unlabelled data and identifies the patterns from the input data and (c) reinforcement learning algorithm allows the ML techniques to capture the learnable things on the basis of rewards or reinforcement. Now, the Deep Learning (DL) technique are the advanced version of machine learning algorithms gained much popularity in the area of image recognition and computer vision. The artificial neural networks (ANNs) clubbed with representation learning are the backbone of the deep learning concepts. These techniques allow a machine to learn patterns in the dataset with multiple levels of abstractions. The DL models are composed of a series of non-linear layers where each of the layer has the capability of transforming the low-level representations into higher-level representations i.e. into a more abstract representations (Le Cun *et al.*, 2015). There are several DL algorithms available now-a-days such as Deep

Convolutional Neural Networks, Deep Recurrent Neural networks, Long Short-term Memory (LSTM) networks that are being applied to different areas of engineering, bioinformatics, agriculture, medical science and many more (Fusco *et al.,* 2021).

2. Applications of Artificial Intelligence in Agriculture:

In present scenario, AI techniques are being exponentially applied in the various areas of the agricultural domain. These areas can be categorized into the following groups: Soil and water management, Crop Health Management, Crop Phenotyping, Recommender-based systems for crops, Semantic web and Ontology driven expert systems for crops and Geo-AI. The application of AI, ML and DL based techniques on these areas are discussed in the following sections.

2.1 Soil and Irrigation Management:

Soil and irrigation are the most viable components of agriculture. The soil and irrigation are the determinant factors for the optimum crop yield. In order to obtain enhanced crop yield and to maintain the soil properties, there is a requirement of appropriate knowledge about the soil resources. The management of irrigation becomes crucial when there are scares of water availability. Therefore, the soil and irrigation related issues should be managed properly and cautiously to ensure a potential yield in crops. In this regards, AI and ML based techniques have shown potential ability to resolve soil and irrigation related issues in crops. A range of machine learning models such as linear regression, support vector machines (or regressors), Artificial neural networks, random forest algorithm and so on are being used. Many researchers have used remote-sensed data with the machine learning techniques for determining soil health parameters. In this section, few significant works in this field are highlighted below:

A. Soil Management:

Besalatpour *et al.,* (2011), Aitkenhead *et al.,* (2012) and Sirsat *et al.,* (2017) used different machine learning techniques such as linear regression, support vector machine, random forests for the prediction of the physical and chemical properties of soil. Rivera *et al.* (2020) and Azizi *et al.,* (2020) worked on estimation and classification of aggregate stability of the soils using conventional machine learning techniques as well as deep learning models. Jha *et al.,* (2018) worked on prediction of microbial dynamics in soils using regression-based techniques. Patil and Dekha (2016) and Mehdizadeh *et Al.* (2017) worked on predicting the evapotranspiration rate in crops using several machine learning techniques. Researchers worked on mapping the soil properties digitally using the remote sensing data with the help of machine learning and deep learning models (Taghizadeh-Mehrjardi *et al.* 2016; Kalambukattu *et al.,* (2018; Padarian *et al.* 2019; Taghizadeh-Mehrjardi *et al.,* 2020).

B. Irrigation management:

Zema *et al.* 2018 applied Data Envelopment Analysis (DEA) with Multiple Regression analysis to improve the irrigation performance Water Users Associations. Ramya *et al.* 2020 and Glória *et al.,* 2021worked on IoT based smart irrigation systems with machine learning models. Agastya *et al*, 2021 and Zhang *et al.* 2018 used deep learning-based CNN models for detection of irrigations using remote sensing data. Jimenez *et al.* 2021 worked on estimating the irrigation based on soil matric potential.

2.2 Crop Health Management:

Every year a significant amount of yield is damaged due to attack of disease causing pathogens and insect-pest infestation. In order to manage the spread of the diseases and insect-pests, proper management practices should be applied at the earliest. Therefore, there is requirement of automatic diseases, pest identification system. In this regard, image-based diagnosis of diseases and pests have become de facto standard of automatic stress identification. This kind of automated detection methodology use sophisticated deep learning-based AI techniques that reduces the intervention of the human experts. There are several attempts have been done to diagnose the diseases as well as insects-pests in crops using deep learning techniques. In this section, some of the significant works in this field have been discussed briefly.

A. Disease identification:

Mohanty *et al.* 2016 worked on disease diagnosis problem using deep CNN models. They used an open-source dataset named PlantVillage (Hughes and Salathe, 2016) containing 54,306 digital images of 26 diseases from 14 crops. Ferentinos, 2018 worked on developing deep CNN-based models for recognising 56 diseases from different crops. Barbedo, 2019 applied transfer learning approach for diagnosis of diseases of 12 different crops. Too et al. 2019, applied pre-trained deep CNN models for identification of diseases of 18 crops using the PlantVillage data. Chen *et al.* 2020 applied a pretrained VGGNet network for classifying the diseases of Rice and Maize crop. Chen *et. al.* 2020 and Rahman *et al.* 2020 worked on identifying the major diseases of Rice crop. Lu *et al.,* 2017; Johannes *et al.* 2017; Picon *et al.* 2019 and Nigam *et al.* 2021 applied deep CNN models for recognising the diseases of wheat crop. Priyadharshini *et al.* 2019; Sibiya & Sumbwanyambe, 2019; Haque *et al.* 2021 used deep learning models for identifying diseases of maize crop.

B. Pest Identification:

Pest Identification problem is inherently different from disease detection. As compared to disease detection there are less number of work has found in the literature. Some of the research of pest identification has been discussed in the following section.

Cheeti *et al.* (2021) developed a model for pest detection and classification of peat using YOLO (You look only once) and CNN. YOLO algorithm is used for detection of pest in an image and Alex net CNN is used for pest classification. Chen et al. (2021) propose an AI-based pest detection system for solving the specific issue of detection of scale pests based on pictures. Deep-learning-based object detection models, such as faster region-based convolutional networks (Faster R-CNNs), single-shot multibox detectors (SSDs), and You Only Look Once v4 (YOLO v4), are employed to detect and localize scale pests in the picture. Taiwan Agricultural Research Institute, Council of Agriculture, has collected images of the three types of pests from the actual fields for decades. Fuentes *et al.* (2017) address disease and pest identification by introducing the application of deep meta-architectures and feature extractors. They proposed a robust deep-learning-based detector for real-time tomato diseases and pests recognition. The system introduces a practical and applicable solution for detecting the class and location of diseases in tomato plants, which in fact represents a main comparable difference with traditional methods for plant diseases classification. Karnik *et al.* (2021) image pre-processing and data augmentation techniques has been performed to get better image.yolov3 classification for classifying plant leaf disease of pepper bell, potato and tomato. This proposed in divided into two stage part first classifier and second stage classifier where in first classifier it will preprocess of median filter and data augmentation is used and trained in yolov3 algorithm and in second stage classifier it will perform the extract plant leaf image output using Resnet50 based. So, it two step classification approach. Based on this research work we achieved 94% accuracy of detection lead diseases. Experiments showed [Li et al. (2020)] that our system with the custom backbone was more suitable for detection of the untrained rice videos than VGG16, ResNet-50, ResNet-101 backbone system and YOLOv3 with our experimental environment. Liu et al.2020 used Yolo V3 model is a little inadequate in the scale when recognizing tomato disease spots and pests.

2.3 Plant Phenomics:

Non-destructive phenotypic measurement with high throughput imaging technique becoming extremely popular. High throughput imaging system produces a large number of images. Deduction of the phenotypic characteristics through image analysis is quick and accurate. A wide range of phenotypic study can be done using phenomics analysis. High throughput imaging system coupled with sophisticated AI technology like deep learning make this field more efficient and accurate. Phenomics is has been used for study of several phenotypic characters like spike detection and counting, yield forecasting, quantification of the senescence in the plant, leaf weight and count, plant volume, convex hull, water stress and many more.

2.4 Recommender Systems:

Recommender systems (RSs) help online users in decision making regarding products among a pile of alternatives. In general, these systems are software solutions which predict liking of a user for unseen items. RSs have been mainly designed to help users in decision making for areas where one is lacking enough personal experience to evaluate the overwhelming number of alternative items that a website has to offer [Resnick & Varian, 1997]. Recommender systems have proved its worth in many different applications like e-commerce, e-library, e-tourism, e-learning, e-business, e-resource services etc. by suggesting suitable products to users [Lu *et al.*, 2015]. RSs are used to introduce new/unseen items to users, to increase user satisfaction etc. Recommendations are generated by processing large amount of historical data on the users and the products to be suggested. Most popular way of gathering users liking on a particular product is in terms of rating either in numerical scale (1 to 5) or ordinal scale (strongly agree, agree, neutral, disagree, strongly disagree). Other techniques of more knowledge – based recommendation are the use of Ontologies [Middleton *et al.*, 2002] of user profiles or item descriptions etc. The core task of a recommendation system is to predict the usefulness of an item to an individual user based on the earlier history of that item or by evaluating the earlier choices of the user. Collaborative way of user modelling [Konstan *et al.*, 1997] is where ratings are predicted for <user,

item> pair, $\overline{R}$<u, i> based on a large number of ratings previously gathered by the system on individual <user, item> pairs. Another way of recommendation is to suggest items that are similar to the ones previously liked by the user, called Content based filtering [Wang *et al.,* 2018; Smyth, 2007]. In a hybrid method of prediction, limitations by the earlier mentioned processes are tackled in various ways.

Agriculture has used recommender systems since 2015 and continues to do so. RSs have been explored to develop crop recommendation strategies based on soil and weather parameters, crop rotation practices, water management, suggestion on suitable varieties, recommendations for management practices etc. It is absolutely essential for the farmers to receive recommendations on the best crop for cultivation. Kamatchi and parvati, 2019 proposed a hybrid RS in combination with Collaborative Filtering, Case-based Reasoning and Artificial Neural Networks (ANN) to predict future climatic conditions and recommendation of crops based on the predicted climate. Crop recommendations have been developed based on season and productivity [Vaishnavi *et al.,* 2021], area and soil type [Pande *et al.,* 2021] by using several machine learning algorithms like Support vector Machine (SVM), Random forest (RF), Multivariate Linear regression (MLR), K- Nearest neighbour (KNN), ANN etc. Ensemble techniques have been used to develop a collaborative system of crop rotation, crop yield prediction, forecasting and fertilizer recommendation [Archana *et al.,* 2020]; to classify soil types into recommended crop types Kharif or Rabi based on specific physical and chemical characteristics, average rainfall and surface temperature [Kulkarni *et al.,* 2018]. Naha and Marwaha, 2020 presented an Ontology driven context aware RS that can recommend land preparation methods, sowing time, seed rate, fertilizer management, irrigation scheduling and harvesting methods to Maize cultivators. Application of RSs has also penetrated in the e-agriculture domain by suggesting parts of agricultural machineries in online ordering [Ballesteros *et al.,* 2021].

2.5 Semantic web, Knowledgebase and Natural Language processing:

Agriculture is vast source of resources and so it is also a vast source of information. The problem with this information is most of the information are unstructured. That unstructured knowledge is merely understandable for machine. It is also has low accessibility for human too. The main objectives of the semantic web and knowledge base system are to make unstructured data into structured one. Semantic web and the knowledgebase mainly facilitated by the ontology in the back end. Ontology is a formal, explicit specification of a shared conceptualization (Gruber, 1993). Making of Ontology that facilitated the semantic web and knowledge base can be made across the agricultural domain to make the unstructured data into structured one. Many ontology has already been developed in accordance with the Bedi and Marwaha, 2004 in the agricultural domain. Saha *et. al.,* (2011) developed an ontology on dynamic maize variety selection in different climatic condition, Sahiram *et. al.,* (2012) developed a ontology on rapeseed and mustard for identification of the variety in multiple languages, Das *et. al.,* (2011) developed a ontology for USDA soil taxonomy and ontology was extended by Deb *et. al*., (2012), Biswas *et. al*., (2012) developed a ontology on microbial taxonomy and was extended by Karn *et. al.* (2014).

2.6 GIS and Remote sensing coupled with AI:

GIS and Remote sensing is helping agricultural community since long. The land use planning, land cover analysis, forest distribution, water distribution, water use pattern, crop rotation and crop calendar analysis can be done by GIS and remote sensing. But when the AI and machine learning coupled with these technology it become more powerful. Machine learning and AI efficiently used for correct land classification and phonological change detection. From Digital soil mapping to yield forecasting, from phenology detection to leaf area index a vast range of the area in agriculture can be handled by GIS and Remote sensing.

References:

Agastya, C., Ghebremusse, S., Anderson, I., Vahabi, H., &Todeschini, A. (2021). Self-supervised Contrastive Learning for Irrigation Detection in Satellite Imagery. arXiv preprint arXiv:2108.05484.

AhilaPriyadharshini, R., Arivazhagan, S., Arun, M., &Mirnalini, A. (2019). Maize leaf disease classification using deep convolutional neural networks. Neural Computing and Applications, 31(12), 8887-8895.

Aitkenhead, M. J., Coull, M. C., Towers, W., Hudson, G., & Black, H. I. J. (2012). Predicting soil chemical composition and other soil parameters from field observations using a neural network. Computers and Electronics in Agriculture, 82, 108-116.

Archana, K., &Saranya, K. G. (2020). Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier. *Seventh Sense Research Group (SSRG) International Journal of Computer Science and Engineering*, 7(5), 1-4.

Azizi, A., Gilandeh, Y. A., Mesri-Gundoshmian, T., Saleh-Bigdeli, A. A., &Moghaddam, H. A. (2020). Classification of soil aggregates: A novel approach based on deep learning. Soil and Tillage Research, 199, 104586.

Ballesteros J. M., Cartujano, A. R., Evaldez, D., Macutay, J., (2021). Online ordering and recommender system of combine harvester parts and equipment with 3D modelling and augmented reality brochure for BLAZE equifarm and general merchandise. *11th International Workshop on Computer Science and Engineering (WCSE 2021)*, 174-179.

Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. Biosystems Engineering, 180, 96-107.

Besalatpour, A., Hajabbasi, M. A., Ayoubi, S., Gharipour, A., &Jazi, A. Y. (2012). Prediction of soil physical properties by optimized support vector machines. International Agrophysics, 26(2).

Biswas, S., Marwaha, S., Malhotra, P. K., Wahi, S. D., Dhar, D. W., & Singh, R. (2013). Building and querying microbial ontology. *Procedia Technology, 10*, 13-19.

Cheeti, S., Kumar, G. S., Priyanka, J. S., Firdous, G., &Ranjeeva, P. R. (2021). Pest Detection and Classification Using YOLO AND CNN. Annals of the Romanian Society for Cell Biology, 15295-15300.

Chen, J. W., Lin, W. J., Cheng, H. J., Hung, C. L., Lin, C. Y., & Chen, S. P. (2021). A smartphone-based application for scale pest detection using multiple-object detection methods. Electronics, 10(4), 372.

Chen, J., Zhang, D., Nanehkaran, Y. A., & Li, D. (2020). Detection of rice plant diseases based on deep transfer learning. Journal of the Science of Food and Agriculture, 100(7), 3246-3256.

Das, B. *et al.* (2017a) "Comparison of different uni-and multi-variate techniques for monitoring leaf water status as an indicator of water-deficit stress in wheat through spectroscopy," *Biosystems Engineering*, 160, pp. 69–83.

Deb, C. K., Marwaha, S., Malhotra, P. K., Wahi, S. D., &Pandey, R. N. (2015, March). Strengthening soil taxonomy ontology software for description and classification of USDA soil taxonomy up to soil series. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1180-1184). IEEE.

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and electronics in agriculture, 145, 311-318.

Fuentes, A., Yoon, S., Kim, S. C., & Park, D. S. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors, 17(9), 2022.

Fusco, R., Grassi, R., Granata, V., Setola, S. V., Grassi, F., Cozzi, D., ...&Petrillo, A. (2021). Artificial Intelligence and COVID-19 Using Chest CT Scan and Chest X-ray Images: Machine Learning and Deep Learning Approaches for Diagnosis and Treatment. *Journal of Personalized Medicine*, **11(10)**, 993.

Glória, A.; Cardoso, J.; Sebastião, P. Sustainable irrigation system for farming supported by machine learning and real-time sensor data. Sensors 2021, 21, 3079.

Janik, L. J., Forrester, S. T., & Rawson, A. (2009). The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. Chemometrics and Intelligent Laboratory Systems, 97(2), 179-188.

Jimenez, A.F.; Ortiz, B.V.; Bondesan, L.; Morata, G.; Damianidis, D. Long short-term memory neural network for irrigation management: A case study from southern Alabama, USA. Precis. Agric. 2021, 22, 475–492

Johannes, A., Picon, A., Alvarez-Gila, A., Echazarra, J., Rodriguez-Vaamonde, S., Navajas, A. D., & Ortiz-Barredo, A. (2017). Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. Computers and electronics in agriculture, 138, 200-209.

Kalambukattu, J. G., Kumar, S., & Raj, R. A. (2018). Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. Environmental earth sciences, 77(5), 1-14.

Kamatchi, S., B. &Parvathi, R. (2019). Improvement of crop production using recommender system by weather forecasts. *Procedia Computer Science*, 165, 724–732.

Karnik, J., &Suthar, A. (2021). Agricultural Plant Leaf Disease Detection Using Deep Learning Techniques. Available at SSRN 3917556.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3), 77-87.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, **521(7553)**, 436-444.

Li, D., Wang, R., Xie, C., Liu, L., Zhang, J., Li, R., ...& Liu, W. (2020). A recognition method for rice plant diseases and pests video detection based on deep convolutional neural network. Sensors, 20(3), 578.

Liu, J., & Wang, X. (2020). Tomato diseases and pests detection based on improved Yolo V3 convolutional neural network. Frontiers in plant science, 11, 898.

Lu, J., Hu, J., Zhao, G., Mei, F., & Zhang, C. (2017). An in-field automatic wheat disease diagnosis system. Computers and electronics in agriculture, 142, 369-379.

Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G., (2015). Recommender system application developments: a survey. *Decision Support Systems*, 74, 12- 32.

Manoranjan, D., Malhotra, P. K., Sudeep, M., &Pandey, R. N. (2012). Building and querying soil ontology. *Journal of the Indian society of agricultural statistics*, *66*(3), 459-464.

Mohanty, S. P., Hughes, D. P., &Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in plant science, 7, 1419.

Naha, S. and Marwaha, S. (2020). Context-Aware Recommender System for Maize Cultivation. *Journal of Community Mobilization and Sustainable Development*, 15(2), 485-490.

Nigam, S., Jain, R., Marwaha, S., & Arora, A. (2021). Wheat rust disease identification using deep learning. De Gruyter.

Padarian, J., Minasny, B., &McBratney, A. B. (2019). Using deep learning for digital soil mapping. Soil, 5(1), 79-89.

Pande, S. M., Ramesh, P. K., Anmol, A., Aishwarya, B. R., Rohilla, K., &Shaurya, K. (2021). Crop recommender system using machine learning approach. *In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1066-1071. IEEE.

Patil, A. P., &Deka, P. C. (2016). An extreme learning machine approach for modelling evapotranspiration using extrinsic inputs. Computers and electronics in agriculture, 121, 385-392.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3), 56- 58.

Rich, E., & Knight, K. (1991). *Artificial Intelligence. 2nd Edn*. New York, NY, United States: McGraw-Hill.

Rich, E., Knight, K., and Nair, S. B. (2009). *Artificial Intelligence*. *3rd Edn*. New Delhi, India: Tata McGraw-Hill.

Rivera, J. I., & Bonilla, C. A. (2020). Predicting soil aggregate stability using readily available soil properties and machine learning techniques. Catena, 187, 104408.

Sibiya, M., &Sumbwanyambe, M. (2019). A computational procedure for the recognition and classification of maize leaf diseases out of healthy leaves using convolutional neural networks. AgriEngineering, 1(1), 119-131.

Sirsat, M. S., Cernadas, E., Fernández-Delgado, M., & Khan, R. (2017). Classification of agricultural soil parameters in India. Computers and electronics in agriculture, 135, 269-279.

Smyth, B. (2007). Case-based recommendation. In The adaptive web. Springer, Berlin, Heidelberg, 342-376.

Taghizadeh-Mehrjardi, R., Ayoubi, S., Namazi, Z., Malone, B. P., Zolfaghari, A. A., &Sadrabadi, F. R. (2016). Prediction of soil surface salinity in arid region of central Iran using auxiliary variables and genetic programming. Arid Land Research and Management, 30(1), 49-64.

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 161, 272-279.

Vaishnavi, S., Shobana, M., Sabitha, R., &Karthik, S. (2021). Agricultural Crop Recommendations based on Productivity and Season. *In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS).* 1, 883-886. IEEE.

Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for computer science publications. *Knowledge-Based Systems*, 157, 1-9.

Zema, D.A.; Nicotra, A.; Mateos, L.; Zimbone, S.M. Improvement of the irrigation performance in water users associations integrating data envelopment analysis and multi-regression models. Agric. Water Manag. 2018, 205, 38–49.

INTRODUCTION TO R SOFTWARE

Soumen Pal, B. N. Mandal
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
E-mail: Soumen.Pal@icar.gov.in

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.R is a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages.

R environment

The R environment provides an integrated suite of software facilities for data manipulation, calculation and graphical display. It has

- a data handling and storage facility,
- a suite of operators for calculations on arrays and matrices,
- a large, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display, and
- a well-developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined functions and input and output facilities.

Origin

R can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus systems. Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland started the project on R in 1995 and hence the name software has been named as 'R'.

R was introduced as an environment within which many classical and modern statistical techniques can be implemented. A few of these are built into the base R environment, but many are supplied as packages. There are a number of packages supplied with R (called "standard" and "recommended" packages) and many more are available through the CRAN family of Internet sites (via http://cran.r-project.org) and elsewhere.

Availability

Since R is an open source project, it can be obtained freely from the website www.r-project.org. One can download R from any CRAN mirror out of several CRAN

(Comprehensive R Archive Network) mirrors. Latest available version of R is R version 4.3.1 and it has been released on June 16 2023.

Installation

To install R in windows operating system, simply double click on the setup file. It will automatically install the software in the system.

Usage

R can work under Windows, UNIX and Mac OS. In this note, we consider usage of R in Windows set up only.

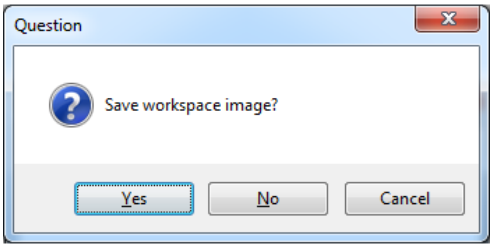
Difference with other packages

There is an important difference between R and the other statistical packages. In R, a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give large amount of output from a given analysis, R will give minimal output and store the results in an object for subsequent interrogation by further R functions.

Invoking R

If properly installed, usually R has a shortcut icon on the desktop screen and/or you can find it under Start| All Programs| R menu.

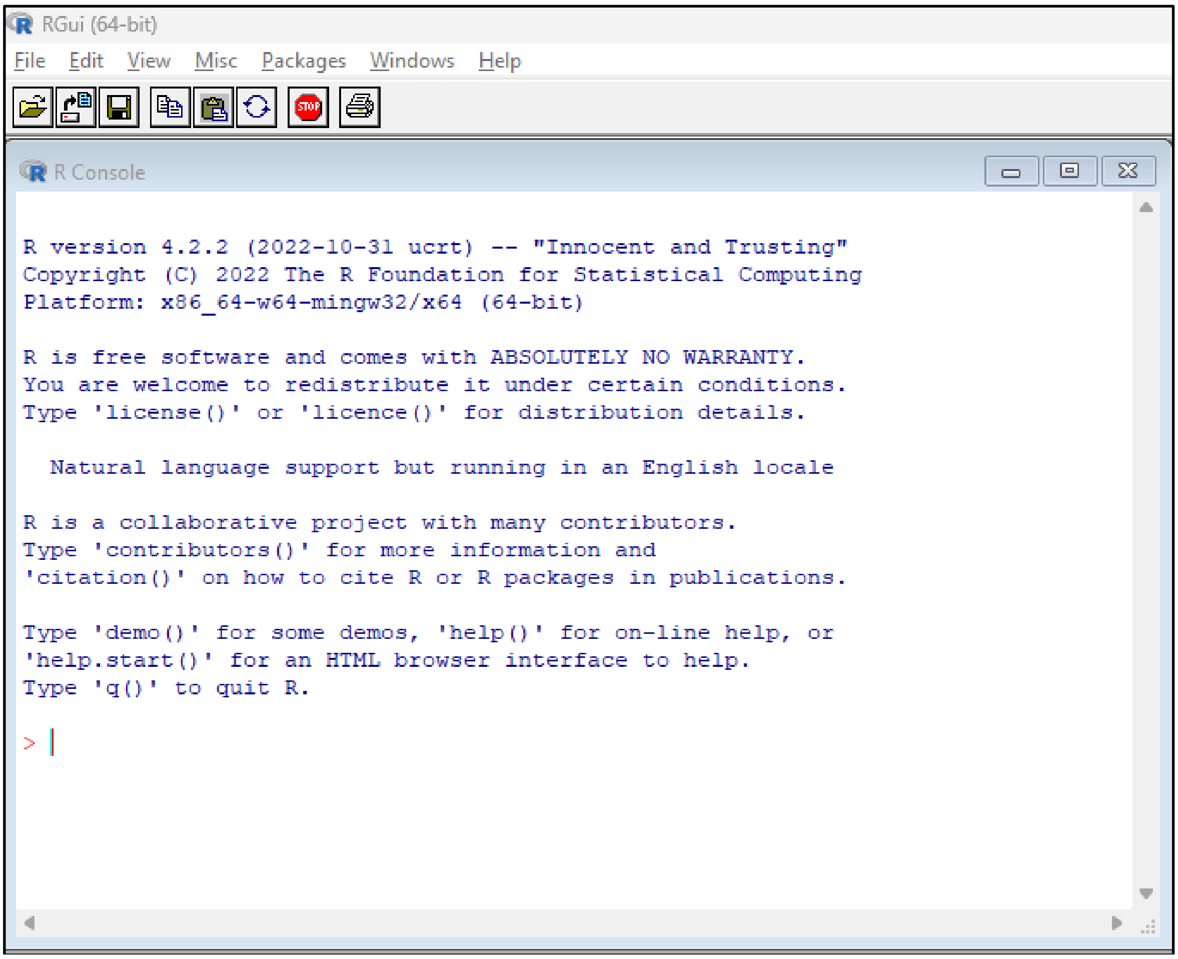
To quit R, type q() at the R prompt (>) and press Enter key. A dialog box will ask whether to save the objects you have created during the session so that they will become available next time when R will be invoked.



Save workspace image?

Windows of R

R has only one window and when R is started it looks like



```
R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

R commands

i. R commands are case sensitive, so X and x are different symbols and would refer to different variables.

ii. Elementary commands consist of either expressions or assignments.

iii. If an expression is given as a command, it is evaluated, printed and the value is lost.

iv. An assignment also evaluates an expression and passes the value to a variable but the result is not automatically printed.

v. Commands are separated either by a semi-colon (';'), or by a newline.

vi. Elementary commands can be grouped together into one compound expression by braces '{' and '}'.

vii. Comments can be put almost anywhere, starting with a hashmark ('#'). Anything written after # marks to the end of the line is considered as a comment.

viii. Window can be cleared of lines by pressing Ctrl + L keys.

Executing commands from or diverting output to a file

If commands are stored in an external file, say ‘D:/commands.txt’ they may be executed at any time in an R session with the command

```
>source("d:/commands.txt")
```

For Windows Source is also available on the File menu.

The function *sink()*,

```
>sink("d:/record.txt")
```

will divert all subsequent output from the console to an external file, ‘record.txt’ in D drive. The command

```
>sink()
```

restores it to the console once again.

Simple manipulations of numbers and vectors

R operates on named data structures. The simplest such structure is the numeric vector, which is a single entity consisting of an ordered collection of numbers. To set up a vector named x, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

The function *c()* assigns the five numbers to the vector x. The assignment operator (<-) ‘points’ to the object receiving the value of the expression. Once can use the ‘=’ operator as an alternative.

A single number is taken as a vector of length one.

Assignments can also be made in the other direction, using the obvious change in the assignment operator. So the same assignment could be made using

```
>c(10.4, 5.6, 3.1, 6.4, 21.7) -> x
```

If an expression is used as a complete command, the value is printed. So now if we were to use the command

```
> 1/x
```

the reciprocals of the five values would be printed at the terminal.

The elementary arithmetic operators

+ addition

– subtraction

* multiplication

/ division

^ exponentiation

Arithmetic functions

log, exp, sin, cos, tan, sqrt,

Other basic functions

max(x) – maximum element of vector x,

min(x)- minimum element of vector x,

range (x) – range of the values of vector x ,

length(x) - the number of elements in x,

sum(x) - the total of the elements in x,

prod(x) – product of the elements in x

mean(x) – average of the elements of x

var(x) – sample variance of the elements of (x)

sort(x) – returns a vector with elements sorted in increasing order.

Logical operators

< - less than

<= less than or equal to

>greater than

>= greater than or equal to

== equal to

!= not equal to.

Other objects in R

Matrices or arrays - multi-dimensional generalizations of vectors.

Lists - a general form of vector in which the various elements need not be of the same type, and are often themselves vectors or lists.

Functions - objects in R which can be stored in the project’s workspace. This provides a simple and convenient way to extend R.

Matrix facilities

A matrix is just an array with two subscripts. R provides many operators and functions those are available only for matrices. Some of the important R functions for matrices are

t(A) – transpose of the matrix A

nrow(A) – number of rows in the matrix A ncol(A) – number of columns in the matrix A

A%*% B– Cross product of two matrices A and B

A*B – element by element product of two matrices A and B

diag (A) – gives a vector of diagonal elements of the square matrix A

diag(a) – gives a matrix with diagonal elements as the elements of vector a

eigen(A) – gives eigen values and eigen vectors of a symmetric matrix A

rbind (A,B) – concatenates two matrix A and B by appending B matrix below A matrix (They should have same number of columns)

cbind(A, B) - concatenates two matrix A and B by appending B matrix in the right of A matrix (They should have same number of rows)

Data frame

Data frame is an array consisting of columns of various mode (numeric, character, etc). Small to moderate size data frame can be constructed by *data.frame()* function. For example, following is an illustration how to construct a data frame from the car data*:

Make	**Model**	**Cylinder**	**Weight**	**Mileage**	**Type**
Honda	Civic	V4	2170	33	Sporty
Chevrolet	Beretta	V4	2655	26	Compact
Ford	Escort	V4	2345	33	Small
Eagle	Summit	V4	2560	33	Small
Volkswagen	Jetta	V4	2330	26	Small
Buick	Le Sabre	V6	3325	23	Large
Mitsubishi	Galant	V4	2745	25	Compact
Dodge	Grand Caravan	V6	3735	18	Van
Chrysler	New Yorker	V6	3450	22	Medium
Acura	Legend	V6	3265	20	Medium

```
> Make<-
c("Honda","Chevrolet","Ford","Eagle","Volkswagen","Buick"
,"Mitsbusihi",
+ "Dodge","Chrysler","Acura")
>
Model=c("Civic","Beretta","Escort","Summit","Jetta","LeSa
bre","Galant",
+ "Grand Caravan","NewYorker","Legend")
```

Note that the plus sign (+) in the above commands are automatically inserted when the carriage return is pressed without completing the list. Save some typing by using *rep()* command. For example, *rep("V4",5)* instructs R to repeat V4 five times.

```
> Cylinder<-c(rep("V4",5),"V6","V4",rep("V6",3))
> Cylinder
 [1] "V4" "V4" "V4" "V4" "V4" "V6" "V4" "V6" "V6" "V6"
> Weight<-
c(2170,2655,2345,2560,2330,3325,2745,3735,3450,3265)
> Mileage<-c(33,26,33,33,26,23,25,18,22,20)
> Type<-
c("Sporty","Compact",rep("Small",3),"Large","Compact","Va
n",rep("Medium",2))
```

Now *data.frame()* function combines the six vectors into a single data frame.

```
> Car<-
data.frame(Make,Model,Cylinder,Weight,Mileage,Type)
> Car
         Make          Model Cylinder Weight
Mileage    Type
1       Honda          Civic        V4   2170    33  Sporty
2   Chevrolet        Beretta        V4   2655    26 Compact
3        Ford         Escort        V4   2345    33   Small
4       Eagle         Summit        V4   2560    33   Small
5  Volkswagen          Jetta        V4   2330    26   Small
6       Buick       LeSabre         V6   3325    23   Large
7  Mitsbusihi         Galant        V4   2745    25 Compact
8  Dodge Grand       Caravan        V6   3735    18     Van
9    Chrysler     New Yorker        V6   3450    22  Medium
10      Acura         Legend        V6   3265    20  Medium
> names(Car)
[1] "Make"     "Model"    "Cylinder"
"Weight"   "Mileage"  "Type"
```

Just as in matrix objects, partial information can be easily extracted from the data frame:

```
>Car[1,]
   Make Model Cylinder Weight Mileage   Type
1 Honda Civic       V4   2170      33 Sporty
```

In addition, individual columns can be referenced by their labels:

```
>Car$Mileage
 [1] 33 26 33 33 26 23 25 18 22 20
>Car[,5]        #equivalent expression
```

```
> mean(Car$Mileage)    #average mileage of the 10
vehicles
[1] 25.9
> min(Car$Weight)
[1] 2170
```

table() command gives a frequency table:

```
>table(Car$Type)
Compact   Large  Medium   Small  Sporty     Van
      2       1       2       3       1       1
```

If the proportion is desired, type the following command instead:

```
>table(Car$Type)/10
Compact   Large  Medium   Small  Sporty     Van
    0.2     0.1     0.2     0.3     0.1     0.1
```

Note that the values were divided by 10 because there are that many vehicles in total.

If you don't want to count them each time, the following does the trick:

```
>table(Car$Type)/length(Car$Type)
```

Cross tabulation is very easy, too:

```
>table(Car$Make, Car$Type)

             Compact Large Medium Small Sporty Van
  Acura      0       0     1      0     0      0
  Buick      0       1     0      0     0      0
  Chevrolet  1       0     0      0     0      0
  Chrysler   0       0     1      0     0      0
  Dodge      0       0     0      0     0      1
  Eagle      0       0     0      1     0      0
  Ford       0       0     0      1     0      0
  Honda      0       0     0      0     1      0
  Mitsbusihi 1       0     0      0     0      0
  Volkswagen 0       0     0      1     0      0
```

What if you want to arrange the data set by vehicle weight? *order()* gets the job done.

```
>i<-order(Car$Weight);i
```

[1] 1 5 3 4 2 7 10 6 9 8

```
> Car[i,]
         Make         Model Cylinder Weight

Mileage    Type

1       Honda          Civic     V4   2170       33  Sporty

5  Volkswagen          Jetta     V4   2330       26   Small

3        Ford         Escort     V4   2345       33   Small

4       Eagle         Summit     V4   2560       33   Small

2   Chevrolet       Beretta     V4   2655       26 Compact

7  Mitsbusihi        Galant     V4   2745       25 Compact
```

```
10        Acura          Legend     V6   3265     20  Medium
6         Buick          LeSabre    V6   3325     23   Large
9      Chrysler        NewYorker    V6   3450     22  Medium
8    Dodge Grand         Caravan    V6   3735     18     Van
```

Creating/editing data objects

```
>y<-c(1,2,3,4,5);y
```

[1] 1 2 3 4 5

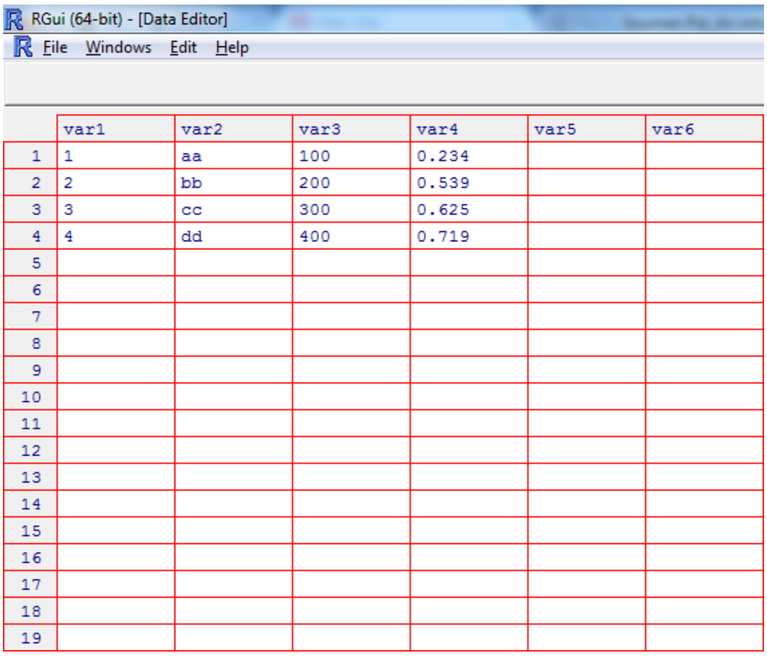
If you want to modify the data object, use *edit()* function and assign it to an object. For example, the following command opens R Editor for editing.

```
> y<-edit(y)
```

If you prefer entering the data.frame in a spreadsheet style data editor, the following command invokes the built-in editor with an empty spreadsheet.

```
> data1<-edit(data.frame())
```

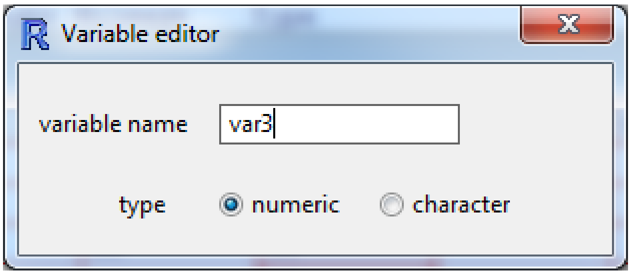
After entering a few data points, it looks like this:



	var1	var2	var3	var4	var5	var6
1	1	aa	100	0.234		
2	2	bb	200	0.539		
3	3	cc	300	0.625		
4	4	dd	400	0.719		
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

You can also change the variable name by clicking once on the cell containing it.

Doing so opens a dialog box:



variable name var3

type ● numeric ○ character

When finished, click ✕ in the upper right corner of the dialog box to return to the

Data Editor window. Close the Data Editor to return to the R command window (R Console). Check the result by typing:

```
> data1
```

Reading data from files

When data files are large, it is better to read data from external files rather than entering data through the keyboard. To read data from an external file directly, the external file should be arranged properly.

The first line of the file should have a name for each variable. Each additional line of the file has the values for each variable.

Input file form with names and row labels:

Price	Floor	Area	Rooms	Age	isNew
52.00	111.0	830	5	6.2	no
54.75	128.0	710	5	7.5	no
57.50	101.0	1000	5	4.2	yes
57.50	131.0	690	6	8.8	no
59.75	93.0	900	5	1.9	yes

...

By default numeric items (except row labels) are read as numeric variables and non-numeric variables, such as isNew in the example, as factors. This can be changed if necessary.

The function *read.table()* can then be used to read the data frame directly

```
>HousePrice<-read.table("d:/houses.data", header = TRUE)
```

Reading comma delimited data

The following commands can be used for reading comma delimited data into R.

read.csv(filename)	This command reads a .CSV file into R. You need to specify the exact filename with path.
read.csv(file.choose())	This command reads a .CSV file but the *file.choose()* part opens up an explorer type window that allows you to select a file from your computer. By default, R will take the first row as the variable names.
read.csv(file.choose(), header=T)	This reads a .CSV file, allowing you to select the file, the header is set explicitly. If you change to header=F then the first row will be treated like the rest of the data and not as a label.

Storing variable names

Through *read.csv()* or *read.table()* functions, data along with variable labels is read into R memory. However, to read the variables' names directly into R, one should use *attach(dataset)* function. For example,

```
>attach(HousePrice)
```

causes R to directly read all the variables' names eg. Price, Floor, Area etc. it is a good practice to use the *attach (datafile)* function immediately after reading the *datafile* into R.

Packages

All R functions and datasets are stored in packages. The contents of a package are available only when the package is loaded. This is done to run the codes efficiently without much memory usage. To see which packages are installed at your machine, use the command

```
>library()
```

To load a particular package, use a command like

```
>library(forecast)
```

Users connected to the Internet can use the *install.packages()* and *update.packages()* functions to install and update packages. Use *search()* to display the list of packages that are loaded.

Standard package

The standard (or base) packages are considered part of the R source code. They contain the basic functions those allow R to work with the datasets and standard statistical and graphical functions. They should be automatically available in any R installation.

Contributed packages and CRAN

There are a number of contributed packages for R, written by many authors. Various packages deal with various analyses. Most of the packages are available for download from CRAN (https://cran.r-project.org/web/packages/), and other repositories such as Bioconductor (http://www.bioconductor.org/). The collection of available packages changes frequently. As on June07, 2019, the CRAN package repository contains 14346 available packages.

Getting Help

Complete help files in HTML and PDF forms are available in R. To get help on a particular command/function etc., type *help (command name)*. For example, to get help on function 'mean', type *help(mean)* as shown below

```
>help(mean)
```

This will open the help file with the page containing the description of the function mean.

Another way to get help is to use "?" followed by function name. For example,

```
>?mean
```

will open the same window again.

In this lecture note, all R commands and corresponding outputs are given in `Courier New` font to differentiate from the normal texts. Since R is case-sensitive, i.e. typing *Help(mean)*, would generate an error message,

```
>Help(mean)
Error in Help(mean) : could not find function "Help"
```

Further Readings

Various documents are available in https://cran.r-project.org/manuals.html from beginners' level to most advanced level. The following manuals are available in pdf form:

1. An Introduction to R
2. R Data Import/Export
3. R Installation and Administration
4. Writing R Extensions
5. The R language definition
6. R Internals
7. The R Reference Index

RStudio

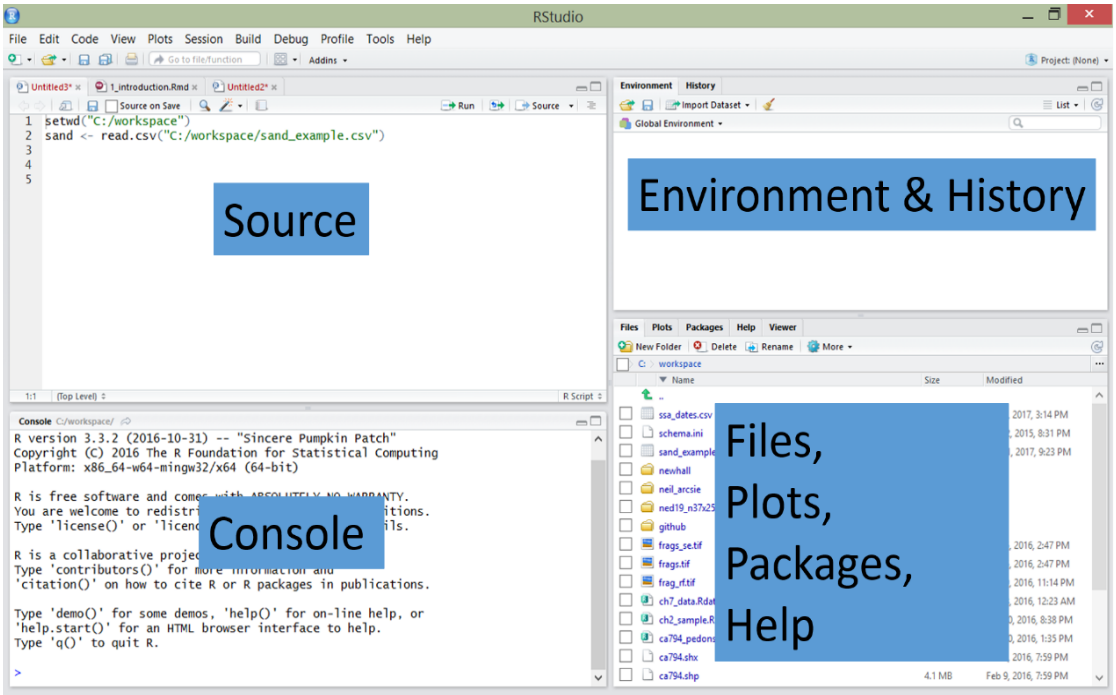
RStudio is an integrated development environment (IDE) that allows to interact with R more readily. RStudio is similar to the standard RGui, but is considerably more user friendly. It has more drop-down menus, windows with multiple tabs, and many customization options.

Installation of RStudio

RStudio requires R 3.0.1+ that means R software should be pre-installed before using RStudio.

RStudio requires a 64-bit operating system, and works exclusively with the 64 bit version of R. If you are on a 32 bit system or need the 32 bit version of R, you can use an older version of RStudio (https://www.rstudio.com/products/rstudio/older-versions/).RStudio free desktop version can be downloaded from the following link: https://www.rstudio.com/products/rstudio/download/#download

The first time RStudio is opened, three windows are seen. A forth window is hidden by default, but can be opened by clicking the **File** drop-down menu, then **New File**, and then **R Script**.

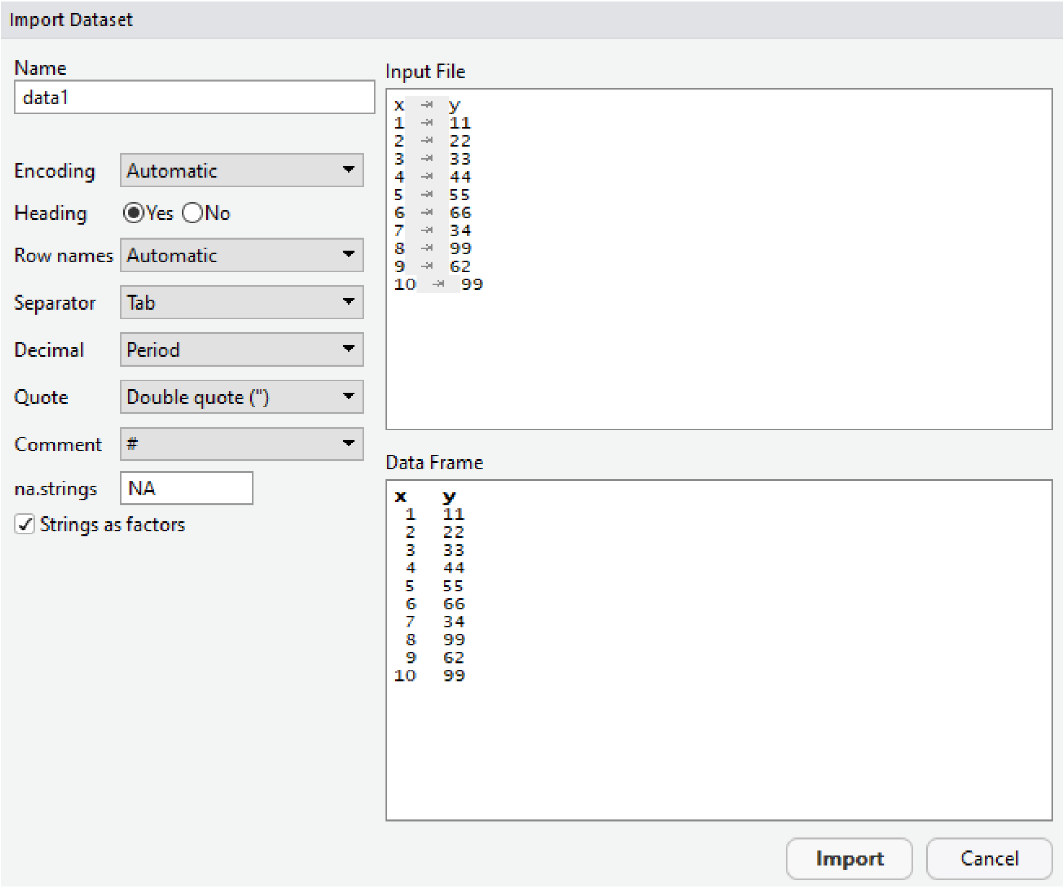


```
setwd("C:/workspace")
sand <- read.csv("C:/workspace/sand_example.csv")
```

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistri                 itions.
Type 'license()' or 'licenc                 ils.

R is a collaborative projec
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Importing Data in R Studio

1. Click on the import dataset button in the top-right section under the environment tab. Select the file you want to import and then click open. The Import Dataset dialog will appear as shown below
2. After setting up the preferences of separator, name and other parameters, click on the Import button. The dataset will be imported in R Studio and assigned to the variable name as set before.



Import Dataset

Name: data1

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Automatic

Separator: Tab

Decimal: Period

Quote: Double quote (")

Comment: #

na.strings: NA

☑ Strings as factors

Input File

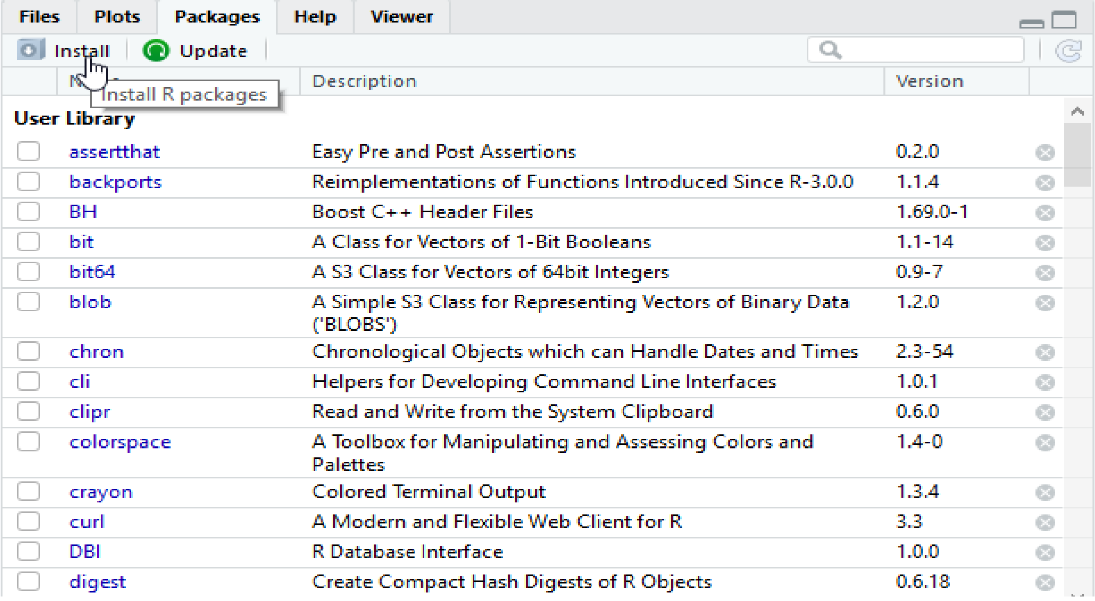
```
x  →  y
1  →  11
2  →  22
3  →  33
4  →  44
5  →  55
6  →  66
7  →  34
8  →  99
9  →  62
10  →  99
```

Data Frame

```
x   y
 1  11
 2  22
 3  33
 4  44
 5  55
 6  66
 7  34
 8  99
 9  62
10  99
```

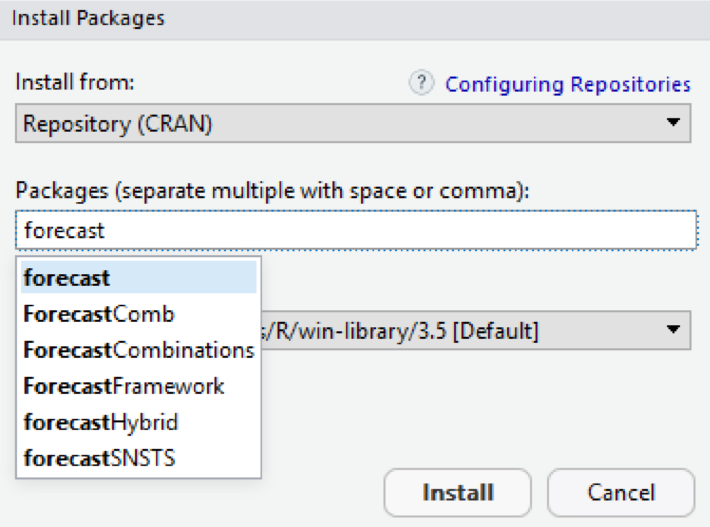
Installing Packages in RStudio

Within the **Packages** tab, a list of all the packages currently installed on the working computer and 2 buttons labeled either "Install" or "Update" are seen. To install a new package simply select the Install button. It is possible to install one or more than one packages at a time by simply separating them with a comma.



User Library

	Name	Description	Version
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.4
□	BH	Boost C++ Header Files	1.69.0-1
□	bit	A Class for Vectors of 1-Bit Booleans	1.1-14
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.0
□	chron	Chronological Objects which can Handle Dates and Times	2.3-54
□	cli	Helpers for Developing Command Line Interfaces	1.0.1
□	clipr	Read and Write from the System Clipboard	0.6.0
□	colorspace	A Toolbox for Manipulating and Assessing Colors and Palettes	1.4-0
□	crayon	Colored Terminal Output	1.3.4
□	curl	A Modern and Flexible Web Client for R	3.3
□	DBI	R Database Interface	1.0.0
□	digest	Create Compact Hash Digests of R Objects	0.6.18



Install Packages

Install from: Repository (CRAN)

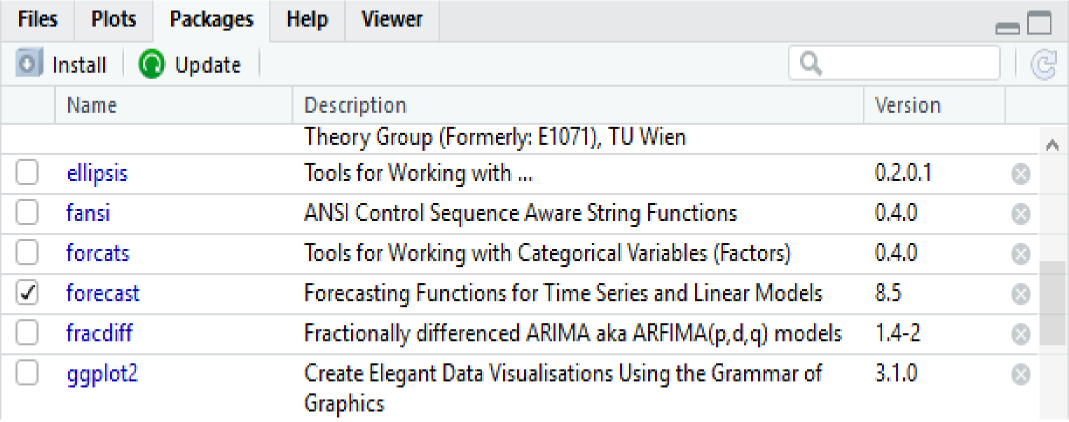
Packages (separate multiple with space or comma): forecast

forecast
ForecastComb
ForecastCombinations
ForecastFramework
forecastHybrid
forecastSNSTS

:/R/win-library/3.5 [Default]

Loading Packages in RStudio

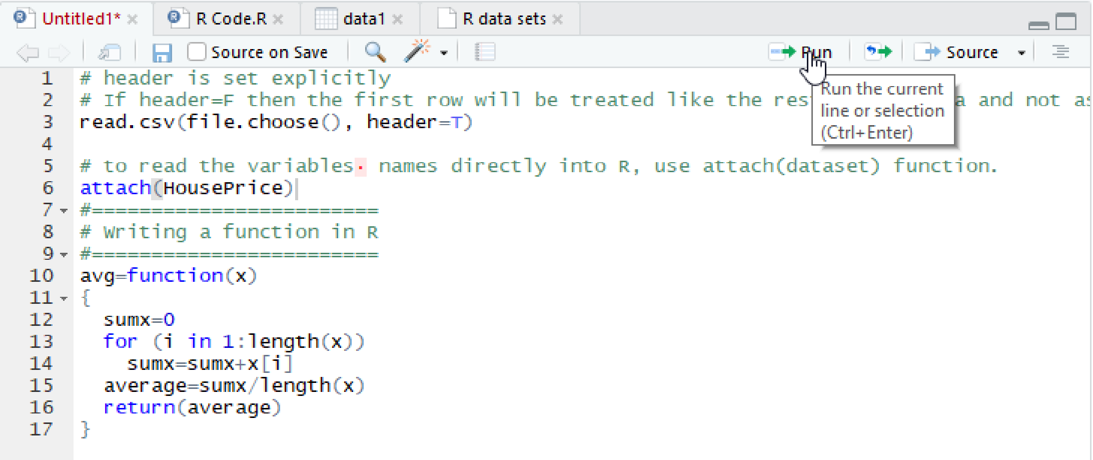
Once a package is installed, it must be loaded into the R session to be used.



	Name	Description	Version
		Theory Group (Formerly: E1071), TU Wien	
□	ellipsis	Tools for Working with ...	0.2.0.1
□	fansi	ANSI Control Sequence Aware String Functions	0.4.0
□	forcats	Tools for Working with Categorical Variables (Factors)	0.4.0
☑	forecast	Forecasting Functions for Time Series and Linear Models	8.5
□	fracdiff	Fractionally differenced ARIMA aka ARFIMA(p,d,q) models	1.4-2
□	ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	3.1.0

Writing Scripts in RStudio

RStudio's Source Tabs serve as a built-in text editor. Prior to executing R functions at the Console, commands are typically written down (or scripted).To write a script, simply open a new R script file by clicking File>New File>R Script.



```
# header is set explicitly
# If header=F then the first row will be treated like the res      a and not as
read.csv(file.choose(), header=T)

# to read the variables' names directly into R, use attach(dataset) function.
attach(HousePrice)
#========================
# writing a function in R
#========================
avg=function(x)
{
  sumx=0
  for (i in 1:length(x))
    sumx=sumx+x[i]
  average=sumx/length(x)
  return(average)
}
```

Within the text editor type out a sequence of functions.

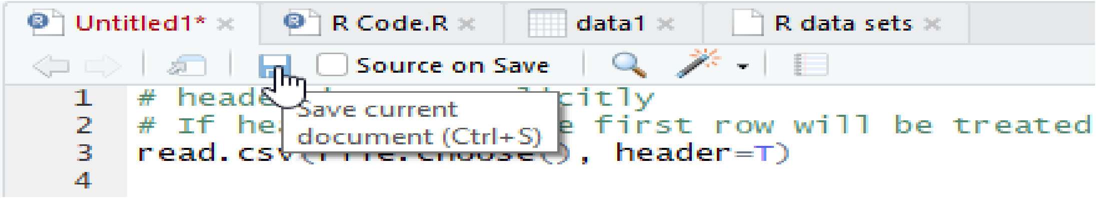
- Place each function (e.g. read.csv()) on a separate line.
- If a function has a long list of arguments, place each argument on a separate line.
- A command can be executed from the text editor by placing the cursor on a line and typing Crtl + Enter, or by clicking the Run button.
- An entire R script file can be executed by clicking the Source button.

Saving R files in RStudio

In R, several types of files can be saved to keep track of the work performed. The file types include: script, workspace, history and graphics.

R script (.R)

An R script is a text file of R commands that have beentyped. To save R scripts in RStudio, click the save button from R script tab. Save scripts with the .R extension.

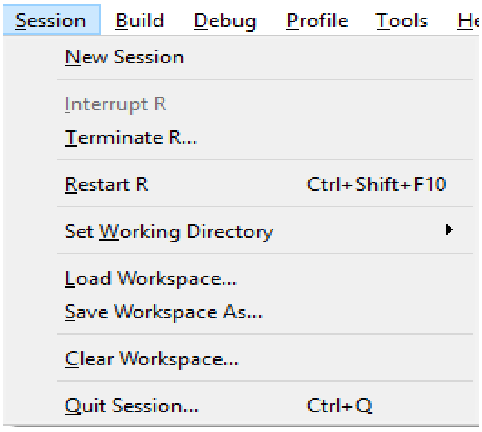


```
# heade            citly
# If he           e first row will be treated
read.csv                 , header=T)
```

To open an R script, click the file icon.

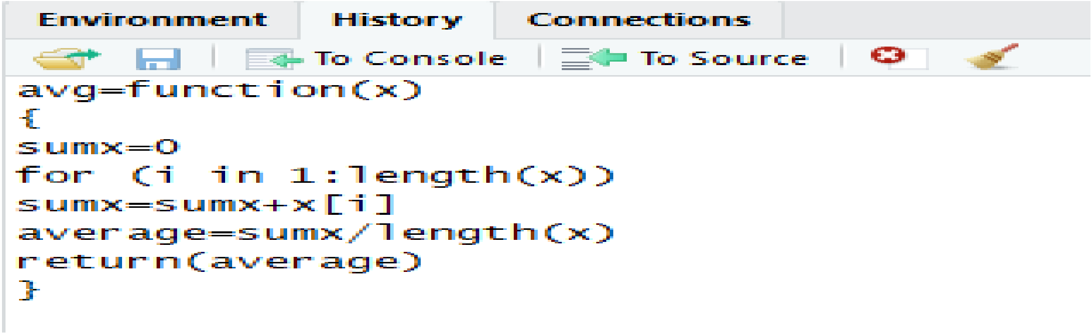
Workspace (.Rdata)

The R workspace consists of all the data objects created or loaded during the R session. It is possible to save or load the workspace at any time during the R session from the menu by clicking Session>Save Workspace As.., or the save button on the Environment Tab.



R history (.Rhistory)

Rhistory file is a text file that lists all of the commands that have been executed. It does not keep a record of the results. To load or save R history from the History Tab click the **Open File** or **Save** button.



```
avg=function(x)
{
sumx=0
for (i in 1:length(x))
sumx=sumx+x[i]
average=sumx/length(x)
return(average)
}
```

R Graphics

Graphic outputs can be saved in various formats like pdf, png, jpeg, bmp etc.

To save a graphic: (1) Click the **Plots** Tab window, (2) click the **Export** button, (3) **Choose** desired format, (4) **Modify** the export settings as desired and (4) click **Save**.

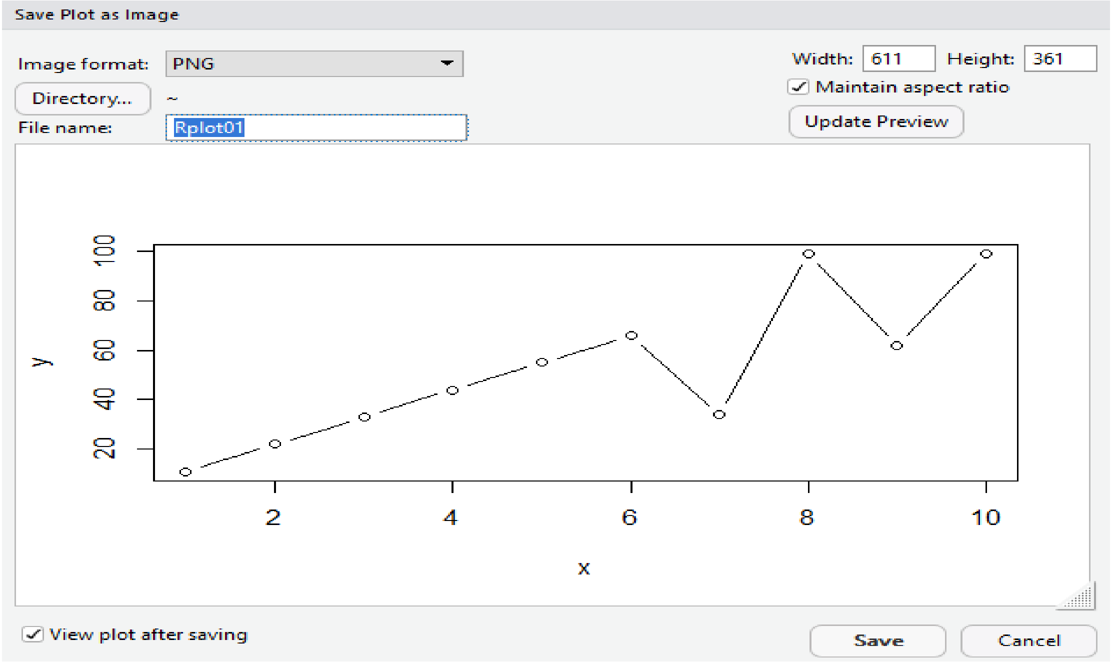


Image format: PNG

Directory... ~

File name: Rplot01

Width: 611 Height: 361

☑ Maintain aspect ratio

☑ View plot after saving

References

1. http://ncss-tech.github.io/stats_for_soil_survey/chapters/1_introduction/1_introduction.html
2. http://web.cs.ucla.edu/~gulzar/rstudio/basic-tutorial.html
3. http://www.gardenersown.co.uk/Education/Lectures/R/index.htm
4. https://www.cran.r-project.org
5. https://www.rstudio.com
6. Matloff, N. (2011). *The art of R programming: A tour of statistical software design*. No Starch Press.
7. Venables, W. N., Smith, D. M. and R Development Core Team (2009). An introduction to R: Notes on R: A programming Environment for Data Analysis and Graphics, version 1.7. 1.

DESCRIPTIVE STATISTICS AND EXPLORATORY DATA ANALYSIS

Md Yeasin
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
E-mail: yeasin.iasri@gmail.com

1. Introduction

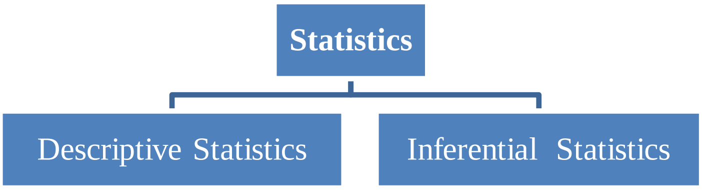
The word ‘Statistics’ has been derived from the Latin word ‘**Status**’ or the Italian word ‘**Statista**’ or the German word ‘**Statistik**’ each of which means ‘political state’. Statistics is a broad concept featuring applications in a wide range of areas. Statistics, in general, can be defined as the process for collecting, analyzing, interpreting, and making conclusions from data. In other terms, statistics is the approach established by scientists and mathematicians for analyzing and deriving conclusions from acquired data. Everything that has anything to do with the collection, processing, interpretation, and presentation of data falls within the scope of statistics.

Definition of statistics: Statistics is a branch of mathematics that deals with collecting, organizing, summarizing, presenting, and analyzing data as well as providing valid results and interpreting towards reasonable decisions.

Statisticians, in other words, give methodologies for

- **Design:** Planning and conducting out research projects.
- **Description:** Data summarization and exploration.
- **Inference:** Making predictions and inferences about the data

Statistics can be divided into two sections; one is descriptive statistics and another is inferential statistics.



Descriptive statistics helps describe, show or summarize data in a meaningful way. Descriptive statistics provides us with tools, tables, graphs, averages, ranges, correlations for organizing and summarizing data. Examples: measures of central tendency, measures of dispersion, skewness, kurtosis etc.

Inferential statistics helps to understand the properties of the population by observing the sample values. Inferential statistics deals with the estimation of parameters and test of hypothesis.

In this section we briefly discussed the descriptive statistics such as measures of central tendency, measures of dispersion, skewness, and kurtosis

2. Measures of central tendency

Central tendency is a statistical measure that determines a single value that accurately describes the center of the distribution. The objective of central tendency is to identify the single value that is the best representative for the entire set of data.

Different measure of central tendency are:

- Mean
    - Arithmetic mean
    - Geometric mean
    - Harmonic mean
- Median
- Mode
- Quartiles
- Deciles
- Percentiles

1.1.Mean (Arithmetic mean: A.M.):

The mean is the most commonly used measure of central tendency. For computation of the mean data should be numerical values measured on an interval or ratio scale. To compute the mean, we add the observation of data sets and then divide by the number of observation.

$$Mean = \frac{Sum\ of\ all\ observation}{Total\ number\ of\ observationa}$$

1.1.1. Simple mean: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The arithmetic mean is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Mean for frequency distribution: Let $X_1, X_2, \dots, X_n$ are observations with correspondingfrequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The arithmetic mean is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$$

Properties of mean:

- It depends on change of origin as well as the change of scale.

$$U = a + hX$$

Where a is origin and h is scale

Then $\bar{U} = a + h\bar{X}$.

- If are$\bar{X}_1$and $\bar{X}_2$ the means of two sets of values with $n_1$and n_2 observations respectively, then their combined mean is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

- Algebraic sum of deviations of set of values from their mean is zero.

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

- The sum of squares of deviation of set of values about its mean is minimum

$\sum_{i=1}^{n}(X_i - A)^2$ is minimum when $A = \bar{X}$

Merits of mean:

- Easy to understand
- Easy to calculate.
- It is rigidly defined.
- It is based on all observations.
- It is least affected by sampling fluctuations.
- It is capable of further mathematical treatment.

Demerits of mean:

- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.
- It cannot be calculated if any observations are missing in the data series.
- It is not suitable for highly skewed distribution.

1.1.2. Geometric mean (G.M.):

For n observations, Geometric mean is the nth root of their product.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The geometric mean is defined as

$$G = (X_1 * X_2 * \dots * X_n)^{1/n}$$

For frequency distribution: Let $X_1, X_2, \dots, X_n$ are observations with correspondingfrequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The geometric mean is defined as

$$G = ({X_1}^{f_1} * {X_2}^{f_2} * \dots * {X_n}^{f_n})^{1/N}$$

Use of geometric mean:

- Measure average relative changes, averaging ratios and percentages
- Best average for construction of index number

Merits of geometric mean:

- It is based on all observations.
- It is not affected by sampling fluctuations.
- It is capable of further mathematical treatment.

Demerits of geometric mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.
- It cannot be calculated if any observations are missing in the data series.

1.1.3. Harmonic mean (H.M.):

Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the observations of the sets.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The harmonic mean is defined as

$$H = \frac{n}{\sum_{i=1}^{n} 1/X_i}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The harmonic mean is defined as

$$H = \frac{N}{\sum_{i=1}^{n} f_i/X_i}$$

Use of harmonic mean:

- Measure the change where the values of a variable are compared with a constant quantity of another variable like time, distance travelled within a given time, quantities purchased or sold over a unit.

Merits of harmonic mean:

- It gives more weight to the small item and less weight to large values.
- It is based on all observations.
- It is not affected by sampling fluctuations.

- It is capable of further mathematical treatment.

Demerits of harmonic mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristics.
- It cannot be calculated if any observations are missing in the data series.

Relation between A.M., G.M. and H.M.:

- For given two observations, $A.M. \geq G.M. \geq H.M.$
- $G.M. = \sqrt{A.M. * H.M.}$
- $A.M. = \frac{G.M.^2}{H.M.}$
- $H.M. = \frac{G.M.^2}{A.M.}$

1.2. Median:

Median is the value situated in the middle position when all the observations are arranged in an ascending/descending order. The median is the central value of an ordered data series. It divides the data sets exactly into two parts. Fifty percent of observations are below the median and 50% are above the median. Median is also known as 'positional average'. The Median is the 50th percentiles, 10th deciles, and 2nd quartiles. Median is also the intersect point of less than and more than ogive curve.

Median for non-frequency data:

Step 1 Order the data from smallest to largest.

Step2 If the number of observations is odd, then (n + 1)/2th observation (in the ordered set) is the median. When the total number of observations is even, the median is given by the mean of n/2th and (n/2 + 1)th observation.

Median for group frequency data:

Step 1 Obtain the cumulative frequencies for the data.

Step 2 Markthe class corresponding to which a cumulative frequency is greater than N/2. That class is the median class.

Step 3 Then median is evaluated by an interpolation formula

$$Median = l + \frac{h}{f}(\frac{N}{2} - C)$$

Where, l = lower limit of the median class

N= Number of observations

C = cumulative frequency of the class proceeding to the median class

f = frequency of the median class

h= magnitude of the median class

Note: Graphically, we can find the median by histogram.

Use of median:

- Qualitative data can be arranged in ascending or descending order of magnitude.
- Find average intelligence, honesty, etc.

Merits of median:

- It is rigidly defined.
- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on an ordinal scale.

Demerits of median:

- It is not based on all observations.
- The calculation is more complex than the mean.
- It is not capable of further mathematical treatment.
- As compared to the mean, it is much affected by sampling fluctuations.

1.3 Mode:

Mode is defined as the value that occurs most frequently in the data. If in the data sets each observation occurs only once, then it does not have mode. When the data set has two or more values equal to the highest frequency than two or more mode are present in the datasets.

Mode for ungroup frequency data: The observation which has the highest frequency in the data sets.

Mode for group (equal width) frequency data:

Step 1 Identify the modal class. Modal class is the class with the largest frequency.

Step 2 Find mode by using interpolated formula.

$$mode = l + \frac{h(f_0 - f_{-1})}{(f_0 - f_{-1}) - (f_1 - f_0)}$$

Where, l = lower limit of the modal class

f_0 = frequency of the modal class

f_{-1}=frequency of the preceding modal class

f_1=frequency of the succeeding modal class

h= magnitude of the modal class

Note: Graphically, we can find mode by histogram.

Use of mode:

- To find ideal consumer preferences for different kinds of products.
- The best measure for the average size of shoes or shirts.

Merits of mode:

- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on a nominal scale.

Demerits of mode:

- It is ill-defined.
- It is not based on all observations.
- The calculation is more complex than the mean.
- It is not capable of further mathematical treatment.
- As compare to the mean, it is much affected by sampling fluctuations.

Quartiles: Quartiles are the three points that divide the whole data into four equal parts.

$$Q_i = l + \frac{h}{f}(\frac{iN}{4} - C)$$

Deciles: Deciles are the nine points that divide the whole data into ten equal parts.

$$D_i = l + \frac{h}{f}(\frac{iN}{10} - C)$$

Percentiles: Percentiles are the ninety-nine point that divides the whole data into hundreds of equal parts.

$$P_i = l + \frac{h}{f}(\frac{iN}{100} - C)$$

Note: $\boldsymbol{Median = 2nd\ Quartles = 5th\ Deciles = 50th\ Percentiles}$

Empirical formula between mean median and mode: If the data sets area symmetric in nature, then

$$\boldsymbol{Mean - \ Mode\ = 3(Mean - \ Median)}$$

The best measure of central tendency:

According to proof. Yule, Mean is the best measure of central tendency. But there are some situations where the other measures of central tendency are preferred.

Scale	Use measure	Best measure
Interval	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ratio	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ordinal	Median, Mode	Median
Nominal	Mode	Mode

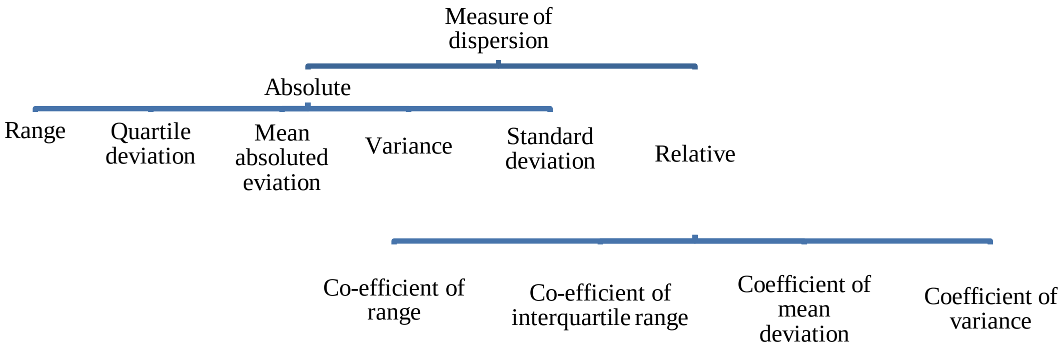
2. Measure of Dispersion

The measure of central tendency such as mean, median, and mode only locate the center of the data. It does not infer anything about the spread of the data. Two data sets can have the same mean but they can be entirely different.

Data 1	38	42	41	44	45
Data 2	50	53	41	35	31

In the above example, two datasets have the same mean. So measures of central tendency are not adequate to describe data. Thus to describe data, one needs to know the measure of scatterness of observations. Dispersion is defined as deviation or scatterness of observations from their central values.

Various measure of dispersion are:



1.2 Range (R):

Range is the simplest measure of dispersion. It is defined as the difference between the highest value and lowest value of the variable. It is a crude measure of dispersion.

$$\boldsymbol{Range = highest\ value\ (H) - lowest\ value\ (L)}$$

Merits of range:

- It is easy to understand and calculate.
- It is not affected by frequency of the data.

Demerits of range:

- It does not depend on all observations.
- It is very much affected by the extreme items.
- It cannot be calculated from open-end class intervals.
- It is not suitable for further mathematical treatment.
- It is the most unreliable measure of dispersion.

1.3 Quartile deviation (Q.D.):

Interquartile range is the difference between the first and third quartile. Hence the interquartile range describes the middle 50% of observations.

$$\boldsymbol{Inter\ quartile\ range = Q3 - Q1}$$

Where,

Q^3=first quartile of the data

Q^1=third quartile of the data

Quartile deviation (Q.D.) is the half of the inter quartile range.

$$\boldsymbol{Quartile\ deviation\ (Q.D.) = \frac{Q3 - Q1}{2}}$$

Merits of Quartile deviation:

- It is easy to understand and calculate.
- It is not affected by extreme values
- It can be calculated for open end frequency data

Demerits of Quartile deviation:

- It does not depend on all observations.
- It is not suitable for further mathematical treatment.
- It is very much affected by sampling fluctuations.

1.4 Mean absolute deviation (MAD):

The absolute deviation of each value from the central value (mean is preferable) is calculated and the arithmetic mean of these deviations is called mean absolute deviation.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observations of a data set. The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^{n} |X_i - A|}{n}$$

The mean absolute deviation (MAD) about mean is given by

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^{n} f_i |X_i - A|}{N}$$

The mean absolute deviation (MAD) about mean is given by

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} f_i |X_i - \bar{X}|}{N}$$

Merits of mean absolute deviation about mean:

- It is easy to understand and calculate.
- It is based on all observations.

Demerits of mean absolute deviation about mean:

- It is not suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.

1.5 Standard deviation (S.D.):

It is the best measure and the most commonly used measure of dispersion. It is defined as the positive square-root of the arithmetic mean of the square of the deviations of the given observation from their arithmetic mean. It takes into consideration the magnitude of all the observations and gives the minimum value of dispersion possible. It is also known as Root Mean Square Deviation about mean.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The standard deviation A is given by

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The standard deviation is given by

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^2}{N}}$$

Properties of standard deviation:

- It is the independent of the change of origin but dependent on the change of scale
  Let $U = a + hX$, then $sd(U) = |h| * sd(x)$
- If all observations are equal standard deviation is zero.
- It is never less than the quartile deviation and mean absolute deviation.

Merits of standard deviation:

- It is based on all observations.
- It is less affected by extreme values.
- It is suitable for further mathematical treatment.

Demerits of standard deviation:

- It is suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.
- It cannot be computed for open-end class data.

1.6 Variance

It is defined as the square of the standard deviation. Unit of the variance is the square of the actual observations, whereas unit of the standard deviation is same as actual observations.

Relations between R, Q.D., M.D. and S.D.

$$\mathbf{9QD = \frac{15}{2}MD = 6SD = R}$$

1.7 Coefficient of Variation (CV):

The Coefficient of variation for a data set defined as the ratio of the standard deviation to the mean and expressed in percentage.

$$\boldsymbol{CV = \frac{SD}{mean} * 100\%}$$

C.V is the relative measure of dispersion. It is the best measure among all the relative measure of dispersion. C.V is used to compare variability or consistency between two or more data series. If C.V. is greater indicate that the group is more variable, less stable, less uniform and less consistent. If the C.V. is less, it indicates that the group is less variable or more stable or more uniform and more consistent.

Example: Consider the data on score of Kohli and Smith in ODI cricket. The mean and standard deviation for Kohli are 55 and 5 respectively. The mean and standard deviation for Smith are 50 and 10 respectively. Find C.V. value for both the data and make compare them.

Solution:

For Kohli, $CV = \frac{5}{55} * 100 = 9\%$

For Smith, $CV = \frac{10}{50} * 100 = 20\%$

The Smith is subject to more variation in score than Kohli. So Kohli is more consistent than Smith.

$$\mathbf{3.6. Coefficient\ of\ range} = \frac{\boldsymbol{H-L}}{\boldsymbol{H+L}} * \mathbf{100\%}$$

$$\mathbf{3.7.\ Coefficient\ of\ inter\ quartile\ range} = \frac{\boldsymbol{Q3-Q1}}{\boldsymbol{Q3+Q1}} * \mathbf{100\%}$$

$$\mathbf{3.8. Coefficient\ of\ mean\ deviation}$$

$$= \frac{\boldsymbol{MAD}}{\boldsymbol{averave\ from\ which\ it\ is\ calculated}} * \mathbf{100\%}$$

Numerical Examples: The marks of 10 students in statistics examination are as follows:

$$10,12,15,12,16, 20, 13,17,15,15$$

Find mean, median, mode, range and standard deviation.

Solution:

X_i	f_i	$f_i X_i$	$f_i(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$f_i(X_i - \overline{X})^2$
10	1	10	-4.5	20.25	20.25
12	2	24	-5	6.25	12.5
13	1	13	-1.5	2.25	2.25
15	3	45	1.5	0.25	0.75
16	1	16	1.5	2.25	2.25
17	1	17	2.5	6.25	6.25
20	1	20	5.5	30.25	30.25
Total	10	145		67.75	74.5

$$mean = \frac{145}{10} = 14.5$$
$$median = 15$$
$$mode = 15$$
$$range = 20 - 10 = 10$$
$$SD = \frac{74.5}{10} = 7.45$$

2 Skewness and kurtosis:

We have discussed measures of central tendency and measure of dispersion which describe the location and scale parameter of the data sets. They do not give any idea about the shape of the data structure. The measure of skewness and kurtosis illustrate the shape of the data sets. The measure of skewness gives the direction and the magnitude of the lack of symmetry and the measure of kurtosis gives the idea of the flatness of the curve.

2.2 Skewness

Skewness measures the degree of asymmetry of the data. Skewness refers to the lack of symmetry.

Skewness is mainly three types: Positive skewness, Negative skewness, and Symmetric data.

Positive Skewness:

A data is said to be positive skew if the long tail is on the right side of the peak. The mean is on the right of the peak value. Here Mean > Median > Mode.

Negative Skewness:

A data is said to be negative skew if the long tail is on the left side of the peak. The mean is on the left of the peak value. Here Mean < Median < Mode.

Symmetric

The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle. When data is symmetrically distributed, the left-hand side and right-hand side, contain the same number of observations. Here Mean = Median = Mode.

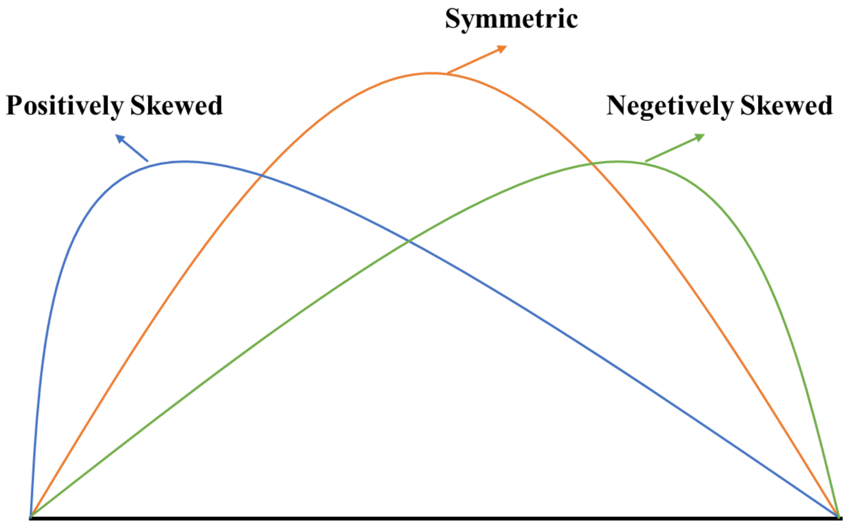


Figure 1. Skewness

The measure of Skewness:

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Interpretation:

1. If $S_k = 0$, then the frequency distribution is normal and symmetrical.

2. If $S_k > 0$, then the frequency distribution is positively skewed.

3. If $S_k < 0$, then the frequency distribution is negatively skewed.

2.3 Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails or outliers. Data sets with low kurtosis tend to have light tails or lack of outliers. A uniform distribution would be the extreme case.

Types of kurtosis: Leptokurtic or heavy-tailed distribution, Mesokurtic, Platykurtic or short-tailed distribution

Leptokurtic

Leptokurtic indicates that distribution is peaked and possesses thick tails.

Platykurtic

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is a flatter (less peaked) when compared with the normal distribution.

Mesokurtic

Mesokurtic is the same as the normal distribution. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

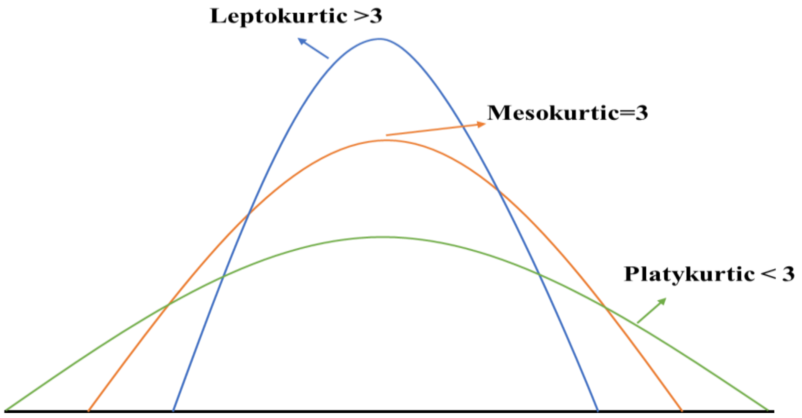


Figure 2. Kurtosis

Measurement of Kurtosis $(\beta_2) = \frac{1}{N-1}\frac{\sum(y_i-\bar{y})^4}{s^4}$

$\gamma_2 = \beta_2 - 3$

Data presentation

Non dimensional diagram	Pictograms
Two dimensional diagram	Bar diagram, Pie diagrams, Histograms, Box Plot
Three dimensional diagram	Cubes, Cylinders diagrams

There are three broad ways of presenting data. These are Textual presentation, Tabular presentation, and Graphic or diagrammatic presentation. We discussed only a few important diagrammatic presentations of data.

2.4 Bar Diagram

2.4.1 Simple Bar Diagram

If the classification is based on attributes and if the attributes are to be compared with respect to a single character we use a simple bar diagram. Simple bar diagrams consist of vertical bars of equal width. The heights of these bars are proportional to the volume or magnitude of the attribute. All bars stand on the same baseline. The bars are separated from each other by equal intervals. The bars may be colored or marked.

2.4.2 Multiple bar diagram

If the data is classified by attributes and if two or more characters or groups are to be compared within each attribute we use multiple bar diagrams. If only two characters are to be compared within each attribute, then the resultant bar diagram used is known as the double bar diagram. The multiple bar diagram is simply the extension of a simple bar diagram. For each attribute, two or more bars representing separate characters or groups are to be placed side by side. Each bar within an attribute will be marked or colored differently in order to distinguish them. The same type of marking or coloring should be done under each attribute. A footnote has to be given explaining the markings or colorings.

2.4.3 Component bar diagram

This is also called a subdivided bar diagram. Instead of placing the bars for each component side by side, we may place this one on top of the other. This will result in a component bar diagram.

2.5 Histogram

Histograms is suitable for continuous class frequency distribution. We mark off class intervals along the x-axis and frequencies (frequency density for unequal frequency data)along the y-axis.

- For equal class intervals, the heights of the rectangles will be proportional to the frequencies, while for unequal class intervals, the heights will be equal (or proportional) to the frequency densities.
- A frequency polygon is a line graph obtained by connecting the midpoints of the tops of the rectangles in the histogram.

Table 1. Differences between bar diagrams and histograms

Characteristics	**Bar Diagrams**	**Histograms**
Frequency is measured by	Height of the bar	Area of the bar
Gaps between the bars	Yes	No
Width of the bar	Equal	May not be equal
Data types	Discrete and Continuous	Continuous only

2.6 Pie diagrams

When we are interested in the relative importance of the different components of a single factor, we use pie diagrams. For the pie diagram, one circle is used and the area enclosed by it being taken as 100. Itis then divided into a number of sectors by drawing angles at the center, the area of each sector representing the corresponding percentage.

2.7 Box Plot

Minimum, maximum, and quartiles ($Q_{1,}$ Median$_{,}$ Q_3) together provide information on the center and variation of the variable in a nice compact way. Written in increasing order, they comprise what is called the five-number summary of the variable. A box plot is based on the five-number summary and can be used to provide a graphical display of the center and variation of the observed values of the variable in a data set. It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

N.B: Examples of graphical presentation have been given in our basic statistics with excel manual.

3 Robust Estimate of Mean and Standard Deviation

The mean and standard deviation provides a correct estimation only if the variable is normally distributed and without outliers. If the variable is skewed and/or has outliers, the mean and standard deviation will be excessively influenced by the extreme observations and provide faulty statistics of data. There are many alternatives to the mean and standard deviation. Alternatives to the mean include the well-known median and trimmed mean, Winsorized mean, and M-estimators and for standard deviation, the alternatives include the Inter-Quartile Range (IQR) and the Median Absolute Deviation (MAD), Trimmed standard deviation, the Winsorized standard deviation, and M-estimators. Median, IQR, MAD are already discussed in the previous section in detail. Here we only discussed the trimmed, Winsorized, and M estimators for mean and standard deviation.

3.2 Trimmed Mean and Standard Deviation

A trimmed mean and standard deviation is similar to a "regular" mean but it trims any outliers from both the side. To obtain the 20% trimmed mean, the 20% lowest and 20 % highest values are removed and the mean is computed on the remaining observations. In our example, these values will be: 4, 4, 5, 5, 6, 6, and the 20% trimmed mean will be equal to 5.

3.3 Winsorized Mean and Standard Deviation

The Winsorized technique is similar to the trimmed technique but the lowest (resp. highest) values are not removed but replaced by the lowest (resp. highest) untrimmed score. In our example, the values of the variables, also called Winsorized scores, will then be: 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, and the 20% Winsorized mean will be equal to 5.

3.4 M estimators

The trimmed mean all either take or drop observations. As for the Winsorized mean, it replaces values with less extreme values. In contrast, the M-estimators, weight each observation according to a function selected for its special properties. The weights depend on a constant that can be chosen by the researcher. The M-estimator solves this problem of assigning a zero value to many observations by down weighting the observations progressively. The only aspect of the M-estimator that could worry substantive researchers is that one must choose the degree of down weighting of the observations.

STATISTICAL DATA ANALYSIS USING MICROSOFT EXCEL

Sanchita Naha
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-12
Sanchita.naha@icar.gov.in

Statistics is the study of collection, analysis, interpretation, presentation, and organization of data. Broadly, two statistical methodologies are used for data analysis, descriptive statistics, and inferential statistics. Statistical analysis can be done using software like MS Excel, SPSS, R but this tutorial is restricted to major statistical analysis methods using Microsoft Excel. Statistical analysis mainly encompasses descriptive statistics and inferential statistics.

1. **Descriptive Statistics:** Descriptive statistics is used to describe or summarize data in a meaningful way. Descriptive statistics provides us with tools, tables, graphs, averages, ranges, correlations for organizing and summarizing data. In descriptive statistics data is summarized with the following major numerical descriptors like

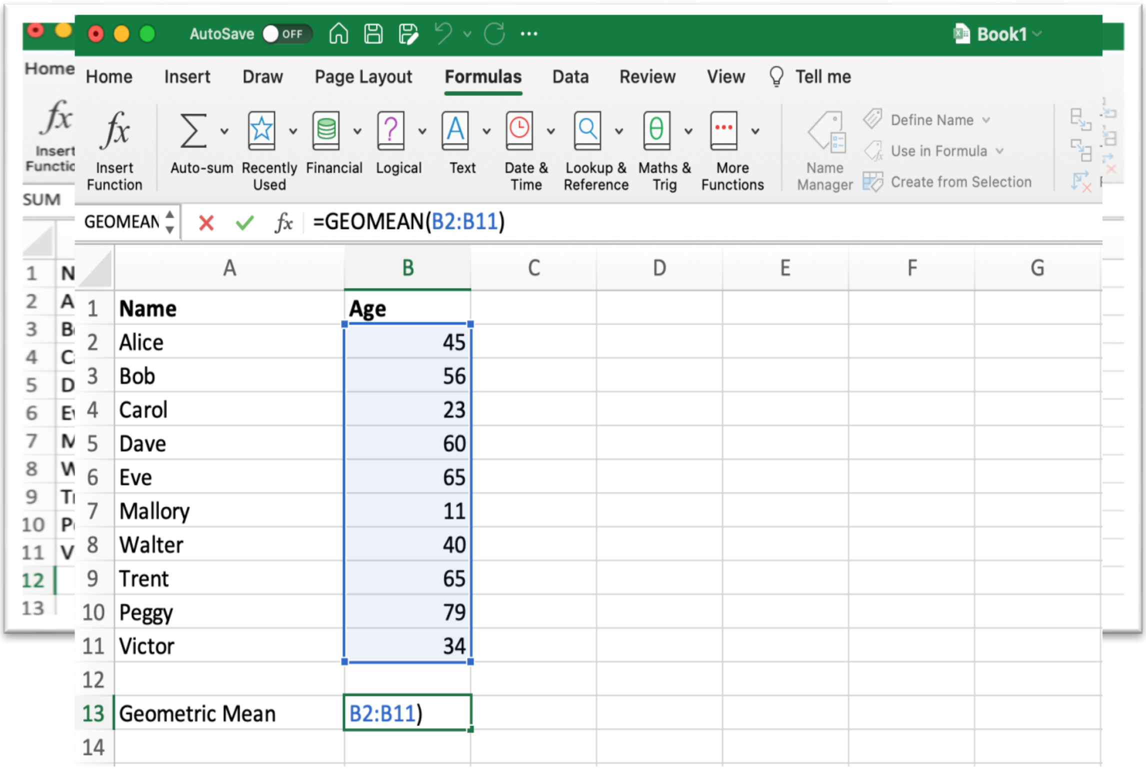
- **Arithmetic Mean:** It is defined as the average of the data values. For mean computation, data must be in numeric form.

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n}$$

n = number of observations

Steps to compute mean in Excel:

Select data points > Click formulas > Expand auto-sum drop down menu >



=GEOMEAN(B2:B11)

	A	B
1	Name	Age
2	Alice	45
3	Bob	56
4	Carol	23
5	Dave	60
6	Eve	65
7	Mallory	11
8	Walter	40
9	Trent	65
10	Peggy	79
11	Victor	34
12		
13	Geometric Mean	B2:B11)
14		

select Average > Enter.

- **Geometric Mean:** It is the n^{th} root of the product of individual data points. Let X_1, X_2, ... , X_n be the n^{th} observation of a data set. The geometric mean is defined as

$$GM = (x_1 * x_2 * \ldots * x_n)^{1/n}$$

Steps to compute geometric mean in Excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'GEOMEAN' > click 'Insert Function' > Enter.

e.g., GEOMEAN (B2:B11)

The geometric mean is used in finance to calculate average growth rates and is referred to as the compounded annual growth rate.

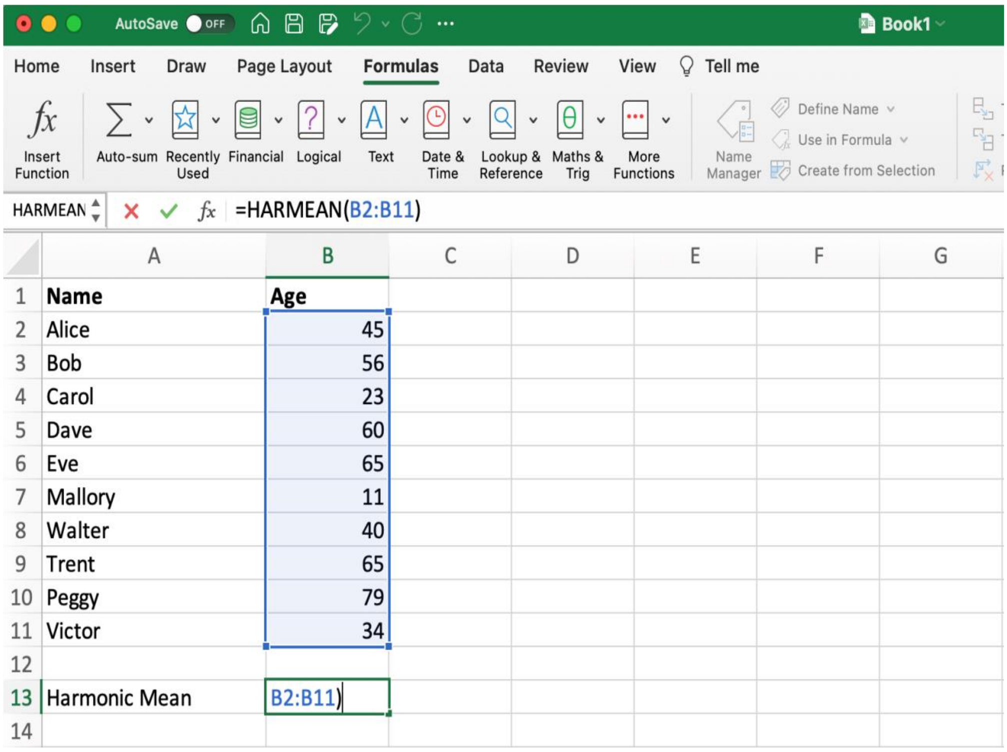
- **Harmonic Mean:** Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the observations of the datasets.

$$HM = \frac{n}{\sum_{i=1}^{n} {}^{1}/{}_{x_i}}$$

Steps to compute harmonic mean in Excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'HARMEAN' > click 'Insert Function' > Enter.

e.g., HARMEAN (B2:B11)



HARMEAN =HARMEAN(B2:B11)

	A	B	C	D	E	F	G
1	Name	Age					
2	Alice	45					
3	Bob	56					
4	Carol	23					
5	Dave	60					
6	Eve	65					
7	Mallory	11					
8	Walter	40					
9	Trent	65					
10	Peggy	79					
11	Victor	34					
12							
13	Harmonic Mean	B2:B11)					
14							

- **Median:** Median is the value in the middlemost position of all the observations when arranged in an ascending/descending order. The median is the central value of an ordered data series. It divides the data sets exactly into two parts. Fifty percent of observations are below the median value and 50% are above the median. Median is also known as 'positional average'.

  Steps to compute median in Excel:

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘MEDIAN’ > click ‘Insert Function’ > Enter.

- **Mode:** Mode is defined as the value that occurs most frequently in the data. If in the data sets each observation occurs only once, then it does not have mode. When the data set has two or more values equal to the highest frequency than two or more mode are present in the datasets.

  Steps to compute median in Excel:

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘MODE’ > click ‘Insert Function’ > Enter.

- **Range:** It is defined as the difference between the highest value and lowest value of the variable.

$$Range = Maximum\ value - Minimum\ value$$

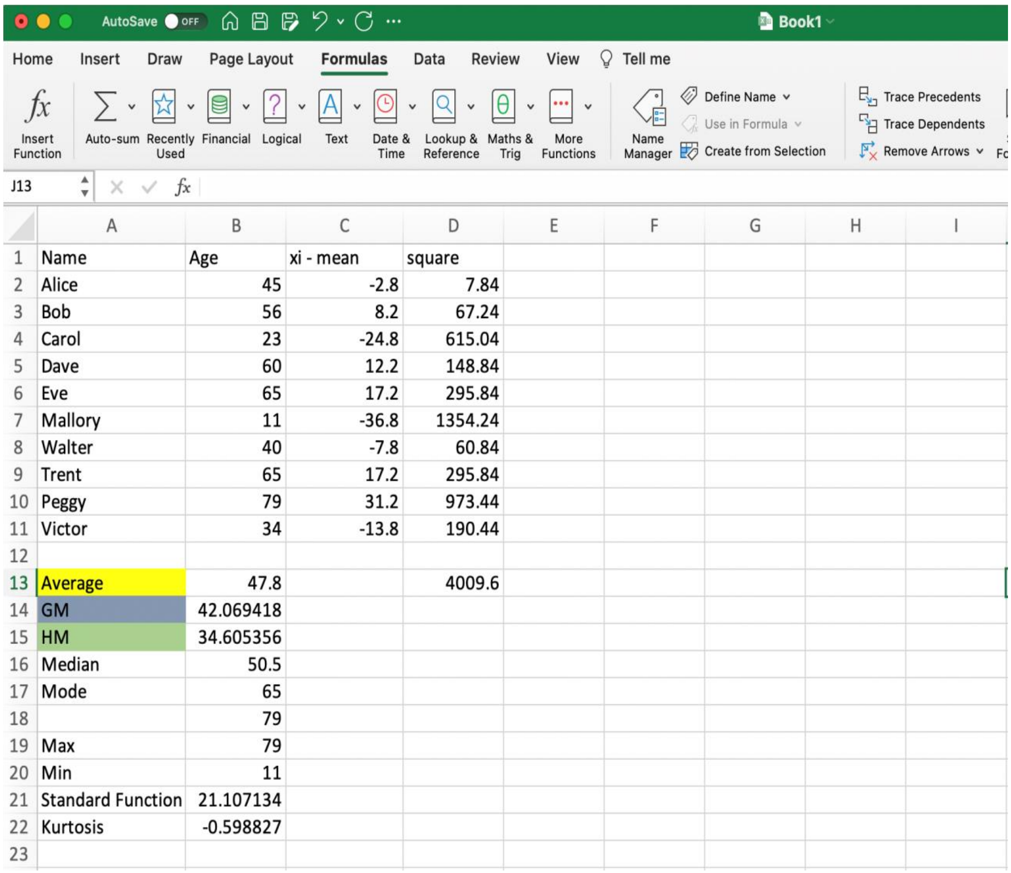
  Steps to compute range in Excel:

  Compute the maximum and minimum value among the data values. Then compute the difference between them to get the range of observations.

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘MAX’ > click ‘Insert Function’ > Enter.

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘MIN’ > click ‘Insert Function’ > Enter.

  Select a cell > write “=(specify the cell where maximum value is stored - specify the cell where minimum value is stored)” in the formula bar > Enter.



	A	B	C	D
1	Name	Age	xi - mean	square
2	Alice	45	-2.8	7.84
3	Bob	56	8.2	67.24
4	Carol	23	-24.8	615.04
5	Dave	60	12.2	148.84
6	Eve	65	17.2	295.84
7	Mallory	11	-36.8	1354.24
8	Walter	40	-7.8	60.84
9	Trent	65	17.2	295.84
10	Peggy	79	31.2	973.44
11	Victor	34	-13.8	190.44
12				
13	Average	47.8		4009.6
14	GM	42.069418		
15	HM	34.605356		
16	Median	50.5		
17	Mode	65		
18		79		
19	Max	79		
20	Min	11		
21	Standard Function	21.107134		
22	Kurtosis	-0.598827		
23				

- **Standard Deviation:** It is defined as the positive square-root of the arithmetic mean of the square of the deviations of the given observations from their arithmetic mean. It takes into consideration the magnitude of all the observations and gives the minimum value of dispersion possible. It is also known as Root Mean Square Deviation about mean.

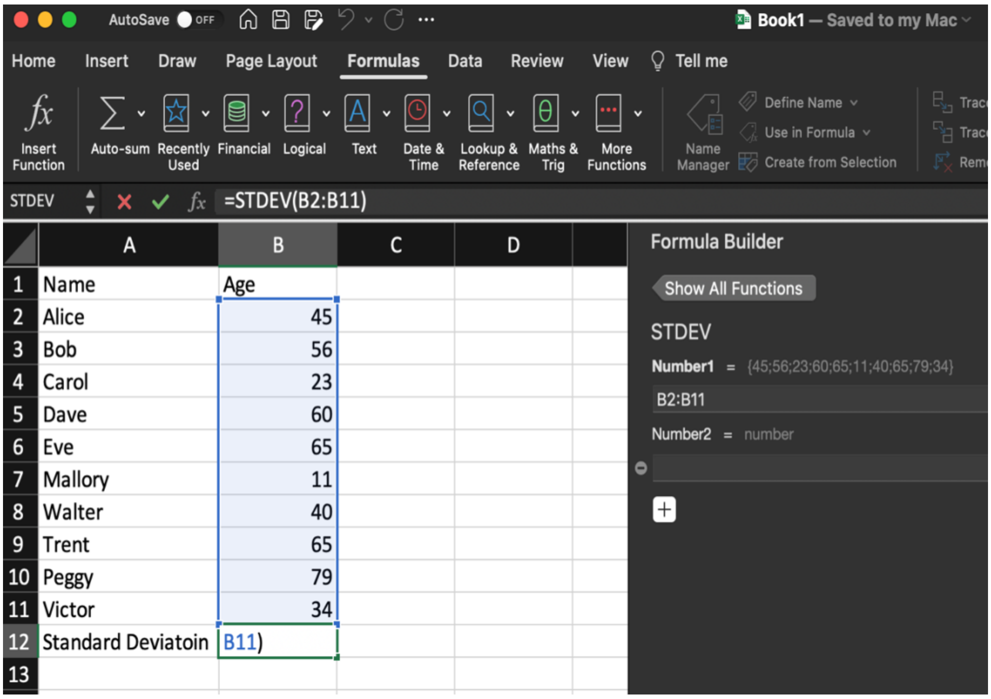
  Let x_1, x_2, …, x_n are the *n* observations in a data set. The standard deviation S.D. is given by,

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n}}$$

- **Variance:** It is defined as the square of the standard deviation. Unit of the variance is the square of the actual observations, whereas unit of the standard deviation is same as the actual observations.

  Steps to calculate Standard Deviation in excel:

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘STDEV’ > click ‘Insert Function’ > Enter.



STDEV =STDEV(B2:B11)

	A	B	C	D
1	Name	Age		
2	Alice	45		
3	Bob	56		
4	Carol	23		
5	Dave	60		
6	Eve	65		
7	Mallory	11		
8	Walter	40		
9	Trent	65		
10	Peggy	79		
11	Victor	34		
12	Standard Deviatoin	B11)		
13				

STDEV

Number1 = {45;56;23;60;65;11;40;65;79;34}

B2:B11

Number2 = number

There are 6 versions of standard deviation formula available which are as following:

STDEV.S: This formula calculates the sample standard deviation based on numeric information alone. It ignores text and logical (TRUE or FALSE) values in the spreadsheet. The denominator in this case is *(n-1)*.

STDEV.P: This formula calculates the standard deviation for an entire population based on numeric information alone. It ignores text and logical values in the spreadsheet. The denominator in this case is *n*.

STDEVA: This formula calculates the sample standard deviation of a dataset but includes text and logical values in the calculation. All FALSE values are represented by 0, and TRUE values are represented by 1.

STDEVPA: This formula calculates the standard deviation for an entire population and includes text and logical values in the calculation. Like STDEVA, all FALSE values are represented by 0, and TRUE values are represented by 1.

STDEV: This is an older version of the STDEV.S formula that Excel used to calculate sample standard deviation before 2007. It still exists for compatibility purposes. This formula acts as the same as STDEV.S

STDEVP: This is an older version of the STDEV.P formula that still exists for compatibility.

Steps to calculate Variance in excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘STDEV’ > click ‘Insert Function’ > Enter.

- **Coefficient of Variation (CV):**The Coefficient of Variation (CV) is defined as the ratio of the standard deviation to the mean, and expressed in percentages,

$$CV = \frac{Standard\ Deviation}{Mean} * 100$$

CV is calculated to have an idea about the consistency/ variability of the series. Higher the CV means the series is more variable, less stable, less uniform, and less consistent. Lesser CV indicates that the series is less variable or more stable or more uniform and more consistent.

- **Skewness and Kurtosis:** Skewness is used to detect outliers in a data set**.** It characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.A data series is said to be positively skewed if the Mean of the data series is greater than Median and is greater than Mode. On the other hand data is said to be negatively skewed if Mean < Median < Mode. Data series is said to be symmetric if Mean = Median = Mode.

$$Pearson's\ Coefficient\ of\ Skewness = \frac{(Mean - Mode)}{Standard\ Deviation}$$

Alternate formula for computing Skewness Coefficient,

$$Coefficient\ of\ Skewness = \frac{3\ (Mean - Median)}{Standard\ Deviation}$$

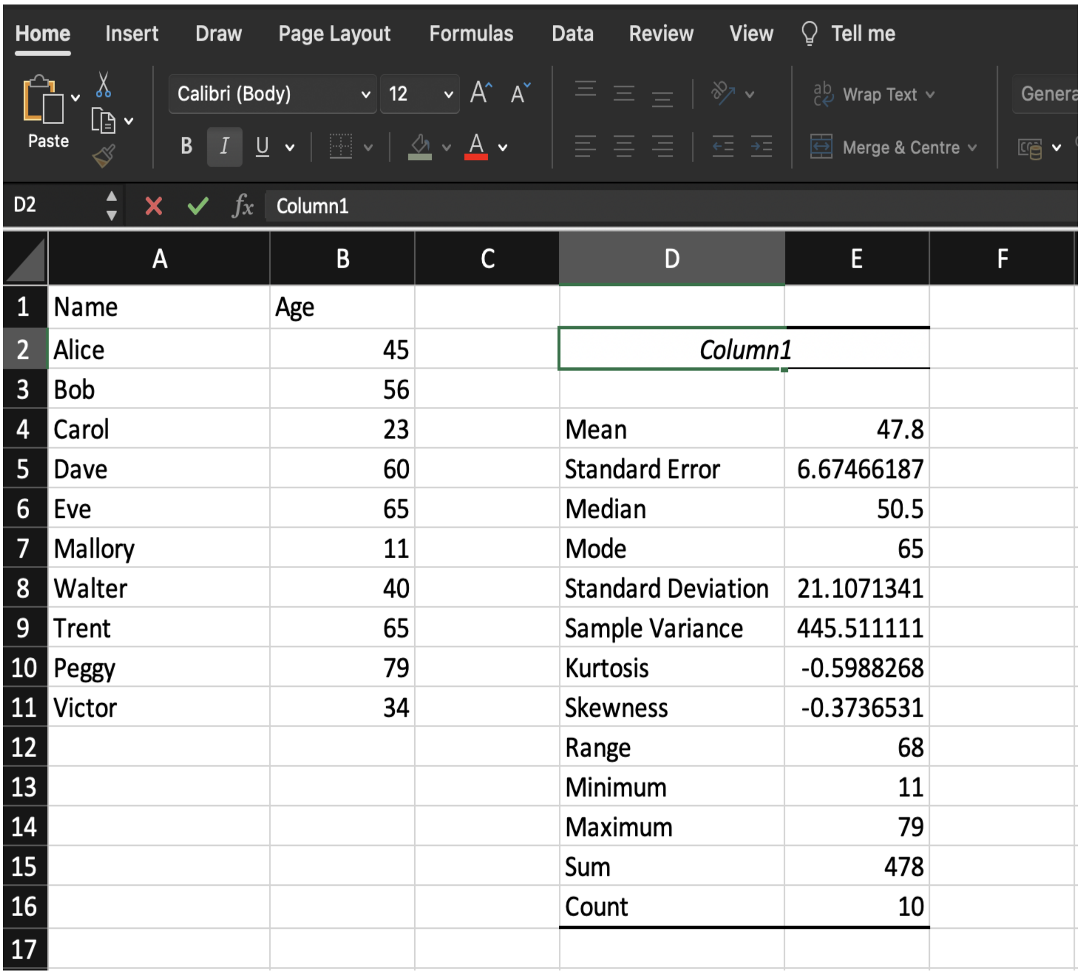
If Skewness coefficient = 0, then the distribution is normal and symmetrical.

If Skewness coefficient > 0, then the frequency distribution is positively skewed.

If Skewness coefficient < 0, then the frequency distribution is negatively skewed.

Steps to calculate Skewness Coefficient in excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘SKEW’ > click ‘Insert Function’ > Enter.



D2 | fx | Column1

	A	B	C	D	E	F
1	Name	Age				
2	Alice	45		*Column1*		
3	Bob	56				
4	Carol	23		Mean	47.8	
5	Dave	60		Standard Error	6.67466187	
6	Eve	65		Median	50.5	
7	Mallory	11		Mode	65	
8	Walter	40		Standard Deviation	21.1071341	
9	Trent	65		Sample Variance	445.511111	
10	Peggy	79		Kurtosis	-0.5988268	
11	Victor	34		Skewness	-0.3736531	
12				Range	68	
13				Minimum	11	
14				Maximum	79	
15				Sum	478	
16				Count	10	
17						

- **Kurtosis:** Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. It is the tailedness of a distribution relative to a normal distribution. Distributions with medium kurtosis (medium tails) are mesokurtic, with low kurtosis are called platykurtic, and distributions with high kurtosis are leptokurtic.

$$\text{Measure of kurtosis, } \gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Kurtosis value equals to 3.0 indicates, the data distribution is mesokurtic, for kurtosis value greater than 3.0, it is called leptokurtic and for a lesser value than 3.0 the distribution is called platykurtic.

Steps to calculate Kurtosis in excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'KURT' > click 'Insert Function' > Enter.

Excel provides an "*Analysis Tool Pak*" add-in under the *Data* tab to generate a report of the Descriptive Statistics on the desired data.

For example, we have examination scores of 10 students in a class like the following. To generate descriptive statistics for these scores, follow the steps below.

Step 1: On the Data tab, in the Analysis group, click Data Analysis.

Step 2: Select Descriptive Statistics and click OK.

Step 3: Select the range B2:B11 as the Input Range.

Step 4: Select cell C1 as the Output Range.

Step 5: Make sure Summary statistics is checked.

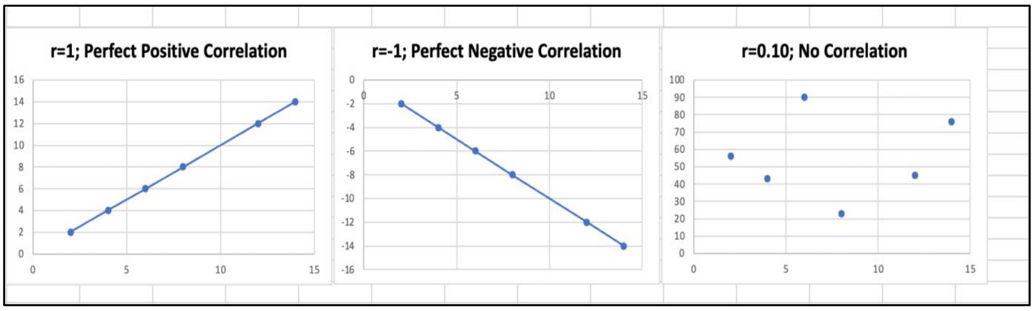
Step 6: Click ok.

2. **Correlation and Regression Analysis:** Correlation is the measurement of linear association between two variables. It is a measure that describes the strength and direction of a relationship between two variables. It is a commonly used measure in statistics, economics and social sciences for budgets, business plans etc. The correlation coefficient is used to measure the correlation between bivariate data which basically denotes the degree of linear association between two random variables.

   In statistics, there are several types of correlation measures depending on the type of data you are working with. Here, we will focus on the most common one.

   Pearson Product Moment Correlation (PPMC), popularly called as Pearson Correlation is used to evaluate linear relationships between data when a change in one variable is associated with a proportional change in the other variable.

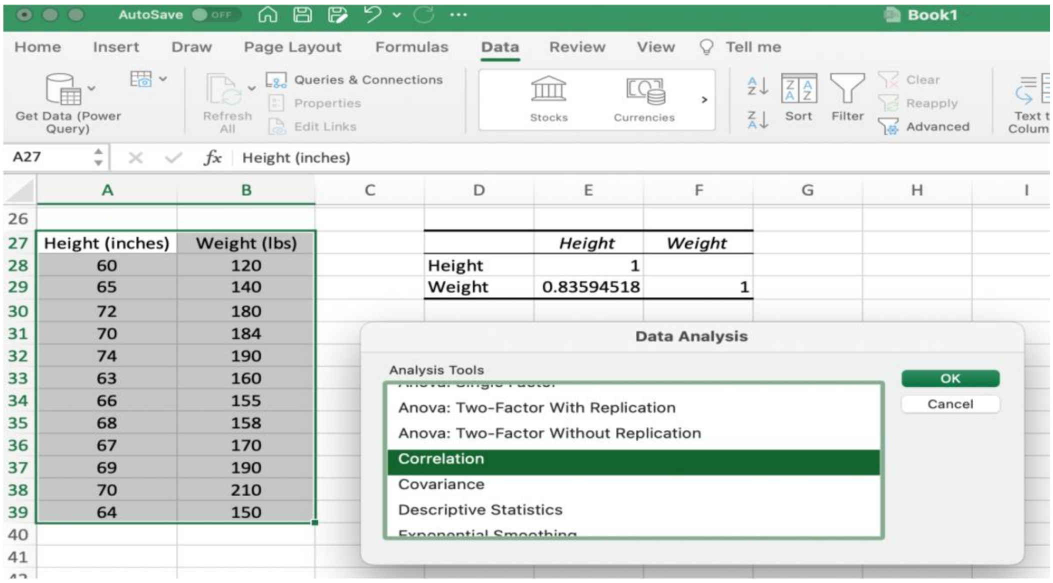
$$\textbf{Pearson Correlation Coefficient, } \boldsymbol{r} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 * \sum(y_i-\bar{y})^2}}$$



The correlation coefficient value always lies between -1 and 1 and it measures both the strength and direction of the linear relationship between the variables. Correlation coefficient of +1 means a perfect positive relationship, as value of one variable increases, value of other variable increases proportionally. Correlation coefficient value of -1 means a perfect negative relationship, with increase in the value of one variable, the other one decreases proportionally. A coefficient of 0 means no linear relationship between the two variables the data points are scattered all over the graph.

Steps to calculate Pearson Correlation Coefficient in Excel:

Select '*Data'* tab > click '*Data Analysis*' > Find Correlation from the given menus > Click ok > Select the input range > select output cell > Grouped by columns > click ok.



Height (inches)	Weight (lbs)
60	120
65	140
72	180
70	184
74	190
63	160
66	155
68	158
67	170
69	190
70	210
64	150

	Height	*Weight*
Height	1	
Weight	0.83594518	1

Regression analysis is used to estimate the relationship between two or more variables. Dependent variable is the main factor you want to study, understand, or predict. Independent variables are the factors that might influence the dependent variable. Regression analysis helps to understand how the dependent variable changes when one of the independent variables vary. Regression analysis can make it easier to predict future variable trends by analyzing the trajectory of the regression line. Simple linear regression model tries to establish a linear association between the dependent and the independent variable so that the outcome of the dependent variable can be predicted using the independent variables. The simple linear regression model uses the following equation:

$$Y = a + bX + \epsilon$$

where, Y = value of the dependent variable

X = value of the independent variable

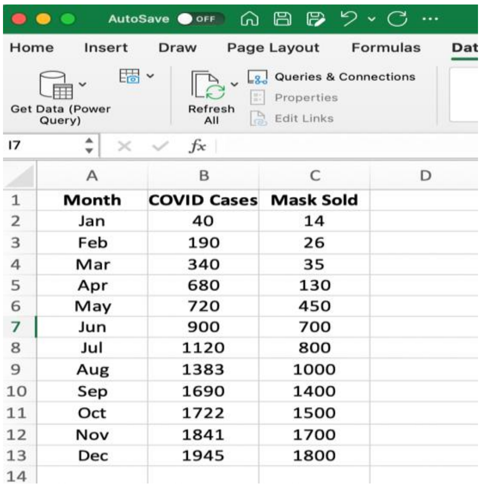
a = intercept

b = slope (regression line steepness)

ϵ = error component

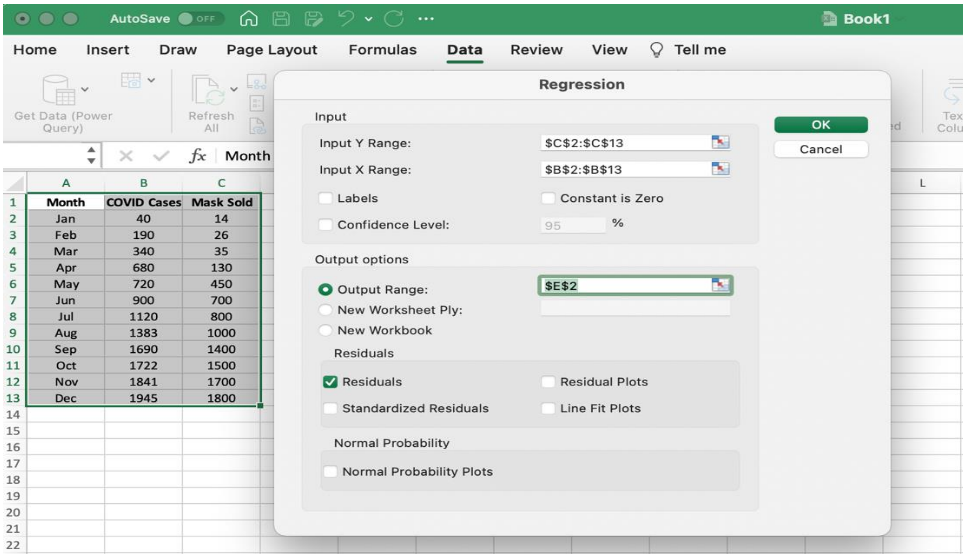
Steps to perform Regression Analysis in Excel:

Step1: Let us consider the data values for the following two variables, COVID cases and masks sold and perform a simple linear regression analysis in Excel considering number of Masks sold as the Y variable and number of COVID cases as X variable on which Y is dependent.



	A	B	C	D
1	Month	COVID Cases	Mask Sold	
2	Jan	40	14	
3	Feb	190	26	
4	Mar	340	35	
5	Apr	680	130	
6	May	720	450	
7	Jun	900	700	
8	Jul	1120	800	
9	Aug	1383	1000	
10	Sep	1690	1400	
11	Oct	1722	1500	
12	Nov	1841	1700	
13	Dec	1945	1800	
14				

Step2: Click on the ‘Data’ tab > Data Analysis > Select ‘Regression’ >click ‘Ok’.



	A	B	C
1	Month	COVID Cases	Mask Sold
2	Jan	40	14
3	Feb	190	26
4	Mar	340	35
5	Apr	680	130
6	May	720	450
7	Jun	900	700
8	Jul	1120	800
9	Aug	1383	1000
10	Sep	1690	1400
11	Oct	1722	1500
12	Nov	1841	1700
13	Dec	1945	1800

Regression

Input

Input Y Range: $C$2:$C$13

Input X Range: $B$2:$B$13

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $E$2

○ New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

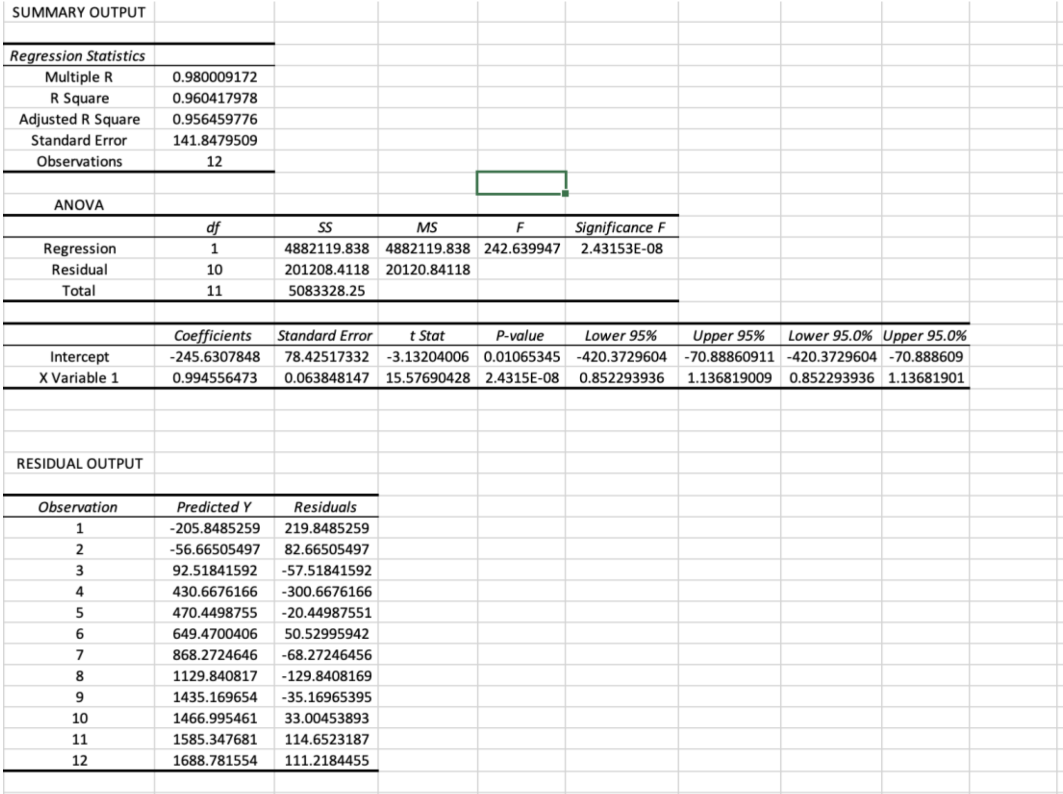
Normal Probability

□ Normal Probability Plots

Step3: In the Regression dialog box select the Input Y Range, which is our dependent variable. In this case it is (C2:C13). Then select the Input X Range, independent variable. In this example, it is the number of COVID cases (B2:B13). Select the desired output range, here E2.

Click ok.

You get the following Output:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980009172
R Square	0.960417978
Adjusted R Square	0.956459776
Standard Error	141.8479509
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4882119.838	4882119.838	242.639947	2.43153E-08
Residual	10	201208.4118	20120.84118		
Total	11	5083328.25			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-245.6307848	78.42517332	-3.13204006	0.01065345	-420.3729604	-70.88860911	-420.3729604	-70.888609
X Variable 1	0.994556473	0.063848147	15.57690428	2.4315E-08	0.852293936	1.136819009	0.852293936	1.13681901

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	-205.8485259	219.8485259
2	-56.66505497	82.66505497
3	92.51841592	-57.51841592
4	430.6676166	-300.6676166
5	470.4498755	-20.44987551
6	649.4700406	50.52995942
7	868.2724646	-68.27246456
8	1129.840817	-129.8408169
9	1435.169654	-35.16965395
10	1466.995461	33.00453893
11	1585.347681	114.6523187
12	1688.781554	111.2184455

- **Interpreting the Out putof Regression Analysis:**
  SUMMARY STATISTICS

  Multiple R is the value of the Correlation Coefficient that measures the strength of a linear relationship between two variables. The larger the absolute value, the stronger the relationship.

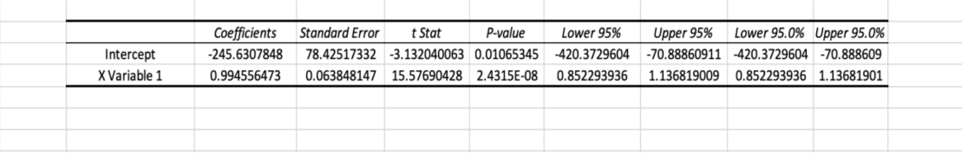
  R Square gives the Coefficient of Determination, which is used as an indicator of the goodness of fit. It shows how many points fall on the regression line. The R2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared deviations of the original data from the mean. In this example, R2 is 0.96, which is very good. It means that 96% of our values fit the regression analysis model. In other words, 96% of the dependent variables (y-values) are explained by the independent variables (x-values). Generally, R Squared of 95% or more is considered a good fit.

Adjusted R Square gives the R square adjusted for the number of independent variables in the model. For multiple regression analysis, adjusted R square value is used instead of R square.

Standard Error is another goodness-of-fit measure that shows the precision of the fitted regression model. The smaller the number, the more certain one can be about the regression equation. It is an absolute measure that shows the average distance that the data points fall from the regression line.

Observations simply provides the total number of observations used to fir the model.

COEFFICIENTS

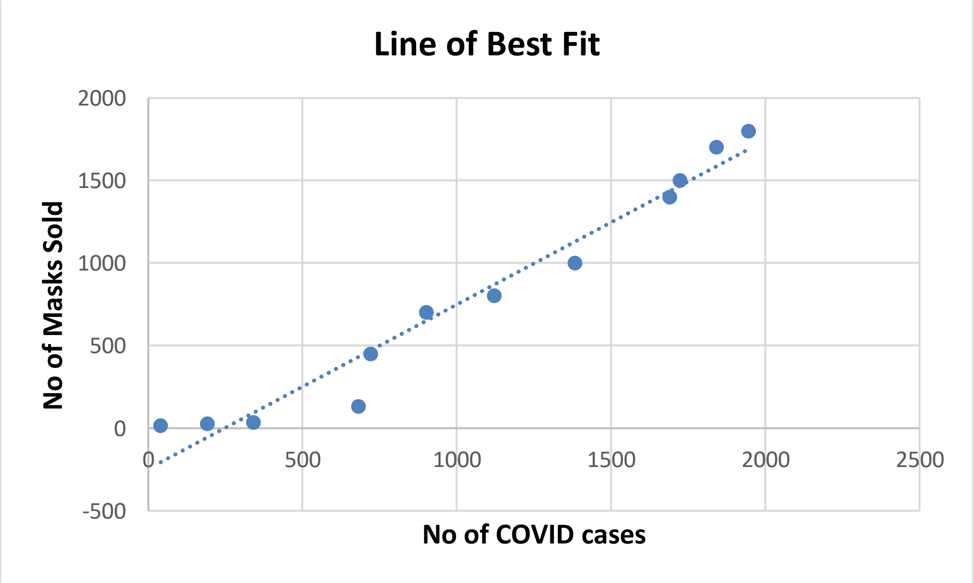


	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-245.6307848	78.42517332	-3.132040063	0.01065345	-420.3729604	-70.88860911	-420.3729604	-70.888609
X Variable 1	0.994556473	0.063848147	15.57690428	2.4315E-08	0.852293936	1.136819009	0.852293936	1.13681901

Linear regression equation fitted was, Y = b*X + a

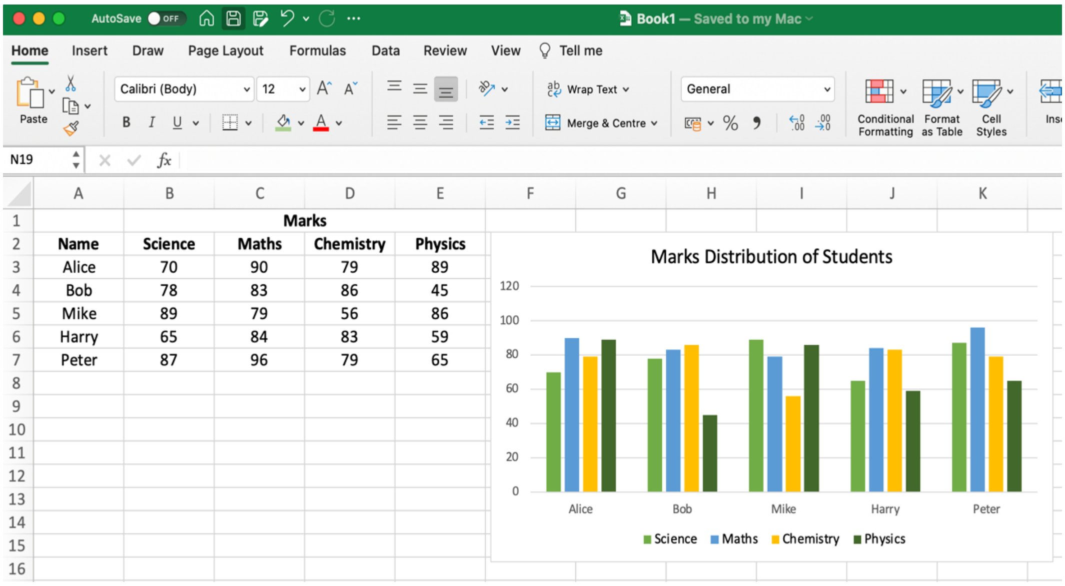
Here, Y = Mask sold; X = COVID cases; b = 0.99; a = -245.63

Therefore, 0.99 * 190 – 245.63 = -57.53



3. Create Charts/ Graphs in MS Excel:

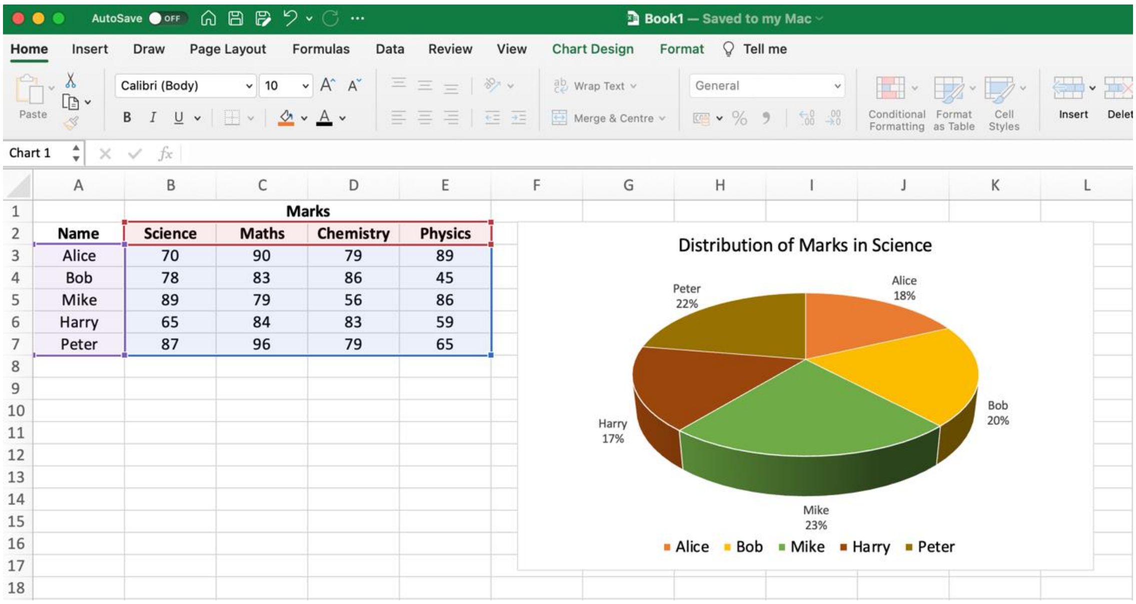
Line Diagram: Select the data for which you want to plot the graph. Click 'Insert' tab > go to insert column chart > pick any chart of your preference. Excel will create the graphical representation as following.



	Marks			
Name	Science	Maths	Chemistry	Physics
Alice	70	90	79	89
Bob	78	83	86	45
Mike	89	79	56	86
Harry	65	84	83	59
Peter	87	96	79	65

Pie chart: Pie chart represents the data in slices of a circle. Each slice represents the percentage contribution of each data section among the sum of individual data values.

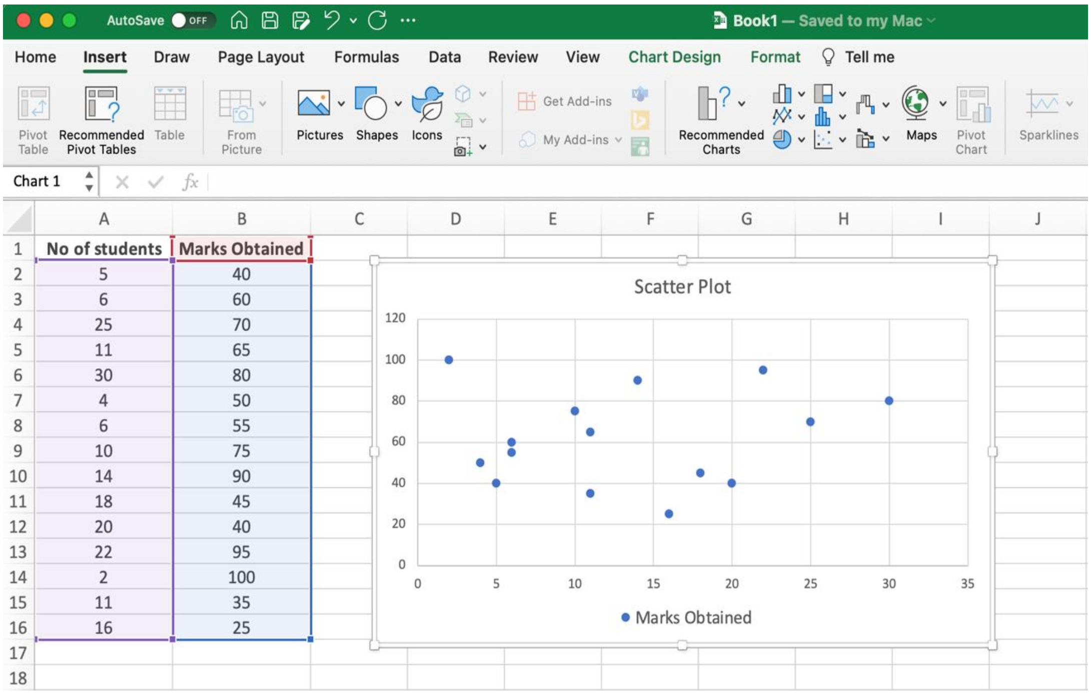
Select the data for which you want to plot the pie chart. Click insert tab > go to insert pie or doughnut chart > pick any chart of your preference. Excel will create the graphical representation as following:



	Marks			
Name	Science	Maths	Chemistry	Physics
Alice	70	90	79	89
Bob	78	83	86	45
Mike	89	79	56	86
Harry	65	84	83	59
Peter	87	96	79	65

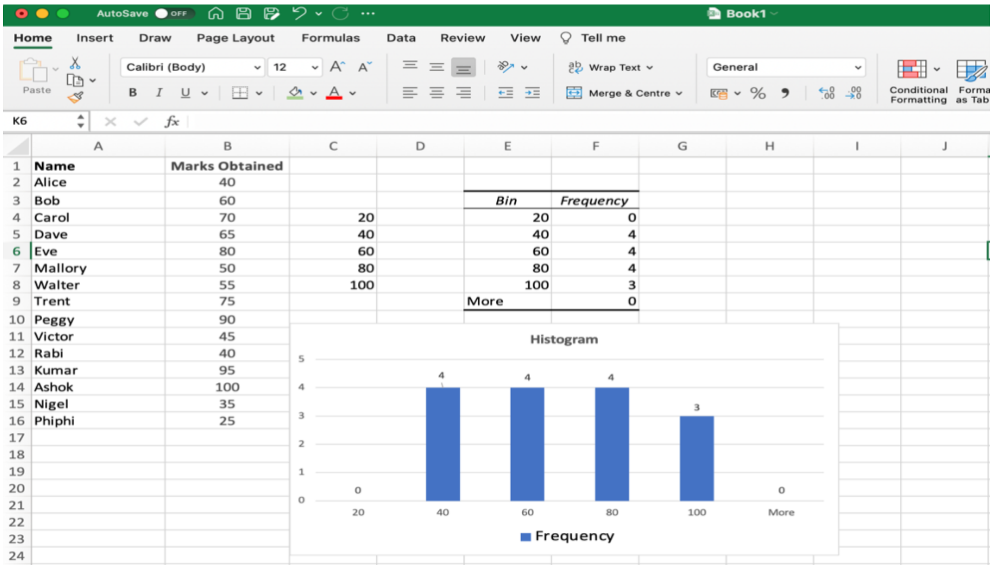
Scatter Diagram: Scatter charts are specifically used to show how one variable is related to another. There are seven scatter chart options: scatter, scatter with smooth lines and markers, scatter with smooth lines, scatter with straight lines and markers, scatter with straight lines, bubble, and 3-D bubble. For plotting a scatter chart, one needs data points for two or more variables.

Select the data> click insert tab > go to X Y Scatter chart > pick any chart of your preference. Excel will create the graphical representation as following:



	A	B
1	No of students	Marks Obtained
2	5	40
3	6	60
4	25	70
5	11	65
6	30	80
7	4	50
8	6	55
9	10	75
10	14	90
11	18	45
12	20	40
13	22	95
14	2	100
15	11	35
16	16	25
17		
18		

Histogram: Select data >click on data tab > select data Analysis >click histogram > select



	A	B	C	D	E	F
1	Name	Marks Obtained				
2	Alice	40				
3	Bob	60			*Bin*	*Frequency*
4	Carol	70	20		20	0
5	Dave	65	40		40	4
6	Eve	80	60		60	4
7	Mallory	50	80		80	4
8	Walter	55	100		100	3
9	Trent	75			More	0
10	Peggy	90				
11	Victor	45				
12	Rabi	40				
13	Kumar	95				
14	Ashok	100				
15	Nigel	35				
16	Phiphi	25				

input range (B2:B16)> select bin (class intervals, here it is C4:C8) > check Chart Output > click ok. Excel will produce the frequency table against the specified bin value and also will create a histogram diagram like following.

4. **Inferential Statistics:**

Inferential statistics is used for estimating the population data by analysing the samples obtained from it. It helps in making generalizations about the population by using different analytical tests and tools. Various sampling techniques are usedto select random samples that will represent the population accurately. Some of the important methods are simple random sampling, stratified sampling, cluster sampling, and systematic sampling techniques.

Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. In inferential statistics, a statistic is taken from the sample data (e.g., sample mean) that used to make inferences about the population parameter (e.g., the population mean). One sample t-test is the most commonly used one and sets a basic understanding of all other kinds of hypothesis testing methods.

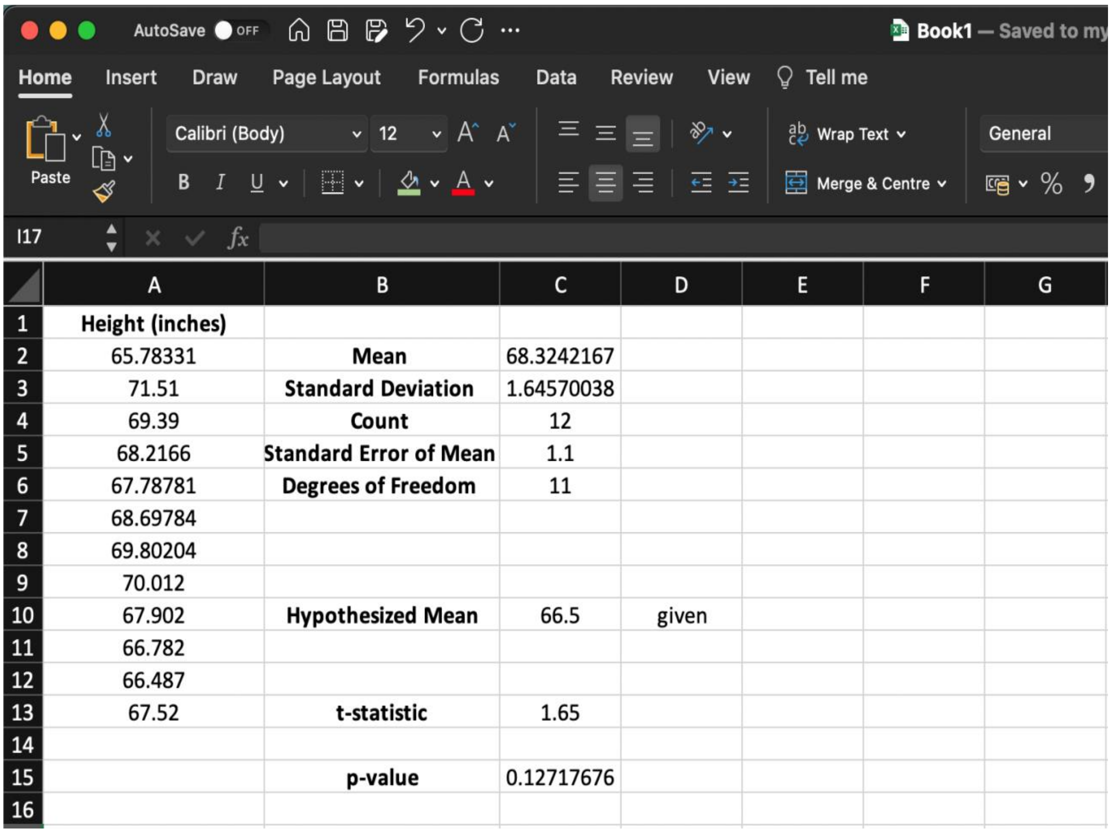
One sample *t*-test:

The one-sample t test compares a given sample mean $\bar{X}$ to a known or hypothesized value of the population mean $\mu_0$provided the population standard deviation σ is unknown. Excel does not have a built-in one-sample t test. However, the use of Excel functions and formulas makes the computations quite simple. The value of *t*-statistic can be calculated from the given formula:

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{x}}}$$

where, $\bar{X}$ is the sample mean, μ_0 is the known or hypothesized population mean and $s_{\bar{x}}$ isthe standard error of mean.To calculate the *t*-statistic in excel we need to first find the following values.

Consider a sample of 12 young female adults, we have the measurement of their heights in inches. Let us assume the national average height of 18-year-old girls is 66.5 inches. We want to perform a one-sample T-test in Excel to determine if there is any significant difference between the heights of the sample data compared with the national average height (66.5 inches).



	A	B	C	D	E	F	G
1	Height (inches)						
2	65.78331	Mean	68.3242167				
3	71.51	Standard Deviation	1.64570038				
4	69.39	Count	12				
5	68.2166	Standard Error of Mean	1.1				
6	67.78781	Degrees of Freedom	11				
7	68.69784						
8	69.80204						
9	70.012						
10	67.902	Hypothesized Mean	66.5	given			
11	66.782						
12	66.487						
13	67.52	t-statistic	1.65				
14							
15		p-value	0.12717676				
16							

The null hypothesis and alternative hypothesis for this test are:

Null hypothesis: There is no significant difference between the heights of the sample, compared with the national average.

Alternative hypothesis: There is significant difference between the heights of the sample, compared with the national average.

First of all, compute mean, standard deviation, standard error, degrees of freedom to calculate the value of the *t*-statistic as shown in the above screenshot then in an empty cell, enter =TDIST (t, df, tails) to compute the p-value.

t – the cell containing the t-statistic

df – The cell containing the degrees of freedom.

tails –1if you want to perform a one-tailed analysis, or 2 if you want to do a two-tailed analysis.p-value for this example is 0.127.

Let us assume alpha level is set at 0.05, then since the p-value is above the alpha level, we will accept the null hypothesis and reject the alternative hypothesis.In other words, there is no significant difference between the heights of the sample, compared with the national average.

TESTS OF SIGNIFICANCE AND NON-PARAMETRIC TEST

Rajeev Ranjan Kumar
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
Rajeev.kumar4@icar.gov.in

In the realm of statistics, the test of significance, also known as hypothesis testing, is a powerful tool used to make informed decisions about population parameters based on sample data. It enables researchers and analysts to assess the validity of assumptions, draw conclusions, and determine the level of confidence in their findings.

The fundamental idea behind the test of significance is to evaluate whether the observed data is strong enough to support or reject a particular hypothesis about a population characteristic. This hypothesis is typically formulated in terms of a null hypothesis (H_0), which assumes no significant difference or relationship, and an alternative hypothesis (H_1), which posits the existence of a meaningful difference or relationship.

To conduct a test of significance, a sample is collected from the population of interest, and relevant statistical techniques are employed to analyze the data. The results are then used to evaluate the likelihood of observing the sample data under the assumption that the null hypothesis is true. If the observed data is highly improbable under this assumption, it provides evidence to reject the null hypothesis in favour of the alternative hypothesis.

The test of significance involves determining a test statistic, which summarizes the data and allows for comparison against a theoretical distribution. The choice of the appropriate test statistic depends on the nature of the research question and the type of data being analyzed. Commonly used test statistics include the z-score, t-statistic, chi-square statistic, and F-statistic, among others.

1.Types of Hypotheses

In scientific research, a hypothesis is a proposed explanation or prediction about a phenomenon or relationship between variables. Hypotheses play a crucial role in guiding research and formulating testable statements that can be supported or refuted by empirical evidence. Depending on the nature of the research question and the specific objectives of the study, different types of hypotheses can be formulated. Here are some common types of hypotheses:

Null Hypothesis (H_0): The null hypothesis represents the absence of an effect, relationship, or difference between variables. It assumes that there is no statistically significant relationship or change in the population being studied. Researchers generally aim to reject the null hypothesis in favour of an alternative hypothesis. For example, the null hypothesis could state that there is no difference in test scores between two groups of students.

Alternative Hypothesis (H_1): The alternative hypothesis is the opposite of the null hypothesis. It suggests that there is a significant effect, relationship, or difference between variables in the population. Researchers seek to gather evidence to support the alternative hypothesis. Building upon the previous example, the alternative hypothesis could state that there is a difference in test scores between the two groups of students.

Directional Hypothesis: A directional hypothesis predicts the specific direction of the relationship or difference between variables. It specifies whether the effect will be positive or negative. For instance, a directional hypothesis may state that Group A will have higher test scores than Group B or that an increase in temperature will lead to a decrease in plant growth. Directional hypotheses are often used when previous research or theoretical considerations provide a basis for predicting the direction of the effect.

Non-Directional Hypothesis: Also known as a two-tailed hypothesis, a non-directional hypothesis does not predict a specific direction of the relationship or difference. It simply states that there is a significant difference or relationship between variables without specifying the direction. Researchers use non-directional hypotheses when they do not have a clear theoretical basis or prior evidence to suggest a specific direction. For example, a non-directional hypothesis may state that there is a difference in test scores between two groups of students, without specifying which group will perform better.

Composite Hypothesis: A composite hypothesis consists of multiple statements or conditions. It encompasses more than one possibility and allows for different outcomes. Composite hypotheses are often used when there are multiple factors or variables involved in the research question. For instance, a composite hypothesis could state that the effect of a particular treatment on patient outcomes varies depending on age, gender, and socioeconomic status.

Simple Hypothesis: In contrast to composite hypotheses, simple hypotheses involve a single statement or condition. They are straightforward and make specific predictions about a single variable or relationship. Simple hypotheses are commonly used when the research question focuses on a single factor or variable. For example, a simple hypothesis could state that there is a positive correlation between study time and exam scores.

2. Types of Errors

Errors can occur due to various sources of uncertainty and can impact the validity and reliability of research findings. Understanding the types of errors is essential for researchers and analysts to properly interpret and draw accurate conclusions from their data. Here are the two primary types of errors in statistics:

(A) Type I Error

Type I error, also known as a false positive, occurs when the null hypothesis (H_0) is mistakenly rejected, indicating the presence of a significant effect or relationship when, in fact, none exists in the population. It represents the probability of observing a statistically significant result due to random chance alone. Type I error is typically denoted by the symbol α (alpha) and is related to the significance level chosen for the hypothesis test.

For example, let's say a researcher conducts a study to determine if a new drug is effective in reducing blood pressure. The null hypothesis states that the drug has no effect. If the researcher rejects the null hypothesis and concludes that the drug is effective when it is actually not, it would be a Type I error. The researcher would have falsely claimed a significant effect.

The significance level chosen for the hypothesis test determines the threshold at which a Type I error is considered acceptable. A lower significance level (e.g., $\alpha = 0.05$) reduces the risk of Type I error but increases the chance of Type II error.

(B) Type II Error:

Type II error, also known as a false negative, occurs when the null hypothesis (H_0) is incorrectly accepted, implying no significant effect or relationship, even when there is one in the population. It represents the failure to detect a true effect or relationship. Type II error is denoted by the symbol β (beta) and is related to the statistical power of the test.

Building upon the previous example, if the researcher fails to reject the null hypothesis and concludes that the drug is not effective, even though it is, it would be a Type II error. The researcher would have missed detecting a real effect.

Type II error is influenced by factors such as the sample size, effect size, variability in the data, and the chosen significance level. To minimize the risk of Type II error, researchers often aim to maximize the statistical power of their study by using larger sample sizes, employing more sensitive measurement techniques, or increasing the significance level.

It's important to note that Type I and Type II errors are inversely related: reducing one type of error increases the likelihood of the other. Researchers need to strike a balance between these two types of errors based on the consequences of each in the specific research context.

(3)Level of Significance in Statistics:

In statistical hypothesis testing, the level of significance, often denoted by the symbol α (alpha), is a predetermined threshold that helps researchers make decisions about the validity of their results. It represents the maximum allowable probability of making a Type I error (rejecting the null hypothesis when it is actually true). The level of significance plays a crucial role in determining the critical region and the acceptance or rejection of the null hypothesis.

The most commonly used level of significance in many fields of research is 0.05 (or 5%). This means that if the calculated probability (p-value) of obtaining the observed data under the null hypothesis is equal to or less than 0.05, the null hypothesis is rejected in favour of the alternative hypothesis. In other words, researchers conclude that there is sufficient evidence to suggest that a relationship, effect, or difference exists in the population being studied. However, the choice of the level of significance is not arbitrary and should be determined based on the specific research question, the consequences of Type I and Type II errors, and the desired level of confidence. Commonly used levels of significance include 0.01 (1%) and 0.10 (10%), depending on the context and the stringency of the decision-making process.

A lower level of significance (e.g., 0.01) reduces the risk of Type I error, providing a more conservative approach to hypothesis testing. It requires stronger evidence to reject the null hypothesis and provides a higher level of confidence in the conclusions drawn from the data. On the other hand, a higher level of significance (e.g., 0.10) increases the risk of Type I error, making it easier to reject the null hypothesis. This approach is less conservative and may be appropriate when the consequences of Type II error are more severe or when exploratory analysis is conducted. It's important to note that the level of significance does not directly indicate the magnitude or practical importance of the observed effect. It solely reflects the strength of evidence against the null hypothesis. Therefore, researchers need to carefully interpret the results in the context of the specific research question and consider the practical implications of their findings.

(4) P-value

In statistical hypothesis testing, the p-value is a measure that helps researchers assess the strength of evidence against the null hypothesis (H_0) and make informed decisions about its rejection or acceptance. The p-value represents the probability of obtaining the observed data, or more extreme data, if the null hypothesis were true. The calculation of the p-value involves comparing the observed test statistic (e.g., t-statistic, z-score, chi-square statistic) with the distribution of the test statistic under the assumption that the null hypothesis is true. The p-value provides a quantitative measure of the likelihood of observing the data under the null hypothesis.

Interpreting the p-value is based on a chosen level of significance (α) that represents the threshold for rejecting the null hypothesis. If the p-value is smaller than the chosen level of significance, typically 0.05 (or 5%), it is considered statistically significant, and the null hypothesis is rejected. This indicates that the observed data is unlikely to occur by random chance alone and provides evidence in favour of the alternative hypothesis (H_1).On the other hand, if the p-value is larger than the chosen level of significance, the null hypothesis is not rejected. This suggests that the observed data is reasonably likely to occur by random chance, and there is insufficient evidence to support the alternative hypothesis. It's important to note that failing to reject the null hypothesis does not prove its truthfulness; it simply suggests that there is not enough evidence to support the alternative hypothesis.

(5) Critical Region

The critical region, also known as the rejection region, is a defined range of values or outcomes of a test statistic that leads to the rejection of the null hypothesis (H_0). The critical region is determined based on the chosen level of significance (α) and the distribution of the test statistic under the assumption that the null hypothesis is true.

The critical region represents the extreme or unlikely values of the test statistic that would cast doubt on the validity of the null hypothesis. If the observed test statistic falls within the critical region, it provides evidence against the null hypothesis and leads to its rejection in favour of the alternative hypothesis (H_1).

To determine the critical region, researchers specify the desired level of significance (α) before conducting the hypothesis test. The level of significance represents the maximum allowable probability of making a Type I error (rejecting the null hypothesis when it is actually true). The critical region is then defined such that the probability of observing a test statistic within that region, assuming the null hypothesis is true, is equal to or less than the chosen level of significance (α).The critical region is determined based on the specific distribution associated with the test statistic being used and the nature of the research question. For example, in a t-test, the critical region is defined by critical values obtained from the t-distribution, while in a z-test, it is determined by the critical values of the standard normal distribution. The critical region is often represented graphically on a probability distribution, showing the area in the tail(s) of the distribution associated with rejection of the null hypothesis. The critical values divide the distribution into the critical region (rejection region) and the non-critical region (non-rejection region).

When the calculated test statistic falls within the critical region, the null hypothesis is rejected, indicating that the observed data is unlikely to occur by random chance alone and supports the alternative hypothesis. Conversely, if the test statistic falls within the non-critical region, the null hypothesis is not rejected, suggesting that the observed data is reasonably likely to occur by random chance, and there is insufficient evidence to support the alternative hypothesis.It's important to note that the size and location of the critical region are influenced by the chosen level of significance. A smaller level of significance (e.g., $\alpha = 0.01$) results in a more stringent critical region, making it more difficult to reject the null hypothesis. On the other hand, a larger level of significance (e.g., $\alpha = 0.10$) widens the critical region, making it easier to reject the null hypothesis.

(6)One-Tailed and Two-Tailed Tests in Statistics:

In statistical hypothesis testing, researchers can choose between one-tailed and two-tailed tests based on the specific research question and the directionality of the effect being investigated. These tests differ in the way they assess the evidence against the null hypothesis (H_0) and the corresponding critical region.

One-Tailed Test

In a one-tailed (or one-sided) test, the alternative hypothesis (H_1) specifies the direction of the effect or difference between variables. It predicts that the observed data will be either significantly greater or significantly less than what would be expected under the null hypothesis. Therefore, the critical region is located entirely in one tail of the distribution of the test statistic.

The one-tailed test is appropriate when there is a clear theoretical or practical basis for predicting the direction of the effect. It allows researchers to focus their analysis on that specific direction and increases the power to detect the effect in that direction. One-tailed tests are often used in situations where previous research or knowledge suggests a particular directionality. For example, in a study investigating whether a new treatment improves test scores, the one-tailed test would focus on determining if the treatment leads to significantly higher test scores, neglecting the possibility of significantly lower scores.

Two-Tailed Test

In a two-tailed (or two-sided) test, the alternative hypothesis does not specify a particular direction of the effect. It predicts that the observed data will be significantly different from what would be expected under the null hypothesis, without specifying whether it will be greater or smaller. Therefore, the critical region is divided into two equal tails, one in each direction of the distribution of the test statistic.

The two-tailed test is appropriate when there is no prior expectation or theoretical basis to predict the direction of the effect. It provides a more conservative approach to hypothesis testing, as it requires stronger evidence to reject the null hypothesis compared to a one-tailed test. For example, in a study investigating whether a new teaching method affects test scores, the two-tailed test would examine if the teaching method leads to significantly different test scores, without specifying whether the scores will be higher or lower. The choice between one-tailed and two-tailed tests should be based on careful consideration of the research question, previous knowledge, and theoretical expectations. While a one-tailed test increases the power to detect an effect in a specific direction, it may miss effects in the opposite direction. A two-tailed test is more conservative but captures effects in both directions.

7. Non-parametric Test

In statistics, non-parametric tests, also known as distribution-free tests, are statistical methods used to make inferences and draw conclusions about populations or samples without assuming a specific probability distribution. Unlike parametric tests, which rely on assumptions about the underlying data distribution, non-parametric tests make fewer assumptions and are more robust to violations of distributional assumptions.

Non-parametric tests are often used when the data does not meet the assumptions required for parametric tests, such as when the data is skewed, have outliers, or when the sample size is small. These tests are also useful when dealing with ordinal or nominal data, as they do not require interval or ratio level measurements. Some common non-parametric tests include:

1. Mann-Whitney U test: This test is used to compare the medians of two independent groups. It is a non-parametric alternative to the independent samples t-test.

2. Wilcoxon signed-rank test: This test is used to compare the medians of two related or paired samples. It is a non-parametric alternative to the paired samples t-test.

3. Kruskal-Wallis test: This test is used to compare the medians of three or more independent groups. It is a non-parametric alternative to the one-way analysis of variance (ANOVA).

4. Friedman test: This test is used to compare the medians of three or more related groups. It is a non-parametric alternative to the repeated measures ANOVA.

5. Spearman's rank correlation coefficient: This test is used to assess the strength and direction of the monotonic relationship between two variables. It is a non-parametric alternative to Pearson's correlation coefficient.

Non-parametric tests rely on ranks or other orderings of the data rather than the actual numerical values. They use statistical techniques that compare the distributions of the data or evaluate the degree of association between variables without assuming a specific probability distribution. Advantages of non-parametric tests include their robustness to outliers and their ability to handle data that does not meet the assumptions of parametric tests. However, they generally have less statistical power than parametric tests when the data does conform to the assumptions of the parametric tests. Non-parametric tests are widely used in various fields, including psychology, sociology, biology, medicine, and environmental science, where the assumptions of parametric tests may not be met or when dealing with categorical or ranked data.

References

Agrawal, B.L. (2006) Basic Statistics, New Age International, India

Gupta, S.C. and Kapoor, V.K. (2020) Fundamentals of Mathematical Statistics, Sultan Chand and Sons, India

MULTIVARIATE STATISTICAL TECHNIQUES

Prabina Kumar Meher, Atmakuri Ramakrishna Rao
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Prabina.Meher@icar.gov.in

Multivariate data consist of observations on several different variables for a number of individuals or subjects. Data of this type arise in all the branches of science, ranging from psychology to biology, and methods of analyzing multivariate data constitute an increasingly important area of statistics. Indeed, the vast majority of data in forestry is multivariate and proper handling of such data is highly essential. Principal components analysis (PCA) and Factor analysis (FA) are multivariate techniques applied to a single set of variables to discover which sets of variables in the set form coherent subsets that are relatively independent of one another. The details of PCA and FA are discussed as below.

Principal Components Analysis

Most of the times the variables under study are highly correlated and as such they are effectively "saying the same thing". To examine the relationships among a set of *p* correlated variables, it may be useful to transform the original set of variables to a new set of uncorrelated variables called *principal components*. These new variables are linear combinations of original variables and are derived in decreasing order of importance so that, for example, the first principal component accounts for as much as possible of the variation in the original data.

Let $x_1, x_2, x_3, . . . , x_p$ are variables under study, then first principal component may be defined as

$$z_1 = a_{11} x_1 + a_{12} x_2 + ... + a_{1p} x_p$$

such that variance of z_1 is as large as possible subject to the condition that

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$$

This constraint is introduced because if this is not done, then Var(z_1) can be increased simply by multiplying any a_{1j}s by a constant factor

The second principal component is defined as

$$z_2 = a_{21} x_1 + a_{22} x_2 + ... + a_{2p} x_p$$

Such that Var(z_2) is as large as possible next to Var(z_1)subject to the constraint that

$a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$ and $\text{cov}(z_1, z_2) = 0$ and so on.

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial p variables in subsequent analysis, thus, reducing the effective dimensionality of the problem. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result. However, Principal Component Analysis is more of a means to an end rather than an end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique which does not require user to specify the statistical model or assumption about distribution of original varieties. It may also be mentioned that principal components are artificial variables and often it is not possible to assign physical meaning to them. Further, since Principal Component Analysis transforms original set of variables to new set of uncorrelated variables, it is worth stressing that if original variables are uncorrelated, then there is no point in carrying out principal component analysis.

Computation of principal component

Let us consider the following data on average minimum temperature (x_1), average relative humidity at 8 hrs. (x_2), average relative humidity at 14 hrs. (x_3) and total rainfall in cm. (x_4) pertaining to Raipur district from 1970 to 1986 for kharif season from 21st May to 7th Oct.

X_1	x_2	x_3	x_4
25.0	86	66	186.49
24.9	84	66	124.34
25.4	77	55	98.79
24.4	82	62	118.88
22.9	79	53	71.88
7.7	86	60	111.96
25.1	82	58	99.74
24.9	83	63	115.20
24.9	82	63	100.16
24.9	78	56	62.38
24.3	85	67	154.40
24.6	79	61	112.71
24.3	81	58	79.63
24.6	81	61	125.59
24.1	85	64	99.87

	24.5	84	63	143.56
	24.0	81	61	114.97
Mean	23.56	82.06	61.00	112.97
S.D.	4.13	2.75	3.97	30.06

with the variance co-variance matrix.

$$\Sigma = \begin{bmatrix} 17.02 & -4.12 & 1.54 & 5.14 \\ & 7.56 & 8.50 & 54.82 \\ & & 15.75 & 92.95 \\ & & & 903.87 \end{bmatrix}$$

Find the eigen values and eigen vectors of the above matrix. Arrange the eigen values in decreasing order. Let the eigen values in decreasing order and corresponding eigen vectors are

$\lambda_1 = 916.902 \quad a_1 = (0.006, \; 0.061, \; 0.103, \; 0.993)$

$\lambda_2 = 18.375 \quad a_2 = (0.955, \; -0.296, \; 0.011, \; 0.012)$

$\lambda_3 = 7.87 \quad a_3 = (0.141, \; 0.485, \; 0.855, \; -0.119)$

$\lambda_4 = 1.056 \quad a_4 = (0.260, \; 0.820, \; -0.509, \; 0.001)$

The principal components for this data will be

$$z_1 = 0.006\, x_1 + 0.061\, x_2 + 0.103\, x_3 + 0.993\, x_4$$

$$z_2 = 0.955\, x_1 - 0.296\, x_2 + 0.011\, x_3 + 0.012\, x_4$$

$$z_3 = 0.141\, x_1 + 0.485\, x_2 + 0.855\, x_3 - 0.119\, x_4$$

$$z_4 = 0.26\, x_1 + 0.82\, x_2 - 0.509\, x_3 + 0.001\, x_4$$

The variance of principal components will be eigen values i.e.

$\text{Var}(z_1) = 916.902$, $\text{Var}(z_2) = 18.375$, $\text{Var}(z_3) = 7.87$, $\text{Var}(z_4) = 1.056$

The total variation explained by original variables is

$$= \text{Var}(x_1) + \text{Var}(x_2) + \text{Var}(x_3) + \text{Var}(x_4)$$

$$= 17.02 + 7.56 + 15.75 + 903.87 = 944.20$$

The total variation explained by principal components is

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 916.902 + 18.375 + 7.87 + 1.056 = 944.20$$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.

The proportion of total variation accounted for by the first principal component is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{916.902}{944.203} = .97$$

Continuing, the first two components account for a proportion

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{935.277}{944.203} = .99$$

of the total variance.

Hence, in further analysis, the first or first two principal components z_1 and z_2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of x_i s in equations of z_i s. For above data, the first two principal components for first observation i.e. for year 1970 can be worked out as

$$z_1 = 0.006 \times 25.0 + 0.061 \times 86 + 0.103 \times 66 + 0.993 \times 186.49 = 197.380$$

$$z_2 = 0.955 \times 25.0 - 0.296 \times 86 + 0.011 \times 66 + 0.012 \times 186.49 = 1.383$$

Similarly for the year 1971

$$z_1 = 0.006 \times 24.9 + 0.061 \times 84 + 0.103 \times 66 + 0.993 \times 124.34 = 135.54$$

$$z_2 = 0.955 \times 24.9 - 0.296 \times 84 + 0.011 \times 66 + 0.012 \times 124.34 = 1.134$$

Thus the whole data with four variables can be converted to a new data set with two principal components.

Note: The principal components depend on the scale of measurement, for example, if in the above example X_1 is measured in ^{0}F instead of ^{0}C and X_4 in mm in place of cm, the data gives different principal components when transformed to original x's. In very specific situations results are same. The conventional way of getting around this problem is to use standardized variables with unit variances, i.e., correlation matrix in place of dispersion matrix. But the principal components obtained from original variables as such and from correlation matrix will not be same and they may not explain the same proportion of variance in the system. Further more, one set of principal components is not simple function of the other. When the variables are standardized, the resulting variables contribute almost equally to the principal components determined from correlation matrix. Variables should probably be standardized if they are measured on scales with widely differing ranges or if measured units are not commensurate. Often population dispersion matrix or correlation matrix are not available. In such situations sample dispersion matrix or correlation matrix can be used.

Applications of principal components:

- The most important use of principal component analysis is reduction of data. It provides the effective dimensionality of the data. If first few components account for most of the variation in the original data, then first few components' scores can be utilized in subsequent analysis in place of original variables.
- Plotting of data becomes difficult with more than three variables. Through principal component analysis, it is often possible to account for most of the variability in the data by first two components, and it is possible to plot the values of first two components scores for each individual. Thus, principal component analysis enables us to plot the data in two dimensions. Particularly detection of outliers or clustering of individuals will be easier through this technique. Often, use of principal component analysis reveals grouping of variables which would not be found by other means.
- Reduction in dimensionality can also help in analysis where no. of variables is more than the number of observations, for example, in discriminant analysis and regression analysis. In such cases, principal component analysis is helpful by reducing the dimensionality of data.
- Multiple regression can be dangerous if independent variables are highly correlated. Principal component analysis is the most practical technique to solve the problem. Regression analysis can be carried out using principal components as regressors in place of original variables. This is known as principal component regression.

Discriminant Analysis

Discriminant analysis and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a separatory procedure, it is often employed on a one - time basis in order to investigate observed differences when causal relationships are not well understood. Classification procedures are less explanatory in the sense that they lead to well- defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination.

Thus, the immediate goals of discrimination and classification, respectively, are as follows.

Goal 1. To describe either graphically (in three or lower dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose numerical values are such that the collections are separated as much as possible.

Goal 2. To sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign a new object to the labeled classes.

We shall follow convention and use the term discrimination to refer to Goal 1. This terminology was introduced by R.A. Fisher in the first modern treatment of separatory problems. A more descriptive term for this goal, however, is separation; we shall refer to the second goal as classification, or allocation.

A function that separates may sometimes serve as an allocation, and conversely, an allocatory rule may suggest a discriminatory procedure. In practice, Goals 1 and 2 frequently overlap and the distinction between separation and allocation becomes blurred.

Here we discuss Fisher's linear discriminant function for two multivariate populations having same dispersion matrix. For more general cases readers are requested to go through the references cited at the end.

Fisher's Discriminant Function

Here Fisher's idea was to transform the multivariate observations $\mathbf{x}$ to univariate observations y such that the y's derived from populations π_1 and π_2 were separated as much as possible. Fisher's approach assumes that the populations are normal and also assumes the population covariances matrices are equal because a pooled estimate of common covariance matrix is used.

A fixed linear combination of the $\mathbf{x}$'s takes the values y_{11}, y_{12}, ..., y_{1n1}, for the observations from the first population and the values y_{21}, y_{22}, ..., y_{2n2}, for the observations from the second population. The separation of these two sets of univariate y's is assessed in terms of the differences between $\bar{y}_1$ and $\bar{y}_2$ expressed in standard deviation units. That is,

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \text{ where } s_y^2 = \frac{\sum_{j=1}^{n1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the variance. The objective is to select the linear combination of the **x** to achieve maximum separation of the sample means $\bar{y}_1$ and $\bar{y}_2$.

Result: The linear combination y = $\hat{\boldsymbol{l}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\mathbf{pooled}}^{-1}\mathbf{x}$ maximizes the ratio

$$\frac{\text{(Squared distance between sample means of y)}}{\text{(Sample variance of y)}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{(\hat{\boldsymbol{l}}'\bar{\mathbf{x}}_1 - \hat{\boldsymbol{l}}'\bar{\mathbf{x}}_2)^2}{\hat{\boldsymbol{l}}'\boldsymbol{S}_{pooled}\hat{\boldsymbol{l}}} = \frac{(\hat{\boldsymbol{l}}'\mathbf{d})^2}{\hat{\boldsymbol{l}}'\boldsymbol{S}_{pooled}\hat{\boldsymbol{l}}}$$

overall possible coefficient vectors $\hat{\boldsymbol{l}}'$ where $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. The maximum of the above ratio is $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'s_{\mathbf{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, the Mahalanobis distance.

Fisher's solution to the separation problem can also be used to classify new observations. An allocation rule is as follows.

Allocate **x₀** to π_1 if

$$y_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'s_{\mathbf{pooled}}^{-1}\mathbf{x}_0 \geq \hat{\mathbf{m}} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'s_{\mathbf{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

and to π_2 if

$$y_0 < \hat{\boldsymbol{m}}$$

If we assume the populations π_1 and π_2 are multivariate normal with a common covariance matrix, then a test of **H₀: μ₁ = μ₂** versus **H₁: μ₁ ≠ μ₂** are accomplished by referring

$$\frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p}\left(\frac{n_1 n_2}{n_1 + n_2}\right)\mathbf{D^2}$$

to an F-distribution with $\upsilon_1 = p$ and $\upsilon_2 = n_1 + n_2 - p - 1$ d.f. If **H₀** is rejected, we can conclude the separation between the two populations is significant.

Example:

To construct a procedure for detecting potential hemophilia 'A' carriers, blood samples were analyzed for two groups of women and measurements on the two variables, x_1 = $\log_{10}$(AHF activity) and x_2 = loh_{10}(AHF-like antigens) recorded. The first group of n_1 = 30 women were selected from a population who do not carry hemophilia gene (normal group). The second group of n_2 = 22 women were selected from known hemophilia 'A' carriers (obligatory group). The mean vectors and sample covariance matrix are given as

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} \text{ and } \mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} 131158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Now the linear discriminant function is

$$y_0 = \hat{\boldsymbol{l}}' \mathbf{x}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0$$

$$= \begin{bmatrix} .2418 & -0.0652 \end{bmatrix} \begin{bmatrix} 131158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 37.61x_1 - 28.92\, x_2$$

Moreover

$$\bar{y}_1 = \hat{\boldsymbol{l}}' \, \bar{\mathbf{x}}_1 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} = 0.88$$

$$\bar{y}_2 = \hat{\boldsymbol{l}}' \, \bar{\mathbf{x}}_2 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.2483 \\ -0.0262 \end{bmatrix} = -10.10$$

and the mid-point between these means is

$$\hat{\mathbf{m}} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{s}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = -4.61$$

Now to classify a women who may be a hemophilia 'A' carrier with x_1 = -.210 and x_2 = -0.044, we calculate

$y_0 = \hat{\boldsymbol{l}}' \mathbf{x}_0 = 37.61x_1 - 28.92\, x_2 = -6.62$

Since $y_0 < \hat{m}$ we classify the women in π_2 population, i.e., to obligatory carrier group.

Factor Analysis

Some Basics

Factor analysis is a data reduction technique, which often requires large sample size to have a valid interpretation. The basic idea in factor analysis is that a large number of explanatory variables having similar type of responses can be captured with a single latent variable that cannot be measured directly. For example, the latent variable (or factor) socioeconomic status is associated with the observed variables income, education, health status, occupation, on which the peoples' responses are of similar type.

In factor analysis, the number of factors is same as the number of variables, where each factor captures a certain amount of variation of all the variations present in the observed variables. The factors are always arranged in the decreasing order of their variances. In factor analysis, one expects three outputs viz., common factor variances, factor loadings and factor scores. The common factor variance is the measure of the amount variation explained by a factor present in the observed variables. Factor loading measures the underlying relationship that an observed variable have with a factor. The factor scores are the transformed data, commonly the weighted sum/mean of the observed variables (or manifest variables).

The factor scores are not the penultimate output rather than act as an intermediate step (dimensionality reduction) for carrying out further statistical analysis, a much important one. In other words, factor scores enable user to use a single variable, instead of set of variables, as a measure of the factor in the other statistical investigation. For example, in case of linear model or mixed model, the factor scores can be used as variable (fixed factors or random factors), but here it refers to the categorical independent variable. Further, technically the factor scores are continuous and hence can be used as covariates in the model rather than as factors.

Type of Factor Analysis

There are two types of factor analysis, one is Exploratory Factor Analysis (EFA) and other is Confirmatory Factor Analysis (CFA). In CFA, one assumption is that there should be prior information about the number of factors likely to be encountered as well as which variables will be loaded onto which factors. On the other hand, CFA allows the researchers to test the hypothesis that whether the relationship between a variable and the underlying factor exits or not. Initially, the researcher postulates a certain a priori relationship pattern based on existing knowledge i.e., published research (empirical and/or theoretical) and then test the hypothesis statistically. In EFA, the researcher tries to find out the number of underlying constructs (factors) without having any a priori information about the number of factors. In other words, in EFA, the number of factors is determined on the basis of the dataset supplied by the user, and also depends upon user interpretation. Linking these two approaches, one can use EFA first to explore the underlying factors and then perform CFA to validate the structure of factors in a new dataset that has not been used for performing EFA. For example, a factor "depression" can be obtained with underlying variables depressed mood, fatigue, exhaustion and social dysfunction through EFA for a sample of rural women, and then the CFA can be used to validate this factor using a sample of urban women. In EFA, the cut-off of loading are much relaxed than that of CFA. In other words, a variable having loading value <|0.7| is disqualified from its loading onto a certain factor (Thumb rule). Generally, the EFA is most commonly used in day-to-day life than that of CFA. So, in this study material we only focused on EFA.

Exploratory Factor Analysis (EFA)

Before carrying out factor analysis, some important points need to be considered. At first, the reliability of the dataset should be checked for factor analysis. In other words, for factor analysis, the values of the variables should be in interval scale, each variable should be normally distributed, pairs of variables should follow bi-variate normal distribution and the dataset as a whole should follow multivariate normal distribution. Further, the sample size should be large. Field (2000) suggested 10-15 observations per variable. Habing (2003) state that there should be at least 50 observations and the number of observations should be at least 5 times as many variables. Comrey (1973) categorized the sample size for its suitability to factor analysis i.e., 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 or more as excellent. Also, one can conduct Kaiser-Meyer-Olkin (KMO) test to check the sample adequacy. The sample is said to be adequate if KMO value is more than 0.5.

As far as correlation matrix is concerned, the observed variables should be linearly related but not highly correlated that may lead to the matrix as singular and create difficulty in determining the unique contribution of the variables to the factors. To check the correlation among variables, one can use Bartlett's test of spherity to test the null hypothesis that the correlation matrix is a identity matrix and the result should come out as significant. After rejecting the null hypothesis, one can validate the presence of multi-collinearity via the determinant of the correlation matrix ie., if the determinant is greater than 0.00001, then there is no multi-collinearity (Field, 2000).

After getting correlation matrix, it is essential to determine whether factor analysis (FA) or principal component analysis (PCA) is to be performed. The main difference between these two lies on the way the eigen values are used. In PCA, all the diagonal elements of the correlation matrix are 1 and all the variance present in the dataset are accounted by the components. However, in FA, the diagonal of the correlation matrix are squared multiple correlation coefficient, which is further used to get the eigen values and thereby the factor scores. Also, all the variances are not accounted by the factors as there is also an error variance. Further, in PCA the sum of square of the factor loadings of a variable provided the variance accounted for by that variable, which is not same in FA as it is assumed that the variables do not account for 100% of the variance. Theoretically, FA is more correct than PCA (Field, 2000) but practically there is little difference and is further decreased with decrease in the number of variables and increase in the value of factor loadings (Rietveld and Van Hout, 1993).

In conducting FA, one of the most important questions is the number of factors to be retained in the model. In PCA, the number of components is same as the number of positive eigen value. However eigen values are sometime positive and close to zero, and in that situation deciding the number of factor is difficult. In literature certain thumb rules are there to take decision about the number of factors. Guttman-Kaiser rule state that the factor with eigen value >1 should be retained in the model. Hair et al, (1995) stated that in the natural sciences the number factors retained in the model should explain at least 95% of the total variance present in the observed variables. In humanities, the number factors that can explain up to 60-70% variation may be retained in the model (Hair et al, 1995; Pett et al, 2003). Besides, another option is that first draw a scree plot (Cattell, 1966) and retained all those factors appeared before reaching the point of inflection.

After extracting the factors, the next task is to name the factors and interpret them. Since, most variable have higher value of loading on the most important factors and less amount of loadings on the remaining factors, it is always a difficult task to interpret about the factors. However, the factor rotation can help in this respect to a large extent. Factor rotation transforms the original loadings and thereby the interpretation becomes easier. Rotation maximizes the high loading items and minimizes the less loading items. There are two rotation techniques viz., orthogonal/

varimax and oblique/promax that are commonly used in factor analysis. Varimax rotation (Thomson, 2004) is the most common rotational technique used in factor analysis that produces uncorrelated factors. On the other hand, in oblique rotation, the factors are correlated. Often, the oblique rotation provides more accurate results when the data does not meet the prior assumptions. Further, to decide the type of rotation technique is almost difficult and therefore first carryout the analysis with oblique rotaions, and if the oblique rotation demonstrates a negligible correlation between the extracted factors then it is reasonable to use orthogonally rotated factors (Field, 2000). Regardless of the rotation techniques uses, the objective is to provide easier interpretation of the results.

Interpretation of EFA is nothing but to determine which variables are attributed to a factor and labeling of that factor. However, the labeling of a factor is a subjective process (Henson and Roberts, 2006), where the meaningful of the factor is dependent on the researchers definition. Moreover, through and systematic factor analysis is nothing but to find those factors that together explain the majority of the responses.

Mathematical aspects of EFA

Consider a dataset with *n* observations and *p* standardized variables $x_1, x_2, ..., x_p$. Then, in EFA the observed variables are expressed as the linear combination of the common factors and unique factor i.e., $x_i = a_{i1}F_1 + a_{i2}F_2 + a_{i3}F_3 + ... + a_{ik}F_k + e_i$, where i=1,2,…, p, k<p and a_{ik} is the factor loading of i^{th} variable on k^{th} factor which is not same as that of eigen vector. The assumptions of this model are $E(e_i) = 0$, $V(e_i) = \psi_i$, $E(e_i e_j) = 0$, $E(e_i F_j) = 0$ and $E(F_i F_j) = 0$. In matrix notation we can write $\mathbf{X}_{p\times n} = \mathbf{L}_{p\times k}\mathbf{F}_{k\times n} + \mathbf{E}_{p\times n}$, where

$$\mathbf{X}_{p\times n} = \begin{bmatrix} x_{11}\ x_{12}\ x_{13}\ ...\ x_{1n} \\ x_{21}\ x_{22}\ x_{23}\ ...\ x_{2n} \\ ...\ ...\ ...\ ...\ ... \\ x_{p1}\ x_{p2}\ x_{p3}\ ...\ x_{pn} \end{bmatrix},\ \mathbf{F}_{k\times n} = \begin{bmatrix} F_{11}\ F_{12}\ F_{13}\ ...\ F_{1n} \\ F_{21}\ F_{22}\ F_{23}\ ...\ F_{2n} \\ ...\ ...\ ...\ ...\ ... \\ F_{k1}\ F_{k2}\ F_{k3}\ ...\ F_{kn} \end{bmatrix},\ \mathbf{L}_{p\times k} = \begin{bmatrix} a_{11}\ a_{12}\ a_{13}\ ...\ a_{1k} \\ a_{21}\ a_{22}\ a_{23}\ ...\ a_{2k} \\ ...\ ...\ ...\ ...\ ... \\ a_{p1}\ a_{p2}\ a_{p3}\ ...\ a_{pk} \end{bmatrix} \text{ and}$$

$$\mathbf{E}_{p\times n} = \begin{bmatrix} e_{11}\ e_{12}\ e_{13}\ ...\ e_{1n} \\ e_{21}\ e_{22}\ e_{23}\ ...\ e_{2n} \\ ...\ ...\ ...\ ...\ ... \\ e_{p1}\ e_{p2}\ e_{p3}\ ...\ e_{pn} \end{bmatrix}.$$

Also, it is assumed that $E(\mathbf{E}) = 0$, $E(\mathbf{F}) = 0$, $\text{cov}(\mathbf{F}, \mathbf{E}) = 0$, $V(\mathbf{E}) = Diag(\psi_1, \psi_2, ..., \psi_p) = \boldsymbol{\psi}(say)$ and $\text{var}(\mathbf{F}) = \mathbf{I}$. The correlation matrix is generally used for performing the factor analysis. Here the diagonal elements are 1 (often described as the variance of the observed variable). In PCA, this matrix is used as such but factor analysis involves the replacing of diagonal element with communality estimate. The communality estimate is the estimated proportion of variance of the variable that is free of error variance and is shared with other variables in the matrix. These estimates reflect the variance of a variable in common with all others together. The initial estimate of the communality is taken as the squared multiple correlation coefficients and then the communalities of the variables are estimated as the sum of the square of the loadings onto different factors. Once the correlation matrix of the observed variables are obtained, the factor analysis can be written as $\mathbf{\Sigma} = \mathbf{LL}' + \boldsymbol{\psi}$, which nothing but $\text{var}(\mathbf{X}_{p\times n}) = \text{var}(\mathbf{L}_{p\times k}\mathbf{F}_{k\times n} + \mathbf{E}_{p\times n})$. So, for the i^{th} variable, one can write $1 = (a_{i1}^2 + a_{i2}^2 + ... + a_{ip}^2) + \psi_i$ or $1 = h_i^2 + \psi_i$ or Total variance=Variance explained by the common factors + Error variance. Here h_i^2 is the communality and 1-h_i^2 is the variance accounted for by the i^{th} unique factor. In this model, there is a need to estimate the common factor loadings (***L***) as well as the factor scores (***F***). For estimating ***L***, there are two methods available one is Principal Axis Factor (PAF) method and other is Maximum Likelihood (ML) method. PAF makes no assumption about the error and minimizes the sum of squares of the residual matrix i.e., $\frac{1}{2}tr\left[(S-\Sigma)^2\right] = \sum_i\sum_j(s_{ij}-\sigma_{ij})^2$, where s_{ij} and σ_{ij} are the observed correlation matrix and implied correlation matrix, respectively (Jöreskog, 2007). The maximum likelihood (ML) estimation is derived from the theory of normal distribution. The ML value is obtained by minimizing $\ln|\Sigma| - \ln|S| + tr[S\Sigma^{-1}] - p$, which similar to minimizing the discrepancy function $\sum_i\sum_j\left[\frac{(s_{ij}-\sigma_{ij})^2}{\psi_i^2\psi_j^2}\right]$ (MacCallum et al, 2007).

For estimation of factor scores, generally three types of methods are used viz., ordinary least squares, weighted least squares and regression method. Let ***x*** be the

(I) ***Ordinary Least Square***

The estimate of $\mathbf{f}_i$ can be obtained by minimizing the error sum of squares i.e., $\sum_{j=1}^{p} e_{ij}^2 = \sum_{j=1}^{p} (x_{ij} - a_{i1} f_1 - a_{i2} f_2 - ... - a_{ik} f)^2 = (\mathbf{x}_i - \mathbf{Lf}_i)'(\mathbf{x}_i - \mathbf{Lf}_i)$. This is like a least squares regression, except in this case we already have estimates of the parameters (the factor loadings). In matrix notations, it can be written as $\hat{\mathbf{f}}_i = (\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}'\mathbf{x}_i$. Using the principal component method with the unrotated factor loadings, the results can be obtained as

$$\hat{\mathbf{f}}_i = \begin{pmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}} \hat{\zeta}_1 \mathbf{x}_i \\ \frac{1}{\sqrt{\hat{\lambda}_2}} \hat{\zeta}_2 \mathbf{x}_i \\ \cdots \\ \frac{1}{\sqrt{\hat{\lambda}_k}} \hat{\zeta}_k \mathbf{x}_i \end{pmatrix},$$

where $\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_k$ are the eigen vectors and $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k$ are the estimate of eigen values.

(II) ***Weighted Least Squares***

In this method, larger weights are given to the variables having low specific variances. Variables with low specific variances are those for which the model fits the data best. In other words, the variable with the low specific variance provides more information regarding the true values for the specific factors. For the above considered model, we wish to minimize

$$\sum_{j=1}^{p} \frac{e_{ij}^2}{\psi_j} = \sum_{j=1}^{p} \frac{(x_{ij} - a_{i1} f_1 - a_{i2} f_2 - ... - a_{ik} f)^2}{\psi_j} = (\mathbf{x}_i - \mathbf{Lf}_i)'\boldsymbol{\psi}^{-1}(\mathbf{x}_i - \mathbf{Lf}_i) \quad , \quad \text{that}$$

resulted in the estimate as $\hat{\mathbf{f}}_i = (\mathbf{L}'\boldsymbol{\psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\psi}^{-1}\mathbf{x}_i$. Both OLS and WLS methods are used for estimating the factor scores, while PAF method is used to estimate the factor loadings.

(III) ***Regression method***

This method is used when maximum likelihood is used for estimating the factor loadings. Now, for standardized variables the joint distribution of $\mathbf{x}_i$

and $\mathbf{f}_i$ can be writes as $\begin{pmatrix}\mathbf{x}_i \\ \mathbf{f}_i\end{pmatrix} \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix},\begin{pmatrix}\mathbf{LL}'+\boldsymbol{\psi} & \mathbf{L} \\ \mathbf{L}' & \mathbf{I}\end{pmatrix}\right]$. Then, we can calculate the conditional expectation of the factor score $\mathbf{f}_i$ given the observed data $\mathbf{x}_i$ as $E(\mathbf{f}_i|\mathbf{x}_i) = \mathbf{L}'(\mathbf{LL}'+\boldsymbol{\psi})^{-1}\mathbf{x}_i$, which is nothing but the estimate of $\mathbf{f}_i$.

Step by step procedure for performing exploratory factor analysis using R

Step 1: Set the working directory. Let my directory is “meher” present in “D” drive. Then, set the directory as

```
setwd(“C:/Documents and Settings/Prabin/Desktop/meher”)
```

Step 2: Read the data from the specified directory. Let my data file is *fact.txt* present in the directory. Then data file can be imported to R as

```
x <- read.table (file= “fact.txt”)
```

Step 3: Check the normality assumption of each variable using Shapiro-Wilk’s test.

```
shapiro.test (x[,i])      # This is for ith variable. If P-value is >level of significance, the variable is normally distributed.
```

Step 4: Check the adequacy of the each variable and sample as a whole for factor analysis using KSA and KMO and test. The desired value of KMO is > 0.5. Variables with MSA being below 0.5 indicate that item does not belong to a group and may be removed from the factor analysis.

```
kmo <- function(x)
{
x <- subset(x, complete.cases(x)) # Omit missing values
r <- cor(x)                                    # Correlation matrix
r2 <- r^2                  # Squared correlation coefficients
i <- solve(r)                 # Inverse matrix of correlation matrix
d <- diag(i)               # Diagonal elements of inverse matrix
p2 <- (-i/sqrt(outer(d, d)))^2          # Squared partial correlation coefficients
diag(r2) <- diag(p2) <- 0        # Delete diagonal elements
KMO <- sum(r2)/(sum(r2)+sum(p2))
MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
return(list(KMO=KMO, MSA=MSA))
}
kmo (x)
```

Step 5: Check that the correlation matrix is not an identity matrix using Bartlett’s sphericity test. The test should come out significant.

```
bst <- function(x)
{
method <- "Bartlett's test of sphericity"
data.name <- deparse(substitute(x))
x <- subset(x, complete.cases(x)) # Omit missing values
n <- nrow(x)
p <- ncol(x)
chisq <- (1-n+(2*p+5)/6)*log(det(cor(x)))
df <- p*(p-1)/2
p.value <- pchisq(chisq, df, lower.tail=FALSE)
names(chisq) <- "X-squared"
names(df) <- "df"
return(structure(list(statistic=chisq, parameter=df,
p.value=p.value,
method=method, data.name=data.name), class="htest"))
}
bst (x)
```

Step 6: Test that there is no presence of high degree of multicollinearity. The determinant of the matrix should come out > 0.0001 to pass the test.

```
det(cor(x))
```

Step 7: Carryout factor analysis to extract the factor loadings (by ML estimate method), common variances and specific variances.

```
factanal (x=swiss, factors=2, rotation= “varimax or
promax”)
or
factanal (~., factors=2, data=swiss, rotation= “varimax
or promax”)
```

# In the result one cannot see the complete factor loadings but it is possible with the following commands.

```
factanal (~., factors=2, rotation= “varimax or
promax”)$loadings[,i] # for complete ith factor loading.
```

Step 8: Estimate the factor scores either by Bartlett’s WLS method or Johnson’s regression method.

```
factanal (~., factors=2, rotation= “varimax or promax”,
scores=”Bartlett or regression”)$scores
```

Step 9: The factor loadings, common variances, specific variances can also be computed by supplying the covariance matrix and number of observations. However, the scores can only be obtained when full data set is available.

```
factanal (factors=2, covmat=cor(swiss),rotation= “varimax
or promax”, n.obs=47)
```

Step 10: Interpretation of the result and conclusion

Note: One can use the “psych” package of R-software for KMO test and Barlett’s test of sphericity using single line code as provided below.

`KMO(r)` # r is the correlation matrix. This will provide the values of both KMO and KSA

`cortest.bartlett(r, n)` # r is the correlation matrix and n is the number of observation in the dataset.

References

Cattell, RB (1966) The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

Chatfield, C. and Collins, A.J. (1990). Introduction to multivariate analysis. *Chapman and Hall publications*.

Comrey, AL (1973) *A First Course in Factor Analysis*. New York: Academic Press, Inc.

Field, A (2000) *Discovering Statistics using SPSS for Windows*. London – Thousand Oaks – New Delhi: Sage publications.

Habing, B (2003) *Exploratory Factor Analysis*. Website: http://www.stat.sc.edu/~habing/courses/530EFA.pdf

Hair, J., Anderson, RE., Tatham, RL., Black, WC (1995) *Multivariate data analysis*. 4th edn. New Jersey: Prentice-Hall Inc.

Henson, RK., Roberts, JK (2006) Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, 66(3).

Johnson, R.A. and Wichern, D.W. (1996). Applied multivariate statistical analysis. *Prentice-Hall of India Private Limited*.

Jöreskog, G (2007) *Factor analysis and its extensions*, in *Factor analysis at 100: Historical Developments and Future Directions*, R. Cudeck and R.C. MacCallum, eds., Lawrence Erlbaum, Mahwah, NJ, pp. 47–77.

MacCallum, RC., Browne, MW., Cai, L (2007) *Factor analysis models as approximations*, in *Factor Analysis at 100: Historical Developments and Future Directions*, R. Cudeck and R.C. MacCallum eds., Lawrence Erlbaum, Mahwah, NJ, pp. 153–175.

Pett, MA., Lackey, NR., Sullivan, JJ (2003) *Making Sense of Factor Analysis: The use of factor analysis for instrument development in health care research*. California: Sage Publications Inc.

Rietveld, T., Van Hout, R (1993) *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin – New York: Mouton de Gruyter.

Thompson, B (2004) *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington, DC: American Psychological Association.

CORRELATION AND REGRESSION ANALYSIS

Dr.Kanchan Sinha
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
kanchan.sinha@icar.gov.in

1. Introduction

Correlation is a powerful statistical concept that enables us to explore the relationships between variables and uncover hidden patterns in complex data. By measuring the extent to which two variables move together, correlation helps us gain insights into the interconnectedness of phenomena. In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). Regression analysis is primarily used for two distinct purposes. First, it is widely used for prediction and forecasting, which overlaps with the field of machine learning. Second, it is also used to infer causal relationships between independent and dependent variables. This methodology is widely used in business, social and behavioral sciences, biological sciences including agriculture. For example, yield of a crop can be predicted by utilizing the relationship between yield and other factors like water temperature, rainfall, quantity of fertilizer, quantity of seeds, irrigation level and relative humidity, etc.

A functional relationship between two variables can be expressed by a mathematical formula. If x denotes the independent variable and y the dependent variable, then y can be related x through a functional relation of the form $y = f(x)$. Given a particular value of x, the function f indicates the corresponding value of y. In regression analysis, the variable x is known as input variable, explanatory variable or predictor variable. This is an exact mathematical relationship. In statistical relation, may not be perfect owing to sampling. The above functional form is made a statistical model by adding an error term as $y = f(x) + \varepsilon$, where ε denotes the error term.

Depending on the nature of the relationships between x and y, regression approach may be classified into two broad categories *viz.*, linear regression models and nonlinear regression models. The response variable is generally related to other causal variables through some parameters. The models that are linear in these parameters are known as linear models; whereas in nonlinear models parameters appear nonlinearly.

2. The Concept of Correlation

2.1 *Defining Correlation*: Correlation refers to the statistical association between two or more variables, indicating the degree to which they tend to change together. It measures the direction (positive or negative) and strength (weak or strong) of the relationship.

2.2Significance of Correlation

Identifying Associations: Correlation helps us identify relationships between variables, providing a foundation for further analysis.

Prediction: Correlated variables can be used to make predictions about one variable based on the other(s).

Variable Selection: Correlation assists in selecting relevant variables for analysis, weeding out redundant or irrelevant ones.

2.3 Measuring Correlation

2.3.1 Pearson's Correlation Coefficient

The Pearson correlation coefficient (r) quantifies the linear relationship between two continuous variables and can be expressed as:

$$r = \frac{\Sigma \quad (x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\Sigma \quad (x_i - \underline{x})^2 \Sigma \quad (y_i - \underline{y})^2}}$$

Where, r is the correlation coefficient

x_i are the values of the x-variable.

y_i are the values of the y-variable.

$\underline{x}$ is the mean of the values of x-variable.

$\underline{y}$ is the mean of the values of y-variable.

Range and Interpretation: r ranges from -1 to 1, where -1 denotes a perfect negative correlation, 1 signifies a perfect positive correlation, and 0 indicates no linear relationship.

Strength of Correlation: Various criteria, such as effect size or correlation coefficient magnitude, determine the strength of the relationship.

2.3.2 Spearman's Rank Correlation Coefficient

Spearman's rho (ρ) measures the monotonic relationship (increasing or decreasing) between variables, especially when the relationship is not strictly linear and can be expressed as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} \quad d_i^2}{n(n^2-1)}$$

where, ρ is the Spearman's Rank Correlation Coefficient.

d_i is the difference between the two ranks of each observation.

n is the number of observations.

Advantages: It is robust to outliers and can handle ordinal or non-normal data.

Interpretation: Similar to Pearson's r, ρ ranges from -1 to 1, with the same interpretations.

2.4 Types of Correlation

2.4.1 Positive Correlation

Definition: Positive correlation exists when an increase in one variable corresponds to an increase in the other, and vice versa.

Examples: Height and weight, income and education level.

2.4.2 Negative Correlation

Definition: Negative correlation occurs when an increase in one variable corresponds to a decrease in the other, and vice versa.

Examples: Temperature and heating costs, exercise duration and body weight.

2.4.3 Zero Correlation

Definition: Zero correlation indicates no discernible relationship between variables.

Examples: Shoe size and IQ, number of siblings and favourite colour.

2.4.4 Interpreting Correlation

2.4.4.1 Causation vs. Correlation

Correlation does not imply causation; a strong relationship between two variables does not necessarily mean one variable causes the other.

Spurious Correlation: Be cautious of coincidental associations without a meaningful underlying connection.

2.4.4.2 Scatterplots

Visualizing Correlation: Scatter plots are graphical representations that help us assess the relationship between variables.

Patterns: Scatterplots can exhibit various patterns, such as linear, nonlinear, or clusters, aiding in understanding the correlation visually.

2.4.4.3 Applications of Correlation

Finance and Economics

Analyzing stock market trends and investment portfolios.

Examining relationships between economic indicators, such as GDP and unemployment rates.

Social Sciences

Investigating relationships between variables like crime rates and income levels.

Studying the impact of education on health outcomes.

Medicine and Health

Exploring the correlation between risk factors and disease prevalence.

Assessing the effectiveness of treatments or interventions.

Agriculture

Crop Yield and Environmental Factors

Pest and Disease Management

Crop Nutrient Requirements

Crop-Livestock Interactions

Climate Change Impact Assessment

Water Management

Market Analysis and Price Forecasting, etc.

3. Simple Linear Regression (SLR) Model

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

The simple linear regression model is usually written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i (3)$$

where the ε_i's are normal random variables with mean 0 and variance σ^2. The model implies (i) The average y-value at a given $x-$value is linearly related to x.

(ii) The variation in responses y at a given x value is constant.

(iii) The population of responses y at a given x is normally distributed.

(iv) The observed data are a random sample.

Regression model (3) is said to be simple and linear regression model. It is “simple” in the sense that there is only one predictor variable and “linear” in the sense that all parameters appeared linearly with the predictor variables. The parameters $\beta_0$and β_1 in regression model (3) are called regression coefficients, β_1 is the slope of the regression line. It indicates the change in the mean of the probability distribution of y per unit increase in x. The parameter β_0 is the y intercept of the regression line.

3.1 Estimation of Parameters in a Simple Linear Regression Model

In the above models the variables y and x are known, these are observed. The only unknown quantities are the parameters β’s. In regression analysis, our main concern is how precisely we can estimate these parameters. Once these parameters are estimated, our model becomes known and we can use it for further analysis. The method of least squares is generally used to estimate these parameters. For each observations (x_i, y_i), the method of least squares considers the error of each observation, i.e, for a simple model $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$. The method of least squares requires the sum of the n squared errors. This criterion is denoted by S:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

According to the method of least squares, the estimators of β_0 and β_1 are those values of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, that minimize the criterion S for the given observations. To minimize S, we differentiate S with respect to each parameter and equate to zero. We get as many equations as the number of parameters. Solving these equations simultaneously, we get the estimates of parameters. For example, for the regression model (3) the values of $\hat{\beta}_0$ and $\hat{\beta}_1$that minimizes S for any particular set of sample data are given by the following simultaneous equations:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \qquad (6)$$

These two equations are called normal equations and can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$ as follows

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \ (x_i-\underline{x})(y_i-\underline{y})}{\sum_{i=1}^{n} \ (x_i-\underline{x})^2} \qquad (7)$$

$$\hat{\beta}_0 = \frac{1}{n}(\sum_{i=1}^{n} \ y_i - \beta_1 \sum_{i=1}^{n} \ x_i) = \underline{y} - \beta_1\underline{x} \ (8)$$

where,$\underline{y}$ and $\underline{x}$ are the means of the y_i and x_i observations, respectively.

3. Multiple Linear Regression Model (MLR) Model

A regression model that involves more than one regressor variable is called a multiple regression model i.e., the multiple linear regression model is used to study the relationship between a dependent variable and one or more independent variables. The generic form of the linear regression model is

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad (9)$$

where,$y$is the dependent or explained variable and $x_1, x_2, \dots, x_p$are the independent or explanatory variables. The regression model in the equation describes above is linear in the sense, it is a linear function of the unknown parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$. In general, any regression model that is linear in the parameters (β's) is a linear regression model, regardless of the shape of the surface that it generates. We have also assumed that the expected value of the error term ε is zero. The parameter β_0 is the intercept of the regression model. If the range of the data includes $x_1 = x_2 = \cdots = x_p = 0$, then β_0 is the mean of y when $x_1 = x_2 = \cdots = x_p = 0$. Otherwise β_0 has no physical interpretation. The parameter β_1 indicates the expected change in response (y) per unit change in x_1 when $x_2, \dots, x_p$ are held constant. Similarly $\beta_2$measures the expected change in response (y) per unit change in x_2 when $x_1, \dots, x_p$ are held constant. For this reason the parameters $\beta_i, \forall\ i = 1,2, \dots, p$ are often called as partial regression coefficients.

A. Assumptions of the Multiple Linear Regression Model

1. Linearity

The model defined by the following equation

$y = f(x_1, x_2, \dots, x_p) + \varepsilon = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$ specifies a linear relationship between y and $x$and our primary interest is in estimation and inference about the parameter vector $\boldsymbol{\beta}$. For the regression to be linear in the sense described here, it must be of the form in the original variables or after some suitable transformation.

i. Full rank

There are no exact linear relationships among the variables in the model. x is an $n \times p$ matrix with rank p. Hence x has full column rank; the columns of x are linearly independent and there are at least p observations $(n \geq p)$.

i. Exogeneity of the independent variables:

The disturbance is assumed to have conditional expected value zero at every observation, which we can write as $E[x] = 0$.

In this equation, the left hand side states, in principle, that the mean of each ε_i conditioned on all observations $x$is zero. This strict exogeneity assumption states, in words, that no observations on x convey information about the expected value of the disturbance.

i. Homoscedasticity:

The fourth assumption concerns the variances and covariance of the disturbances:

$$Var(x) = \sigma^2, \forall\, i = 1, \ldots, n$$

$$Cov(x) = 0\ \forall\, i \neq j \qquad (10)$$

Constant variance is labelled **homoscedasticity**. Consider a model that describes the profits of firms in an industry as a function of, say, size. Even accounting for size, measured in dollar terms, the profits of large firms will exhibit greater variation than those of smaller firms. The homoscedasticity assumption would be inappropriate here. Survey data on household expenditure patterns often display marked **heteroscedasticity**, even after accounting for income and household size. The two assumptions imply that

$$E[x] = [\sigma^2\ 0\ \cdots\ 0\ 0\ \sigma^2\ \cdots\ 0\ \vdots\ \vdots\ \ddots\ \vdots\ 0\ 0\ \cdots\ \sigma^2\ ] = \sigma^2 I \ (11)$$

i. Data generating process for the regressors

It is common to assume that $x_i$is nonstochastic, as it would be in an experimental situation. Here the analyst chooses the values of the regressors and then observes y_i. This process might apply, for example, in an agricultural experiment in which y_i is yield and $x_i$is fertilizer concentration and water applied.

i. Normality

It is convenient to assume that the disturbances are normally distributed with zero mean and constant variance. This is a convenience that we will dispense with after some analysis of its implications. The normality assumption is useful for defining the computations behind statistical inference about the regression, such as confidence intervals and hypothesis tests. For practical purposes, it will be useful then to extend those results and in the process develop a more flexible approach that does not rely on this specific assumption.

$\varepsilon|x{\sim}N(0,\sigma^2 I)(12)$

The validity of these assumptions is needed for the results to be meaningful. If these assumptions are violated, the result can be incorrect and may have serious consequences. If these departures are small, the final result may not be changed significantly. But if the deviations are large, the model obtained may become unstable in the sense that a different sample could lead to an entirely different model with opposite conclusions. So such underlying assumptions have to be verified before attempting to regression modeling. One crucial point to keep in mind is that these assumptions are for the population, and we work only with a sample. So the main issue is to make a decision about the population on the basis of a sample of data. Several diagnostic methods to check the violation of regression assumption are based on the study of model residuals and also with the help of various types of graphics.

4.1Estimation of Parameters in a Multiple Linear Regression (MLR) Model

The method of least squares can be used to estimate the regression coefficients in Eq. (9). Suppose that $n > p$ observations are available, and let y_i denote the ith observed response and x_{ij} denote ith observation or level of regressor x_j. The data will appear in the following table 1. We also assume that the error term ε in the model has $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, and the errors are uncorrelated.

Table 1: Data for Multiple Linear Regression

		Regressors		
Observation, i	Response, y	x_1	x_2	x_p
1	y_1	x_{11}	x_{12}	x_{1p}
2	y_2	x_{21}	x_{22}	x_{2p}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	y_n	x_{n1}	x_{n2}	x_{np}

We may write the sample regression model corresponding to (9) as

$$y = \beta_0{+}\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon$$

$= \beta_0 + \sum_{j=1}^{p} \quad \beta_j x_{ij} + \varepsilon_i, \forall\, i = 1,2, \dots, n$

The least - squares function is then used to estimate the model parameters, which are obtained by minimizing the error sum of squares with respect to the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.

It is more convenient to deal with multiple regression models if they are expressed in matrix notation. This allows a very compact display of the model, data, and results. In matrix notation, we can express the multiple regression model as

$y = X\beta + \varepsilon$(14)

Where

$$y = [y_1\, y_2 \,.\,.\,. y_n\ ]X = \left[1\; x_{11}\, \cdots\, x_{1p}\; 1\; x_{21}\, \cdots\, x_{2p}\ \vdots \vdots \ddots \vdots\ 1\; x_{n1}\, \cdots\, x_{np}\ \right]\beta$$

$$= \left[\beta_0\, \beta_1 \,.\,.\,. \beta_p\ \right]\varepsilon = [\varepsilon_1\, \varepsilon_2 \,.\,.\,. \varepsilon_n\ ]$$

$y$is a $n \times 1$ vector of responses

$X$is a $n \times p$ matrix of the regressor variables

$\beta$is a $n \times 1$ vector of unknown constants, and

$\varepsilon$is a $n \times 1$ vector of random errors with $\varepsilon_i \sim NID(0, \sigma^2)$

We wish to find the vector of least-squares estimators, $\hat{\beta}$that minimizes

$S(\beta) = \sum_{i=1}^{n} \quad \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$

Note that $S(\beta)$ may be expressed as

$$S(\beta) = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta$$

$= y'y - 2\beta'X'y + \beta'X'X\beta$(16)

Since$\beta'X'y$is a1×1matrix, or a scalar, and its transpose $(\beta'X'y)' = y'X\beta$is the same scalar. The least square estimators must satisfy

$$\frac{\partial S}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0$$

Which simplifies

$X'X\hat{\beta} = X'y$ (17)

To solve the normal equations, multiply both sides of (iv) by the inverse of $X'X$. Thus the least squares estimator of

$\hat{\beta} = (X'X)^{-1}X'y$(18)

So, the vector of fitted values $\hat{y}_i$ corresponding to the observed value y_i is

$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$(19)

The difference between the observed value y_i and the corresponding fitted values $\hat{y}_i$ is the residual i.e., $e_i = y_i - \hat{y}_i$. The $n$residuals may be conveniently written in matrix notation as

$e = y - \hat{y}$(20)

3. Estimation of Error Term Variance (σ^2)

The variance σ^2 of the error terms ε_i in regression model needs to be estimated to know the variability of the probability distribution of y. In addition, a variety of inferences concerning the regression function and the prediction of y require an estimate of σ^2. Denote by $SSE = \sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \quad r_i^2$, is the residual sum of squares. Then an estimate of σ^2 is given by,

$\hat{\sigma}^2 = \frac{SSE}{n-p}$ (21)

where p is the total number of parameters involved in the model including the intercept term, if the model contains it. We also denote this quantity by MSE.

3. Inferences in Linear Regression Models

In multiple linear regression model, all variables may not be contributing significantly to the model. In other word, each of the parameters may not be significant. Therefore, these parameters must be tested whether they are significantly different from zero or not. That is, we test the null hypothesis (H_0) against the alternative hypothesis $(H_1)$for a parameter β_i (say) as follows:

$H_0: \beta_i = 0$

$H_1: \neq 0$

when $H_0: \beta_i = 0$is accepted we infer that there is no linear association between y and x_i. For normal error regression model, the condition β_i implies even more than no linear association between y and x_i. $\beta_i = 0$ for the normal error regression model implies not only that there is no linear association between y and x_i but also that there is no relation of any kind between y and x_i, since the probability distribution of $y$are then identical at all levels of x_i. The test is based on t test

$t = \frac{\beta_i}{s(\beta_i)}$ (23)

where $s(\beta_i)$ is the standard error of $\beta_i$and calculated as $s(\beta_i) = \sqrt{\frac{MSE}{\sum_{i=1}^{n} \quad (x_i - \underline{x})^2}}$

The decision rule with this test statistic when controlling level of significance at α is

if $|t| \leq t\left(1 - \frac{\alpha}{2}; n - p\right)$ conclude H_0,

if $|t| > t\left(1 - \frac{\alpha}{2}; n - p\right)$ conclude H_1.

Similarly testing for other parameters can be carried out.

3. Measures of Fitting (R^2)

The overall fitting of a regression line can be judged by the F-statistic by carrying out an analysis of variance. If the F-statistic is significant, we say that our model is fitted well. However, there are times when the degree of linear association is of interest. A frequently used statistic is R^2. We describe this descriptive measure to describe the degree of linear association between y and x.

Denote by $TSS = \sum_i^n \quad \left(y_i - \underline{y}\right)^2$, total sum of squares which measures the variation in the observation y_i, or the uncertainty in predicting y, when no account of the predictor variable x is taken. Thus TSS is a measure of uncertainty in predicting y when x is not considered. Similarly, SSE measures the variation in the y_i when a regression model utilizing the predictor variable x is employed. A natural measure of the effect of x in reducing the variation in y, i.e., in reducing the uncertaintity in predicting y, is to express the reduction in variation ($TSS - SSE = SSR$ as a proportion of the total variation and it is denoted by

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad (24)$$

The measure R^2 is called coefficient of determination and $0 \leq R^2 \leq 1$. In practice R^2 is not likely to be 0 or 1 but somewhere between these limits. The closer it is to 1, the greater is said to be the degree of linear association between x and y. Remember that R^2 statistic should be used only when in the model an intercept term is involved. For the model with no intercept, $R^2$is not a good statistic. In case of "no intercept" model, sum of all residuals may not be equal to 0, making R^2 inflated.

3. An Illustration of a MLR model

Consider the following data:

Table 2: y as a response variable and x's as explanatory variables

Case No.	x_1	x_2	x_3	y		*Case No.*	x_1	x_2	x_3	Y
1	12.98	0.317	9.99	57.70		14	14.23	10.40	1.04	41.89
2	14.29	2.028	6.77	59.29		15	15.22	1.220	6.14	63.26
3	15.53	5.305	2.94	56.16		16	15.74	10.61	-1.91	45.79
4	15.13	4.738	4.20	55.76		17	14.95	4.815	4.11	58.69
5	15.3	7.038	2.05	51.72		18	14.12	3.153	8.45	50.08
6	17.14	5.982	-0.0	60.44		19	16.39	9.698	-1.7	48.89
7	15.46	2.737	4.65	60.71		20	16.45	3.912	2.14	62.21
8	12.80	10.66	3.04	37.44		21	13.53	7.625	3.85	45.62
9	17.03	5.132	0.25	60.97		22	14.19	4.474	5.11	53.92
10	13.17	2.039	8.73	55.27		23	15.83	5.753	2.08	55.79
11	16.12	2.271	2.10	59.28		24	16.56	8.546	8.97	56.74
12	14.34	4.077	5.54	54.02		25	13.32	8.589	4.01	43.14
13	12.92	2.643	9.33	53.19		26	15.94	8.290	-0.2	50.70

In the present example, we have 3 three predictor variables x_1, x_2 and x_3 and there are 26 observations. The response variable denoted by y. Applying least square method we obtain the parameter estimates as follows:

Table 3: ANOVA of a MLR model

Source	**Degrees of freedom**	**Sum of Square**	**Mean Square**	**F-value**	**Prob. > F**
Model	3	1062.34	354.11	109.69	<0.0001
Error	22	71.02	3.22		
Corrected Total	25	1133.37			

Table 4: Parameter Estimates of a MLR model

Variable	**Degrees of freedom**	**Parameter Estimates**	**Standard Error**	**t-value**	**Prob. > |t|**
Intercept	1	8.19	6.29	1.30	0.2060
x_1	1	3.56	0.36	9.86	<.0001
x_2	1	-1.64	0.15	-10.28	<.0001
x_3	1	0.33	0.17	1.88	0.0741

The value of $R^2$of this model is 0.93. From Table 3, we see that F-statistic is highly significant, indicating that overall model fitting is good. $R^2$is also very high. The fitted regression line is $\hat{y} = 8.19 + 3.56x_1 - 1.64x_2 + 0.33x_3$. The corresponding standard errors are given in the 4th column of Table 3. However, while testing the significance of the parameter estimates, we find that the intercept and the parameter for the variable x_3, i.e., are not significant at 5% level of significance (probability values for these parameters are greater than 0.05).

3. Practical Applications of regression analysis

Economics and Finance

Predicting stock market returns based on various economic indicators.

Analyzing the impact of interest rates on housing prices.

Marketing and Consumer Behavior

Understanding the factors influencing consumer purchasing decisions.

Predicting sales based on advertising expenditure and market demographics.

Healthcare and Medicine

Assessing the relationship between risk factors and disease outcomes.

Predicting patient outcomes based on treatment protocols and patient characteristics.

Agriculture

Crop Yield Prediction

Soil Fertility Assessment

Pest and Disease Management

Livestock Production

Economic Analysis and Market Forecasting, etc.

3. Conclusion

Correlation serves as a fundamental tool for analyzing relationships and unveiling hidden associations in data. By understanding the concept, measuring techniques, types, and interpretation of correlation, we can gain valuable insights and make informed decisions across a wide range of fields. Embracing correlation empowers us to unlock the intricate connections underlying the phenomena we observe, fostering a deeper understanding of the complex world around us.

Regression analysis serves as a versatile tool for understanding and predicting the relationship between variables. By comprehending the principles, assumptions, and types of regression analysis, we can harness its power to uncover patterns, make predictions, and inform decision-making across diverse fields. Embracing regression analysis empowers us to unravel the dynamics of complex systems, enabling us to navigate the intricacies of the world we inhabit with greater clarity and confidence.

3. References

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example*, New York: John Wiley & Sons.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, New York: Wiley Eastern Ltd.

Montgomery, D. C., Peck, E. and Vining, G. (2003). *Introduction to Linear Regression Analysis*, 3rd Edition, New York: John Wiley and Sons.

OVERVIEW OF SURVEY SAMPLING

Ankur Biswas
ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012
Ankur.Biswas@icar.gov.in

1. Introduction

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conservation and controversy require numerical data for their resolution. The data collected and analyzed in an objective manner and presented suitably serve as a basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include government, industry, business, research institutions, public organizations and international agencies and organizations. To discharge its various responsibilities, the government needs variety of information regarding different sectors of economy, trade, industrial production, health and mortality, population, livestock, agriculture, forestry, environment and available resources. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. For instance, the information on cost of living for different categories of people, living in various parts of the country is of importance in shaping its policies in respect of wages and price levels. Data on agricultural production are of immense use to the state for planning to feed the nation. In case of industry and business, the information is to be collected on labour, cost and quality of production, stock and demand and supply positions for proper planning of production levels and sales campaigns.

1.1 Complete enumeration

One way of obtaining the required information at regional and country level is to collect the data for each and every unit (person, household, field, factory, shop etc. as the case may be) belonging to the population which is the aggregate of all units of a given type under consideration and this procedure of obtaining information is termed as complete enumeration. The effort, money and time required for the carrying out complete enumeration to obtain the different types of data will, generally, be extremely large. However, if the information is required for each and every unit in the domain of study, a complete enumeration is clearly necessary. Examples of such situations are preparation of “voter list” for election purposes and recruitment of personnel in an establishment, etc. But there are many situations, where only summary figures are required for the domain of study as a whole or for group of units.

1.2 Need for sampling

An effective alternative to a complete enumeration can be sample survey where only some of the units selected in a suitable manner from the population are surveyed and an inference is drawn about the population on the basis of observations made on the selected units. It can be easily seen that compared to sample survey, a complete enumeration is time-consuming, expensive, has less scope in the sense of restricted subject coverage and is subject to greater coverage, observational and tabulation errors. In certain investigations, it may be essential to use specialized equipment or highly trained field staff for data collection making it almost impossible to carry out such investigations. It is of interest to note that if a sample survey is carried out according to certain specified statistical principles, it is possible not only to estimate the value of the characteristic of the population as a whole on the basis of the sample data, but also to get a valid estimate of the sampling error of the estimate. There are various steps involved in the planning and execution of the sample survey. One of the principal steps in a sample survey relates to methods of data collection.

1.3. Various concepts and definitions

i. Element:

An element is a unit about which we require information. For example, a field growing a particular crop is an element for collecting information on the yield of a crop.

ii. Population

The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

iii. Sampling unit

Elementary units or group of such units which besides being clearly defined, identifiable and observable, are convenient for the purpose of sampling are called sampling units. For instance, in a family budget enquiry, usually a family is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit.

iv. Sampling frame

A list of all the sampling units belonging to the population to be studied with their identification particulars or a map showing the boundaries of the sampling units is known as sampling frame. Examples of a frame are a list of farms and a list of suitable area segments like villages in India or counties in the United States. The frame should be up to date and free from errors of omission and duplication of sampling units.

v. Random sample

One or more sampling units selected from a population according to some specified procedures are said to constitute a sample. The sample will be considered as random or probability sample, if its selection is governed by ascertainable laws of chance. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection. For example, if a population consists of the N sampling units$U_1,U_2,\ldots,U_i,\ldots,U_N$ then, we may select a sample of n units by selecting them unit by unit with equal probability for every unit at each draw with or without replacing the sampling units selected in the previous draws.

vi. Non-random sample

A sample selected by a non-random process is termed as non-random sample. A non-random sample, which is drawn using certain amount of judgment with a view to get a representative sample, is termed as judgment or purposive sample. In purposive sampling units are selected by considering the available auxiliary information more or less subjectively with a view to ensuring a reflection of the population in the sample. This type of sampling is seldom used in large-scale surveys mainly because it is not generally possible to get strictly valid estimates of the population parameters under consideration and of their sampling errors due to the risk of bias in subjective selection and the lack of information on the probabilities of selection of the units.

vii. Population parameters

Suppose a finite population consists of the N units $U_1,U_2,\ldots,U_N$ and let Y_i be the value of the variable y, the characteristic under study, for the i^{th} unit U_i, (i=1,2,…,N). For instance, the unit may be a farm and the characteristic under study may be the area under a particular crop. Any function of the values of all the population units is known as a population parameter or simply a parameter. Some of the important parameters usually required to be estimated in surveys are population total and population mean.

viii. Statistic, estimator and estimate

Suppose, a sample of n units is selected from a population of N units, according to some probability scheme and let, the sample observations be denoted by $y_1, y_2, \ldots, y_n$. Any function of these values which is free from unknown population parameters is called a statistic.An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample and the samples are selected with specified probabilities. The particular value, which the estimator takes for a given sample, is known as an estimate.

ix. Sampling and non-sampling error

The error arises due to drawing inferences about the population on the basis of observations on a part (sample) of it, is termed sampling error. The sampling error is non-existent in a complete enumeration survey since the whole population is surveyed. On the contrary, the errors other than sampling errors such as those arising through non-response, in- completeness and inaccuracy of response are termed non-sampling errors and are likely to be more wide-spread and important in a complete enumeration survey than in a sample survey. Non-sampling errors arise due to various causes right from the beginning stage when the survey is planned and designed to the final stage when the data are processed and analyzed.

The sampling error usually decreases with increase in sample size (number of units selected in the sample) while the non-sampling error is likely to increase with increase in sample size.

As regards the non-sampling error, it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former.

2. Simple Random Sampling

Simple random sampling (SRS) can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. Simple random sampling is a method of selecting n units out of the N such that every one of the $\binom{N}{n}$ distinct samples has an equal chance of being drawn. In practice a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. Since a number that has been drawn is removed from the population for all subsequent draws, this method is also called random sampling without replacement. In case of a random sampling with replacement, at any draw all N members of the population are given an equal chance of being drawn, no matter how often they have already been drawn. The with-replacement assumption simplifies the estimation under complex sampling designs and is often adopted, although in practice sampling is usually carried out under a without replacement type scheme. Obviously, the difference between with replacement and without replacement sampling becomes less important when the population size is large and the sample size is noticeably smaller than it.

2.1 Procedure of selecting a random sample

Since probability sampling theory is based on the assumption of random sampling, the technique of random sampling is of basic significance. Some of the procedures used for selecting a random sample are as follows:

i) Lottery method

ii) Use of random number tables

i) Lottery Method:

Each unit in the population may be associated with a chit/ticket such that each sampling unit has its identification mark from 1 to N. All the chits are placed in a container, drum or metallic spherical device, in which a thorough mixing is possible before each draw. Chits may be drawn one by one and may be continued until a sample of the required size is obtained. When the size of population is large, this procedure of numbering units on chits and selecting one after reshuffling becomes cumbersome. In practice, it may be too difficult to achieve a thorough shuffling. Human bias and prejudice may also creep in this method.

ii) Use of Random Number Tables:

A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern where each position is filled with one of these digits. A Table of random numbers is so constructed that all numbers 0, 1, 2,…,9 appear independent of each other. Some random number tables in common use are:

- Tippett's random number Tables
- Fisher and Yates Tables
- Kendall and Smith Tables
- A million random digits Table

A practical method of selecting a random sample is to choose units one-by-one with the help of a Table of random numbers. By considering two-digit numbers, we can obtain numbers from 00 to 99, all having the same frequency. Similarly, three or more digit numbers may be obtained by combining three or more rows or columns of these Tables. The simplest way of selecting a sample of the required size is to select a random number from 1 to N and then taking the unit bearing that number. This procedure involves a number of rejections since all numbers greater than N appearing in the Table are not considered for selection. The procedure of selection of sample through the use of random numbers is, therefore, modified and one of these modified procedures is:

- **Remainder Approach:**

Let N be an r-digit number and let its r-digit highest multiple be N'. A random number k is chosen from 1 to N' and the unit with serial number equal to the remainder obtained on dividing k by N is selected, *i.e.* the selected number is reduced mod (N). If the remainder is zero, the last unit is selected. As an illustration, let N = 123, then highest three-digit multiple of 123 is 984. For selecting a unit, one random number from 001 to 984 has to be selected. Let the random number selected be 287. Dividing 287 by 123 gives the remainder as 41. Hence, the unit with serial number 41 is selected in the sample. Suppose that another random number selected is 245. Dividing 245 by 123 leaves 122 as remainder. So the unit bearing the serial number 122 is selected. Similarly, if the random number selected is 369, then dividing 369 by 123 leaves remainder as 0. So the unit bearing serial number 123 is selected in the sample.

2.2 Estimation of Population Total

Let Y be the character of interest and $Y_1, Y_2, \cdots, Y_i, \cdots, Y_N$ be the values of the character from*N* units of the population. Further, let $y_1, y_2, \cdots, y_i, \cdots, y_n$ be the sample of size n selected by simple random sampling without replacement. For the total $Y = \sum_{i=1}^{N} Y_i$ we have an estimator

$$\hat{Y} = N \sum_{i=1}^{n} y_i / n = N\overline{y}_n$$

i.e., the sample mean $\overline{y}_n$ multiplied by the population size N.

The estimator can be expressed as

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i = \left(N/n\right) \sum_{i=1}^{n} y_i \text{ , where } w_i = N/n.$$

The constant N/n is the sampling weight and is the inverse of the sampling fraction $n/N.$

The estimator has the statistical property of unbiasedness in relation to the sampling design. Variance of the estimator $\hat{Y}$ of the population total is given by

$$V_{SRS}\ (\ \hat{Y}\ ) = \ \frac{N^2}{n} \left(1\text{-}\frac{n}{N}\right) \sum_{i=1}^{N} \ (Y_i - \overline{Y})^2 \big/ (N-1)$$

where $\overline{Y} = \ \sum_{i=1}^{N} Y_i / N$ is the population mean and $S^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 / (N-1)$ is the population mean square.

An unbiased estimator of variance of the estimator $\hat{Y}$ of the total, $V_{SRS}(\hat{Y})$ is given by

$$\begin{aligned} \hat{V}_{SRS}(\ \hat{Y}\ ) &= N^2 \left(1\text{-}\frac{n}{N}\right) \sum_{i=1}^{n} \ (y_i - \overline{y}_n)^2 / n(n-1) \\ &= N^2 \left(1\text{-}\frac{n}{N}\right) s^2 / n \end{aligned}$$

where $\overline{y}_n = \sum_{i=1}^{n} y_i / n$ is the sample mean and s^2 is an unbiased estimator of the population mean square S^2.

3. Use of Auxiliary Information

In sampling theory if the auxiliary information, related to the character under study, is available on all the population units, then it may be advantageous to make use of this additional information in survey sampling. One way of using this additional information is in the sample selection with unequal probabilities of selection of units. The knowledge of auxiliary information may also be exploited at the estimation stage. The estimator can be developed in such a way that it makes use of this additional information. Ratio estimator, difference estimator, regression estimator, generalized difference estimators are the examples of such estimators. Obviously, it is assumed that the auxiliary information is available on all the sampling units. In case the auxiliary information is not available then it can be obtained easily without much burden on the cost.

Another way the auxiliary information can be used is at the stage of planning of survey. An example of this is the stratification of the population units by making use of the auxiliary information.

4. Sampling with Varying Probability

Under certain circumstances, selection of units with unequal probabilities provides more efficient estimators than equal probability sampling, and this type of sampling is known as unequal or varying probability sampling. In the most commonly used varying probability sampling scheme, the units are selected with probability proportional to a given measure of size (PPS) where the size measure is the value of an auxiliary variable x related to the characteristic y under study and this sampling scheme is termed as probability proportional to size sampling. For instance, the number of persons in some previous period may be taken as a measure of the size in sampling area units for a survey of socio-economic characters, which are likely to be related to population. Similarly, in estimating crop characteristics the geographical area or cultivated area for a previous period, if available, may be considered as a measure of size, or in an industrial survey, the number of workers may be taken as the size of an industrial establishment.

Since a large unit, that is, a unit with a large value for the study variable y, contributes more to the population total than smaller units, it is natural to expect that a scheme of selection which gives more chance of inclusion in a sample to larger units than to smaller units would provide estimators more efficient than equal probability sampling. Such a scheme is provided by pps sampling, size being the value of an auxiliary variable x directly related to y. It may appear that such a selection procedure would give biased estimators as the larger units are over-represented and the smaller units are under-represented in the sample. This would be so, if the sample means is used as an estimator of population mean. Instead, if the sample observations are suitably weighted at the estimation stage taking into consideration their probabilities of selection, it is possible to obtain unbiased estimators. Mahalanobis (1938) has referred to this procedure in the context of sampling plots for a crop survey and this procedure has been discussed in detail by Hansen and Hurwitz (1943).

5. Stratified Random Sampling

The basic idea in stratified random sampling is to divide a heterogeneous population into sub-populations, usually known as strata, each of which is internally homogeneous in which case a precise estimate of any stratum mean can be obtained based on a small sample from that stratum and by combining such estimates, a precise estimate for the whole population can be obtained. Stratified sampling provides a better cross section of the population than the procedure of simple random sampling. It may also simplify the organization of the field work. Geographical proximity is sometimes taken as the basis of stratification. The assumption here is that geographically contiguous areas are often more alike than areas that are far apart. Administrative convenience may also dictate the basis on which the stratification is made. For example, the staff already available in each range of a forest division may have to supervise the survey in the area under their jurisdiction. Thus, compact geographical regions may form the strata. If the characteristic under study is known to be correlated with a supplementary variable for which actual data or at least good estimates are available for the units in the population, the stratification may be done using the information on the supplementary variable. For instance, the volume estimates obtained at a previous inventory of the forest area may be used for stratification of the population.

In stratified sampling, the variance of the estimator consists of only the ‘within strata’ variation. Thus the larger the number of strata into which a population is divided, the higher, in general, the precision, since it is likely that, in this case, the units within a stratum will be more homogeneous. For estimating the variance within stratum, there should be a minimum of 2 units in each stratum. The larger the number of strata the higher will, in general, be the cost of enumeration. So, depending on administrative convenience, cost of the survey and variability of the characteristic under study in the area, a decision on the number of strata will have to be arrived at.

6. Cluster Sampling

A sampling procedure presupposes division of the population into a finite number of distinct and identifiable units called the sampling units. The smallest units into which the population can be divided are called the elements of the population, and group of elements the clusters. A cluster may be a class of students or cultivators' fields in a village. When the sampling unit is a cluster, the procedure of sampling is called cluster sampling.

For many types of population a list of elements is not available and the use of an element as the sampling unit is, therefore, not feasible. The method of cluster or area sampling is available in such cases. Thus, in a city a list of all the houses may be available, but that of persons is rarely so. Again, list of farms are not available, but those of villages or enumeration districts prepared for the census are. Cluster sampling is, therefore, widely practiced in sample surveys.

For a given number of sampling units cluster sampling is more convenient and less costly than simple random sampling due to the saving time in journeys, identification and contacts etc., but cluster sampling is generally less efficient than simple random sampling due to the tendency of the units in a cluster to be similar. In most practical situations, the loss in efficiency may be balanced by the reduction in the cost and the efficiency per unit cost may be more in cluster sampling as compares to simple random sampling.

7. Multistage Sampling

Cluster sampling is a sampling procedure in which clusters are considered as sampling units and all the elements of the selected clusters are enumerated. One of the main considerations of adopting cluster sampling is the reduction of travel cost because of the nearness of elements in the clusters. However, this method restricts the spread of the sample over population which results generally in increasing the variance of the estimator. In order to increase the efficiency of the estimator with the given cost it is natural to think of further sampling the clusters and selecting more number of clusters so as to increase the spread of the sample over population. This type of sampling which consists of first selecting clusters and then selecting a specified number of elements from each selected cluster is known as sub-sampling or two stage sampling, since the units are selected in two stages. In such sampling designs, clusters are generally termed as first stage units (fsu's) or primary stage units (psu's) and the elements within clusters or ultimate observational units are termed as second stage units (ssu's) or ultimate stage units (usu's). It may be noted that this procedure can be easily generalized to give rise to multistage sampling, where the sampling units at each stage are clusters of units of the next stage and the ultimate observational units are selected in stages, sampling at each stage being done from each of the sampling units or clusters selected in the previous stage. This procedure, being a compromise between uni-stage or direct sampling of units and cluster sampling, can be expected to be (i) more efficient than uni-stage sampling and less efficient than cluster sampling from considerations of operational convenience and cost, and (ii) less efficient than uni-stage sampling and more efficient than cluster sampling from the view point of sampling variability, when the sample size in terms of number of ultimate units is fixed.

It may be mentioned that multistage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is prohibitive or where the cost of locating and physically identifying the usu's is considerable. For instance, for conducting a socio-economic survey in a region, where generally household is taken as the usu, a complete and up-to-date list of all the households in the region may not be available, whereas a list of villages and urban blocks which are group of households may be readily available. In such a case, a sample of villages or urban blocks may be selected first and then a sample of households may be drawn from each selected village and urban block after making a complete list of households. It may happen that even a list of villages is not available, but only a list of all tehsils (group of villages) is available. In this case a sample of households may be selected in three stages by selecting first a sample of tehsils, then a sample of villages from each selected tehsil after making a list of all the villages in the tehsil and finally a sample of households from each selected village after listing all the households in it. Since the selection is done in three stages, this procedure is termed as three stage sampling. Here, tehsils are taken as first stage units (fsu's), villages as second stage units (ssu's) and households as third or ultimate stage units (tsu's).

8. Systematic Sampling

In all other sampling methods, the successive units (whether elements or clusters) are selected with the help of random numbers. But a method of sampling in which only the first unit is selected with the help of random number while the rest of the units are selected according to a pre-determined pattern, is known as systematic sampling. The systematic sampling has been found very useful in forest surveys for estimating the volume of timber, in fisheries surveys for estimating the total catch of fish, in milk yield surveys for estimating the lactation yield etc.

9. Conclusion

Simple random sampling and probability proportional size designs are most important uni-stage design. In most of the practical situations, complex sampling designs are utilized on the basis of these uni-stage sampling designs. Stratified random sampling, multistage sampling, multiphase sampling, etc. are efficient complex designs widely used in agricultural and socio-economic surveys.

References

Cochran, W.G. (1977). *Sampling techniques*. Wiley Eastern Ltd.

Des Raj, (1968). *Sampling theory*. Tata-Mcgraw-Hill Publishing Company Ltd.

Hansen, M.H. and Hurwitz, W.H. (1943). On the theory of sampling from finite populations.*Ann. Math. Statist.*, **14**, 333-362.

Hansen, M.H., Hurwitz, W.H. and Madow, W.G. (1993). *Sample survey methods and theory*. Vol. 1 and Vol. 2, John Wiley & Sons, Inc.

Murthy, M.N. (1977). *Sampling theory and methods*. Statistical Publishing Society.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984). *Sampling theory of surveys with applications*. Indian Society of Agricultural Statistics.

INTRODUCTION AND OVERVIEW OF THE NONLINEAR GROWTH MODEL

Mrinmoy Ray
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
mrinmoy.ray@icar.gov.in

1. Introduction

Growth is defined as "an irreversible increase in size and volume that occurs as a result of differentiation and distribution in the plant/animal." A model is a schematic representation of a system's conception, an act of mimicry, or a set of equations that represents a system's behaviour. A model is also defined as "a representation of an object, system, or idea in a form other than that of the entity itself." Its purpose is typically to aid in the explanation, comprehension, or improvement of a system's performance.

TYPES OF MODELS

Models are classified into different groups or types based on the purpose for which they are designed. Among them are a few:

a. Statistical models: These models describe the relationship between. Relationships are measured in a system using statistical techniques in these models. Example: regression model, Time series model, etc.

b. Mechanistic models: These models explain not only the relationship between variables, but also how these models work (explains the relationship of influencing dependent variables). Physical selection is the basis for these models.

c. Deterministic models: The exact value of the dependent variable is estimated using these models. These models have defined coefficients as well.

d. Stochastic models: Each output has a probability element attached to it. Different outputs, along with probabilities, are provided for each set of inputs. At a given rate, these models define the state of the dependent variable.

e. Dynamic models: Time is accounted for as a variable. Both dependent and independent variables have values that remain constant over a given time period.

f. Static: Time is not considered a variable. Dependent and independent variables with values that remain constant over time.

g. Simulation models: In general, computer models are mathematical representations of real-world systems. Crop simulation models' primary goal is to estimate agricultural production as a function of weather and soil conditions, as well as crop management. These models employ one or more sets of differential equations to compute rate and state variables over time, typically from planting to harvest maturity or final harvest.

Statistical Modelling

A fundamental problem in statistics is developing models based on a sample of observations and making inferences based on the model. Huge amounts of data pertaining to crop production/productivity, import-export of various agricultural commodities, and so on are being collected sequentially over time in almost all branches of agriculture, including animal sciences and fisheries. One feature of such data is that successive observations are dependent on one another. Each observation of the observed data series, Y_t,may be considered as a realization of a stochastic process $\{Y_t\}$, which is a family of random variables $\{Y_t,\ t \in T\}$, where $T = \{\ 0, \pm1, \pm2, \ldots\}$, and apply standard time-series approach to develop an ideal model which will adequately represent the set of realizations and also their statistical relationships in a satisfactory manner. Forecasting of time-series data is critical for planners and policymakers. Over the last few decades, a new field known as "Nonlinear time-series modelling" has emerged. There are essentially two approaches available here: parametric or nonparametric. Obviously, we should use the former if we are certain about the functional form in a given situation; otherwise, the latter may be used.

Parametric and Nonparametric Approaches

Regression analysis has grown in popularity as a tool for statistical modelling and data analysis over the last several decades. This information describes the relationship between a response variable and one or more predictor variables. The primary goal is to express the mean of the response as a function of the predictor variables. The general regression model takes the following form:

$$Y \quad = \quad m\,(X) + \ \varepsilon$$

Where Y is the response variable, $m(X) = E\ (Y|\ X)$ is the mean response or regression function and ε is the error. The regression function $m(X)$ is usually unknown and the objective is to obtain a suitable estimator of $m(X)$ using a sample of observations.

In the linear regression, it is assumed that the mean of the response variable Y is a linear function of predictor variable(s) X of the form

$$E\ (Y|\ X) \ = \ X\beta$$

i.e. m (X) is linear in parameters. The parameter vector β is usually estimated by the Method of least squares. In nonlinear regression, it is assumed that the mean of the response variable is a nonlinear function of the predictor variable (s) X of the form

$E(Y|X)=m(X,\beta)$

i.e.$m(X)$ is nonlinear in parameters. Generally, there will be no closed form expression for the estimates of β and iterative procedures are required for estimation of parameters.

A parametric regression model (linear or nonlinear) assumes that the form of m is known with the exception of some unknown parameters, and that the shape of the regression function is entirely dependent on the parameters. It is frequently difficult to guess the most appropriate functional form simply by looking at the data. There may be times when no suitable parametric form exists to express the regression function. In such cases, the nonparametric regression approach is very useful because it does not require strong assumptions about the shape of the regression function. A nonparametric regression model only assumes that m is part of an infinitely large collection of functions. One limitation of the preceding approach is that it generally relies on certain assumptions about the smoothness of the function being estimated, which may or may not be true in practice. As a result, the data under consideration may be over smoothed.

LINEAR MODEL

A mathematical model is an equation or set of equations that represents a system's behaviour. It can be 'linear' or 'nonlinear.' A linear model is one in which all of the parameters appear linearly.

NONLINEAR MODELS

Any type of statistical investigation in which principles from a body of knowledge are seriously considered in the analysis is likely to result in a 'Nonlinear model.' Such models are critical in understanding the complex interrelationships between variables. A ‘nonlinear model’ is one in which at least one of the parameters appears nonlinearly. More formally, in a ‘nonlinear model’, at least one derivative with respect to a parameter should involve that parameter.

- Examples of a nonlinear model are:

$$Y(t) = \exp(at+bt^2) \quad (1a)$$

$$Y(t) = at + \exp(-bt) \quad (1b)$$

Note. Some authors use the term ‘intrinsically nonlinear’ to indicate a nonlinear model which can be transformed to a linear model by means of some transformation. For example, the model given by Eq. (1a) is ‘intrinsically nonlinear’ in view of the transformation $X(t) = \log_e Y(t)$.

a. MALTHUS MODEL:

Thomas R. Malthus, an Englishman, proposed a mathematical model of population growth in 1798. Despite its simplicity, the model has become the foundation for most future modelling of biological populations. His essay, "An Essay on the Principle of Population," contains an excellent discussion of the limitations of mathematical modelling and should be required reading for all serious students of the subject. Malthus observed that, if not restrained by environmental or social constraints, human populations appeared to double every twenty-five years, regardless of initial population size. In other words, he proposed that populations increased by a fixed proportion over a given period of time and that, in the absence of constraints, this proportion was unaffected by population size. According to Malthus, if a population of 100 people increased to a population of 135 people over the course of, say, five years, then a population of 1000 people would increase to 1350 people over the same period of time. Malthus' model is an example of a one-variable, one-parameter model. The quantity we are interested in observing is referred to as a variable. They typically evolve over time. Parameters are quantities known to the modeller before the model is built. They are frequently constants, though a parameter can change over time. The variable in the Malthusian model is population, and the parameter is population growth rate.

If N(t) denotes the population size or biomass at time t and r is the intrinsic growth rate, then the rate of growth of population size is given by

$dN/dt = rN$

Therefore, $N(t) = N_o \exp(rt)$

Note : Malthus model can be used for describing growth of simplistic organisms, which begin to grow by binary splitting of cells.

Drawback: $N(t) \to \infty$ as $t \to \infty$, which cannot happen in reality.

Malthus predicted that unchecked population growth would quickly outstrip carrying capacity, resulting in overpopulation and social problems.

a. MONOMOLECULAR MODEL:

Because the monomolecular model assumes a carrying capacity of one, which means that the maximum level of disease is one, disease severity or incidence is measured as a proportion. Plant tissue that is diseased may only have a value between zero (healthy) and one (complete disease).It also assumes the absolute rate of change is proportional to the healthy tissue i.e., (1-*y*).

It describes growth progress in which it is assumed that the rate of growth at any point in time is proportional to the resources yet to be obtained, i.e.

$dN/dt = r(K–N)$,

where K is the carrying capacity.

or $N(t) = K– (K–N_o) \exp(–rt)$

Drawback: No point of inflexion.

a. LOGISTIC MODEL:

Logistic model was developed by Belgian mathematician Pierre Verhulst (1838) who suggested that the rate of population increase may be limited, i.e., it may depend on population density. Population growth rate declines with population numbers, N, and reaches 0 when N = K. Parameter K is the upper limit of population growth and it is called carrying capacity. It is commonly interpreted as the amount of resources expressed in the number of organisms that these resources can support. If the population exceeds K, the population growth rate becomes negative and the population decreases.

The differential equation represents this model:

$dN/dt = rN (1–N/K)$ (1)

Therefore, $N(t) = K/[1+(K/N_o–1) \exp(–rt)]$. The graph of N(t) versus t is elongated S-shaped and the curve is symmetrical about its point of inflexion.

a. GOMPERTZ MODEL

This is another model with sigmoid behaviour that has been found to be quite useful in biological work. Benjamin Gompertz developed the Gompertz curve to estimate human mortality (Gompertz, B. "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies." Phil. Trans. Roy. Soc. London 123, 513-585, 1832). An early description of the use of this equation to describe growth processes is given by

CharlesWinsor (1932). However, unlike the logistic model, this does not have a symmetric point of inflexion.

This model's differential equation is

$dN/dt = rN \log_e (K/N)$ (2)

or $N(t) = K \exp[\log_e (N_o / K) \exp(-rt)]$

a. RICHARDS MODEL:

The Richards curve, also known as generalised logistic, is a popular growth model that can fit a wide range of S-shaped growth curves. Both 4 and 5 parameter versions are commonly used. The logistic curve is symmetrical about its point of inflection. Richards (1959) introduced an additional parameter to deal with asymmetrical growth curves.

This model is given by

$$N(t) = K N_o / [N_o + (K^m - N_o^m) \exp(-rt)]^{1/m} \quad . \qquad (4)$$

However, unlike the earlier models, this model has four parameters.

Drawback. Number of parameters is more.

a. MIXED-INFLUENCE MODEL:

This is a mixture of ‘Monomolecular’ and ‘Logistic’ Models. It is given by

$dN/dt = r (K-N) + s N (1-N/K)$,

FITTING OF NONLINEAR MODELS

The models presented above have been posed deterministically. This is obviously unrealistic, so we replace these deterministic models with statistical models by including an error term on the right hand side and making appropriate assumptions about them. This produces a 'Nonlinear statistical model.' The 'Method of least squares' can be used to estimate parameters in non-linear regression, just as it can in linear regression. However, minimising the residual sum of squares produces normal equations with nonlinear parameters. Because exact solutions to nonlinear equations are not possible, iterative procedures are used to obtain approximate analytic solutions.

- Four main methods of this kind are:

  i) Linearization (or Taylor Series) method

  ii) Steepest Descent method

  iii) Levenberg-Marquardt’s method

  iv) Do not use Derivatives method

Draper and Smith discuss the specifics of these methods, as well as their benefits and drawbacks (1998). Neither the Linearization nor the Steepest descent methods are perfect. The Levenberg-Marquardt method is the most widely used method for computing nonlinear least squares estimates. This method is a compromise between the other two methods, successfully combining the best features of both while avoiding their significant disadvantages. It's good because it almost always converges and doesn't' slow down' at the end of the iterative process.

CHOICE OF INITIAL VALUES

All nonlinear estimation procedures require initial parameter values, and selecting good initial values is critical. There is, however, no standard procedure for obtaining preliminary estimates. The use of prior information is the most obvious method for making initial guesses. Estimates based on previous experiments, known values for similar systems, and values derived from theoretical considerations all combine to form ideal first guesses.

Some other methods are:

(i) Linearization:

After ignoring the error term, check the form of the model to see if it could be transformed into a linear form by means of some transformation. In such cases, linear regression can be used to obtain initial values.

(ii) Solving a system of equations:

If there are p parameters, substitute for p sets of observations into the model ignoring the error. Solve these equations for the parameters, if possible. Widely separated x_i often work best.

R code

Monomolecular growth model

```
z=read.csv(file.choose(), header=TRUE)
head(z)
kk=data.frame(z)
grz1=nls(y~k-(k-y0)*exp(-r*t),data=kk,  start=list(k=1 ,y0=0.03,r=0.1))
summary(grz1)
 fitted=kk$y-resid(grz1)
kkk=data.frame(fitted)
MSE.nn<- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors<- c("blue","red")
```

```
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,1),axes=FALSE, ann=FALSE)
axis(1, at=1:20, lab=c(0:19))
axis(2, las=1, at=0.2*0:5)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22, lty=1:2);
zz=resid(grz1)
predicted= 0.99651-(0.99651-0.08844)*exp(-0.26727*20)
```

Gompertz model

```
z=read.csv(file.choose(), header=TRUE)
 head(z)
 kk=data.frame(z)
gr1=nls(y~k*exp(log(y0/k)* exp(-r*t)),data=kk,  start=list(k=50,y0=11.72,r=0.1))
summary(gr1)
fitted=kk$y-resid(gr1)
kkk=data.frame(fitted)
MSE.nn<- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors<- c("blue","red")
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,35),axes=FALSE, ann=FALSE)
axis(1, at=1:38, lab=c(0:37))
axis(2, las=1, at=5*0:8)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22, lty=1:2);
```

logistic model

```
z=read.csv(file.choose(), header=TRUE)
```

```
 head(z)
 kk=data.frame(z)
gr2=nls(y~k/(1+(k/y0-1)* exp(-r*t)), data=kk,  start=list(k=50,y0=11.72,r=0.1))
summary(gr2)
fitted=kk$y-resid(gr2)
kkk=data.frame(fitted)
MSE.nn<- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors<- c("blue","red")
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,35),axes=FALSE, ann=FALSE)
axis(1, at=1:38, lab=c(0:37))
axis(2, las=1, at=5*0:8)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22, lty=1:2);
```

LOGIT AND PROBIT ANALYSIS

Himadri Shekhar Roy
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
himadri.roy@icar.gov.in

1. Introduction

Regression analysis is a technique used to examine the relationships between variables. These relationships are expressed through equations or models that connect a response or dependent variable with one or more explanatory or predictor variables. Typically, the variables involved in regression analysis are quantitative in nature. The estimation of parameters in this type of analysis relies on four key assumptions. The first assumption is that the response variable is linearly related to the explanatory variables. In other words, there is a linear relationship between the dependent variable and the predictors. The second assumption is that the errors in the model are independently and identically distributed, following a normal distribution with a mean of zero and a common variance. This assumption ensures that the errors are random and have a consistent distribution. The third assumption assumes that the explanatory variables are measured without any errors. This means that the predictor variables are accurate and reliable. The last assumption relates to the equal reliability of observations. It assumes that each observation used in the analysis is equally reliable and contributes equally to the analysis. In cases where the response variable in the model is qualitative, instead of directly modeling the response variable itself, probabilities of belonging to different categories can be modelled using the same regression framework. However, this approach comes with additional constraints and assumptions for multiple regression models. The first constraint is that probabilities range between 0 and 1, while the right-hand side function in multiple regression models is unbounded. This means that adjustments need to be made to ensure that the predicted probabilities remain within the valid range. The second constraint is related to the error term of the model. In this case, the error term can only take limited values, and the variance of the errors is not constant but depends on the probability of the response variable falling into a particular category. There are several notable references available that provide a comprehensive overview of logistic regression, such as the works of Fox (1984) and Klienbaum (1994). For Probit analysis, a useful resource is Finney (1971).

2. Assumptions of Linear Regression Model if Response is Qualitative

To illustrate the limitations of using linear regression when the response variable is qualitative, let's examine a simple linear regression model that involves a single predictor variable and a binary response variable.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \ , \ i = 1, 2, \ldots, n$$

where, the outcome Y_i is binary (taking values 0,1), $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and are independent and n is the number of observations.

Let π_i denote the probability that Y_i =1 when X_i = x, i.e.

$$\pi_i = P(Y_i = 1 | X_i = x) = P(Y_i = 1)$$

thus $\quad P(Y_i = 0) = 1 - \pi_i$.

Under the assumption $E(\varepsilon_i) = 0$, the expected value of the response variable is

$$E(Y_i) = 1.(\pi_i) + 0.(1-\pi_i) = \pi_i$$

If the response is binary, then the error terms can take on two values, namely,

$$\varepsilon_i = 1 - \pi_i \qquad \text{when } Y_i = 1$$

$$\varepsilon_i = -\pi_i \qquad \text{when } Y_i = 0$$

Because the error is dichotomous (discrete), normality assumption is violated. Moreover, the error variance is given by:

$$\begin{aligned} V(\varepsilon_i) &= \pi_i (1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2 \\ &= \pi_i (1-\pi_i) \end{aligned}$$

It can be seen that variance is a function of π_i's and it is not constant. Therefore, the assumption of homoscedasticity (equal variance) does not hold.

3. Logistic regression

3.1 Binary Logistic regression

Logistic regression is often recommended when the multivariate normality assumption is not met by the independent variables and the response variable is qualitative. This situation, where the response variable is qualitative and the independent variables include a mix of categorical and continuous variables, is commonly encountered in statistical applications such as agriculture and medical science. The binary logistic regression model, developed by researcher Cox in the late 1950s, is the preferred statistical model for analysing binary (dichotomous) responses. Agricultural data often exhibit sigmoidal or elongated S-shaped curves, making logistic regression models more appropriate. These models can capture non-linear relationships between the response variable and the qualitative and quantitative factors that influence it. Logistic regression addresses similar questions as discriminant function analysis and multiple regression, but it does not rely on distributional assumptions for the predictors. In other words, the predictors do not need to follow a normal distribution, the relationship between the response and predictors can be non-linear, and the observations do not need to have equal variance in each group. For a comprehensive understanding of logistic regression, informative resources can be found in the works of Fox (1984) and Kleinbaum (1994).

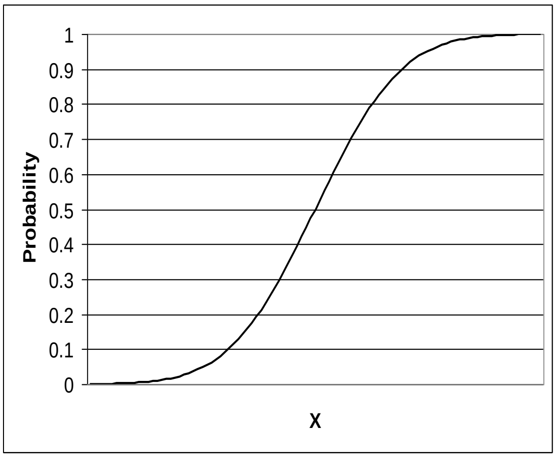
The issue of non-normality and heteroscedasticity, as discussed in section 2, renders least square estimation unsuitable for the linear probability model. When attempting to use weighted least square estimation as an alternative, the resulting fitted values may not be constrained within the interval (0, 1), making them inappropriate for interpretation as probabilities. Furthermore, there is a possibility of negative error variances arising. To address this problem, one solution is to constrain the values of π (the response variable) to the unit interval while still maintaining the linear relationship between π and the regressor X within that interval. By doing so, we can ensure that the predicted values of π remain within the valid range of probabilities.

$$\pi = \begin{cases} 0 & ,\beta_0 + \beta_1 X < 0 \\ \beta_0 + \beta_1 X & ,0 \le \beta_0 + \beta_1 X \le 1 \\ 1 & ,\beta_0 + \beta_1 X > 1 \end{cases}$$

However, this constrained linear probability model has certain unattractive features such as abrupt changes in slope at the extremes 0 and 1 making it hard for fitting the same on data. A smoother relation between π and X is generally more sensible. To correct this problem, a positive monotone (i.e. non-decreasing) function is required to transform ($\beta_0 + \beta_1 x_i$) to unit interval. Any cumulative probability distribution function (CDF) P, meets this requirement. That is, respecify the model as $\pi i = P\ (\beta_0 + \beta_1 x_i)$. Moreover, it is advantageous if P is strictly increasing, for then, the transformation is one-to-one, so that model can be rewritten as $P^{-1}(\pi i) = (\beta 0 + \beta 1xi)$, where P^{-1} is the inverse of the CDF P. Thus the non-linear model for itself will become both smooth and symmetric, approaching $\pi = 0$ and $\pi = 1$ as asymptotes. Thereafter maximum likelihood method of estimation can be employed for model fitting.

3.2 Properties of Logistic Regression Model

The logistic response function exhibits a characteristic S-shaped curve, which can be visualized in the accompanying figure. As X increases, the probability π initially experiences a gradual increase, followed by a rapid acceleration. Eventually, the increase in probability tapers off and stabilizes, but it never exceeds the value of 1.



The shape of the S-curve can be reproduced if the probabilities can be modeled with only one predictor variable as follows:

$\pi = P(Y=1|X=x) = 1/(1+e^{-z})$

where $z = \beta_0 + \beta_1 x$, and e is the base of the natural logarithm. Thus for more than one (say r) explanatory variables, the probability π is modeled as

$\pi = P(Y=1|X_1 = x_1 ... X_r = x_r)$

$=1/(1+e^{-z})$

where $z = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r$.

This equation is called the logistic regression equation. It is nonlinear in the parameters β_0, β_1… β_r. Modeling the response probabilities by the logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression. The method of estimation generally used is the maximum likelihood estimation method.

To explain the popularity of logistic regression, let us consider the mathematical form on which the logistic model is based. This function, called f (z), is given by

$$f(z) = 1/(1+e^{-z}), \ -\infty < z < \infty$$

Now when z = -∞, f (z) =0 and when z = ∞, f (z) =1. Thus the range of f (z) is 0 to1. So the logistic model is popular because the logistic function, on which the model is based, provides. Estimates that lie in the range between zero and one.

An appealing S-shaped description of the combined effect of several explanatory variables on the probability of an event.

3.6 Multinomial logistic regression modeling

Let **X** is a vector of explanatory variables and π denotes the probability of binary response variable then logistic model is given by

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \mathbf{X}\beta = g(\pi)$$

where, 'alpha' is the intercept parameter and 'beta' is a vector of slope parameters. In case response variable has ordinal categories say 1,2,3,--------, I, I+1 then generally logistic model is fitted with common slope based on cumulative probabilities of response categories instead of individual probabilities. This provides parallel lines of regression model with following form

g [Prob ($\mathbf{y} \le \mathbf{i}(\mathbf{x})$)] $\bar{\alpha}_i + \chi\beta$, $1 \le i \le I$

where, $\alpha_1, \alpha_2, ------- \alpha_\mathbf{k}$, are k intercept parameters and β is the vector of slope parameters.

Multinomial logistic regression (taking qualitative response variable with three categories, for simplicity) is given by

logit[Pr(Y ≤ j – 1 / **X**)] = $\alpha_j + \boldsymbol{\beta}^T \mathbf{X}$, j = 1,2

where α_j are two intercept parameters ($\alpha_1 < \alpha_2$), $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots\ldots., \beta_k)$ is the slope parameter vector not including the intercept terms, $\mathbf{X}^T = (X_1, X_2, \ldots., X_k)$ is vector of explanatory variables. This model fits a common slope cumulative model i.e. 'parallel lines' regression model based on the cumulative probabilities of the response categories.

$$\text{logit}(\pi_1) = \log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots.. + \beta_k X_k,$$

$$\text{logit}(\pi_1 + \pi_2) = \log\left(\frac{\pi_1 + \pi_2}{1-\pi_1-\pi_2}\right) = \alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots.. + \beta_k X_k$$

π_j (X) denotes classification probabilities Pr(Y=j-1 / X) of response variable Y, j = 1,2,3, at X^T.

These models can be fitted through maximum likelihood procedure.

4. Probit analysis

4.1 Introduction

Probit analysis is widely utilized in various fields when the response variable is qualitative. One of its main applications is observed in toxicological studies, where it transforms the sigmoid dose-response curve into a linear relationship that can be analyzed using regression techniques like least squares or maximum likelihood. In essence, probit analysis is a methodology that converts the complex relationship between the percentage affected and the dose response into a linear relationship between probit and the dose response. The probit values can then be translated back into percentages. This approach is appropriate because of the typical shape exhibited by dose-response curves. While the method is approximate, it enables the quantification of consequences resulting from exposure. The term "probit" originates from the phrase "probability unit" and was coined by Bliss. It was the first model developed and studied for analyzing data such as the percentage of pests killed by a pesticide.

4.2 Probit Model

In the realm of probability theory and statistics, the probit function represents the inverse of the cumulative distribution function (CDF) linked to the standard normal distribution. Alternatively, one can consider the logistic distribution, which results in the logit or logistic model. Both the logistic and probit curves are highly similar, producing almost indistinguishable outcomes. In practice, they provide estimated probabilities that exhibit very little variation (Aldrich and Nelson, 1984). The selection between the logistic and probit approaches is primarily based on practical preferences and prior experience.

For the standard normal distribution N (0, 1), the CDF is commonly denoted by Φ (*z*) (continuous, monotone increasing sigmoid function) given by,

$$\Phi(z) = \mathrm{P}(Z \le z) = \int_{-\infty}^{\mathrm{z}} \varphi(u)\mathrm{du} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathrm{z}} e^{-\frac{\mathrm{u}^2}{2}} \mathrm{du}$$

As an example, considering the familiar fact that the N (0, 1) distribution places 95% of probability between -1.96 and 1.96, and is symmetric about zero, it follows that

$$\Phi(-1.96) = 0.025 = 1 - \Phi(1.96)$$

The probit function gives the 'inverse' computation, generating a value of an N (0, 1) random variable, associated with specified cumulative probability. Formally, the probit function is the inverse of Φ (z), denoted by Φ^{-1}(p). Continuing the example,

$$\Phi^{-1}(0.025) = -1.96 = -\Phi^{-1}(0.975)$$

In general,

$$\Phi(\text{probit}(p)) = p \text{ and } \text{probit}(\Phi(z)) = z$$

In statistics, a probit model is a popular specification of a generalized linear model. If Y be a binary response variable, and let X be the single predictor variable, then the probit model assumes that,

$$P(Y_i = 1|X_i = x) = \Phi(\alpha+\beta x_i)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta x_i} e^{-\frac{1}{2}z^2} dz$$

where Φ is the CDF of the standard normal distribution. The parameters β are estimated by maximum likelihood.

In any dose-response scenario, there are two key components: the stimulus (such as a vitamin, drug, mental test, or physical force) and the subject (which could be an animal, plant, human volunteer, etc.). The stimulus is administered to the subject at a specific dose or intensity, measured in units such as concentration, weight, time, or other appropriate metrics, within a controlled environmental setting. Consequently, the subject exhibits a response. The response in this context is quantal, meaning it can either occur or not occur depending on the intensity of the stimulus. Under controlled conditions, a response is observed when the stimulus intensity surpasses a certain threshold or limen. However, the term "tolerance" is now more commonly used to refer to this value. The tolerance value varies among individuals within the population being studied. For quantal response data it is therefore necessary to consider distribution of tolerance over the population studied. If the dose or intensity of stimulus is measured by z, the distribution of tolerance may be expressed by $dP = f(z)dz$.

5. Classificatory ability of the models

There are several different classification accuracy measures that are commonly used to assess the performance of a classification model. Here are a few examples:

Accuracy: This is the most basic measure and represents the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances.

Precision: Precision focuses on the proportion of correctly classified positive instances (true positives) out of all instances predicted as positive (true positives plus false positives). It measures the model's ability to avoid false positives.

Recall (Sensitivity or True Positive Rate): Recall calculates the proportion of correctly classified positive instances (true positives) out of all actual positive instances (true positives plus false negatives). It quantifies the model's ability to capture true positives and avoid false negatives.

Specificity (True Negative Rate): Specificity evaluates the proportion of correctly classified negative instances (true negatives) out of all actual negative instances (true negatives plus false positives). It measures the model's ability to identify true negatives and avoid false positives.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that combines both precision and recall into a single value, useful when there is an imbalance between positive and negative instances.

Area Under the Receiver Operating Characteristic curve (AUC-ROC): The AUC-ROC measure quantifies the overall performance of a binary classifier by considering the trade-off between true positive rate (sensitivity) and false positive rate across different classification thresholds. It provides a single value that represents the model's ability to distinguish between positive and negative instances.

References:

Finney, D.J. (1971). *Probit Analysis* (3rd edition). Cambridge University Press, Cambridge, England.

Fox, J. (1984). *Linear statistical models and related methods with application to social research*, Wiley, New York.

Kleinbaum, D.G. (1994). *Logistic regression*: A self learning text, New York: Springer.

OVERVIEW OF TIME SERIES ANALYSIS

Mrinmoy Ray
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
mrinmoy.ray@icar.gov.in

Introduction:

A data set containing sequence of observations on a single phenomenon observed over time is called time-series data. In time series, past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship.

Time Series Components:

Trend: A trend exists when there is a long-term increase or decrease in the data. It does not have to be lin-ear. Some-times we will refer to a trend "changing direction" when it might go from an increas-ing trend to a decreasing trend.

Seasonal: A seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a fixed and known period.

Cyclic: A cyclic pattern exists when data exhibit rises and falls that are *not of fixed period*. The duration of these fluctuations is usually of at least 2 years.

Irregular component: Unobserved component exhibit in a time series

Exponential Smoothing Methods:

This method is suit-able for forecasting data with no trend or seasonal pattern. For exam-ple, the data in fig-ure do not dis-play any clear trend-ing behav-iour or any sea-son-al-ity, although the mean of the data may be chang-ing slowly over time. Simple moving average method assigns equal weights (1/k) to all k data points. Arguably, recent observations provide more information than do observations in the past. Exponential smoothing methods give larger weights to more recent observations, and the weights decrease exponentially as the observations become more distant. These methods are most effective when the parameters describing the time series are changing slowly over time

Types

- Simple exponential smoothing
- Holt's trend corrected exponential smoothing
- Holt-Winters method

Simple Exponential Smoothing (SES):

The SES method is used forecasting a time series when there is no trend or seasonal pattern, but the mean (or level) of the time series y_t is slowly changing over time

No trend model: $y_t = \beta_0 + \varepsilon_t$

Steps for SES method:

1. Compute the initial estimate of the mean (or level) of the series at time period $t = 0$

$$l_0 = \bar{y} = \sum_{t=1}^{n} y_t \big/ n$$

2. Compute the updated estimate by using the smoothing equation

$$\ell_T = \alpha y_T + (1-\alpha)\ell_{T-1}$$

where α is a smoothing constant between 0 and 1

Note that,

$$\ell_T = \alpha y_T + (1-\alpha)\ell_{T-1}$$

$$= \alpha y_T + (1-\alpha)[\alpha y_{T-1} + (1-\alpha)\ell_{T-2}]$$

$$= \alpha y_T + (1-\alpha)\alpha y_{T-1} + (1-\alpha)^2 \ell_{T-2}$$

$$= \alpha y_T + (1-\alpha)\alpha y_{T-1} + (1-\alpha)^2 \alpha y_{T-2} + \ldots + (1-\alpha)^{T-1}\alpha y_1 + (1-\alpha)^T \ell_0$$

Holt's Trend Corrected Exponential Smoothing

- A smoothing approach for forecasting such a time series that employs two smoothing constants, denoted by αandγ.
- There are two estimates ℓ_{T-1} and b_{T-1}
- ℓ_{T-1} is the estimate of the level of the time series constructed in time period T–1 (This is usually called the permanent component).
- b_{T-1} is the estimate of the growth rate of the time series constructed in time period T–1 (This is usually called the trend component).
- Level estimate

$$\ell_T = \alpha y_T + (1-\alpha)(\ell_{T-1} + b_{T-1})$$

- Trend estimate

$$b_T = \gamma(\ell_T - \ell_{T-1}) + (1-\gamma)b_{T-1}$$

where α = smoothing constant for the level (0 ≤ α ≤ 1)

γ = smoothing constant for the trend (0 ≤ γ ≤ 1)

Holt-Winters Method

- Estimate of the level

$$\ell_T = \alpha(y_T / sn_{T-L}) + (1-\alpha)(\ell_{T-1} + b_{T-1})$$

- Estimate of the growth rate or trend

$$b_T = \gamma(\ell_T - \ell_{T-1}) + (1-\gamma)b_{T-1}$$

- Estimate of the seasonal factors

$$sn_T = \delta(y_T / \ell_T) + (1-\delta)sn_{T-L}$$

where α, γ, and δ are smoothing constants between 0 and 1, *L* = number of seasons in a year (*L* = 12 for monthly data, and *L* = 4 for quarterly data)

ARIMA Model:

Auto Regressive Integrated Moving Average (ARIMA) is a prediction model for time series analysis and forecasting

- **Here the terms indicate:**

Auto Regressive: lags of variables itself

Integrated: Differencing steps required to make time series stationary

Moving Average: lags of previous information shocks

- **ARIMA model is denoted as ARIMA(*p,d,q*)**

where

p=number of autoregressive terms

d=number of non-seasonal differences needed to make time series stationary

q=number of lagged forecast errors in the prediction equation

For ARIMA model building process there is a minimum of 30 data points required

In an autoregressive integrated moving average model, the future value of a variable is assumed to be a linear function of several past observations and random errors. The underlying process that generate the time series has the form

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}$$

where, and are the actual and random error at time period t, respectively; (i= 1, 2, …, p) and (*j= 1, 2, …, q*) are model parameters *p* and *q* are integers and often referred to as orders of the model

Random errors are assumed to be independently and identically distributed with a mean zero and a constant variance of σ^2

If *q*= 0, then the above equation becomes an AR model of order *p*. When *p*= 0, the model reduces to an MA model of order *q*.One central task of the ARIMA (*p, d, q*) model building is to determine the appropriate model order (*p, q*) where *d* is the order of differencing.

ANN approach to time series forecasting:

In the domain of time series analysis, the inputs are typically the past observations series and the output is the future value. The ANN performs the following nonlinear function mapping between the input and output

$$y_t = f(y_{t-1} + y_{t-2}, ..., y_{t-p}, w) + \varepsilon_t$$

where, w is a vector of all parameters and f is a function of network structure and connection weights. Therefore, the neural network resembles a nonlinear autoregressive model.

Single hidden layer multilayer feed forward network is the most popular for time series modeling and forecasting. This model is characterized by a network of three layers of simple processing units. The first layer is input layer, the middle layer is the hidden layer and the last layer is output layer.

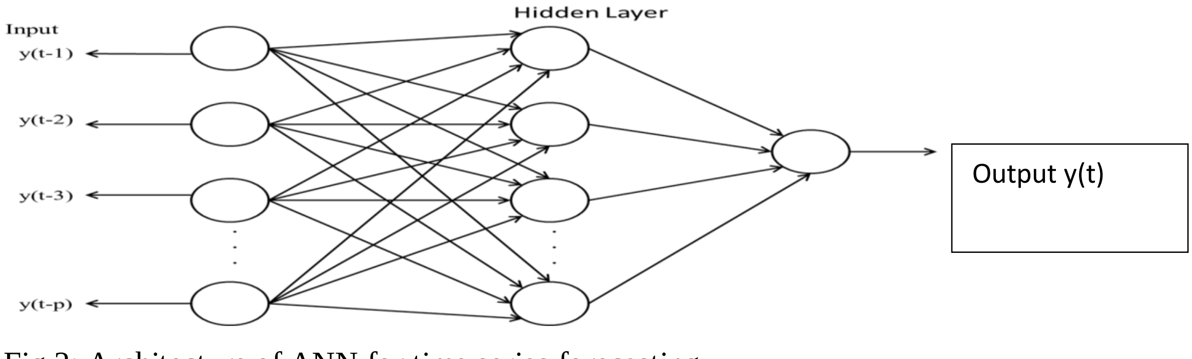


Fig 2: Architecture of ANN for time series forecasting

The relationship between the output (y_t) and the inputs ($y_{t-1}, y_{t-2}, ..., y_{t-p}$) can be mathematically represented as follows:

$$y_t = f\left(\sum_{j=0}^{q} \omega_j g\left(\sum_{i=0}^{p} \omega_{ij} y_{t-i}\right)\right)$$

where, $\omega_j (j = 0,1,2, \ldots\ldots, q)$ and $\omega_{ij} (i = 0,1,2, \ldots\ldots, p,\ j = 0,1,2, \ldots\ldots, q)$ are the model parameters often called the connection weights, p is the number of input nodes and q is the number of hidden nodes, g and f denote the activation function at hidden and output layer respectively. Activation function defines the relationship between inputs and outputs of a network in terms of degree of the non-linearity. Most commonly used activation functions are as follows-

Activation function	**Equation**
Identity	x
Sigmoid	$\frac{1}{1+e^{-x}}$
TanH	$\tanh(x) = \frac{2}{1+e^{-2x}} - 1$
ArcTan	$\tan^{-1}(x)$
Sinusoid	$\sin(x)$
Gaussian	e^{-x^2}

For time series forecasting sigmoid activation function is employed in hidden layer and identity activation function is employed in the output layer.

The selection of appropriate number of hidden nodes as well as optimum number of lagged observation p for input vector is important in ANN modeling for determination of the autocorrelation structure present in a time series. Though there are no established theories available for the selection of p and q, hence experiments are often conducted for the determination of the optimal values of p and q. The connection weights of ANNs are determined by learning method. There are three common learning algorithms for ANN –

1) Supervised Learning

The supervised learning strategy consists of having available the desired outputs for a given set of input signals; in other words, each training sample is composed of the input signals and their corresponding outputs. Henceforth, it requires a table with input/output data, also called attribute/value table, which represents the process and its behavior.

2) Unsupervised Learning

Different from supervised learning, the application of algorithm based on unsupervised learning does not require any knowledge of the respective desired outputs. Thus, the network needs to organize itself when there are existing particularities between the elements that compose the entire sample set, identifying subsets (or clusters) presenting similarities. The learning algorithm adjusts the synaptic weights and thresholds of the network in order to reflect these clusters within the network .itself.

3) Reinforcement Learning

It is the hybrid of supervised and unsupervised learning.

For time series forecasting supervised learning approach is utilized. Gradient decent back propagation algorithm is one of the popular approach of supervised learning.

Gradient decent back propagation algorithm

The objective of training is to minimize the error function that measures the misfit between the predicted value and the actual value. The error function which is widely used is mean squared error which can be written as:

$$E = \frac{1}{N}\sum_{n=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{n=1}^{N}\left\{y_t - f\left(\sum_{j=0}^{q}\omega_j g\left(\sum_{i=0}^{p}\omega_{ij} y_{t-i}\right)\right)\right\}^2$$

Where N is the total number of error terms. The parameters of the neural network are ω_j and ω_{ij} estimated by iteration. Initial connection weights are taken randomly from uniform distribution. In each iteration the connection weights changed by an amount $\Delta\omega_j$

$$\Delta\omega_j(t) = -\eta\frac{\partial E}{\partial\omega_j} + \delta\Delta\omega_j(t-1)$$

where, η is the learning rate and $\frac{\partial E}{\partial\omega_j}$ is the partial derivative of the function E with respect to the weight ω_j. δ is the momentum rate. The $\frac{\partial E}{\partial\omega_j}$ can be represented as follows-

$$\frac{\partial E}{\partial w_j} = -e_j(n)\times f'(x)\times y_j(n)$$

where $e_j(n)$ is the residual at n^th^ iteration

$f'(x)$= derivative of the activation function in the output layer. As in time series forecasting the activation function in the output layer is identity function hence $f'(x)$ =1. $y_j(n)$ is the desired output. Now connection weights in from input to hidden nodes changed by an amount $\Delta\omega_{ij}$

$$\Delta\omega_{ij}(t) = -\eta\frac{\partial E}{\partial\omega_{ij}} + \delta\Delta\omega_{ij}(t-1)$$

where

$$\frac{\partial E}{\partial w_{ij}} = g'(x)\times\sum_{j=0}^{q} e_j(n) * w_j(n)$$

where $g'(x)$ is the activation function in the hidden layer. For sigmoid activation function

$$g'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

Learning rate is user defined parameter known as tuning parameter of neural network which determine how slow or fast the optimal weight is obtained. The learning rate must be set small enough to avoid divergence. The momentum term prevents the learning process from setting in a local minimum. Though there are no established theories available for the selection of learning rate and momentum, hence experiments are often conducted for the determination of the learning rate and momentum.

Step by Step Modeling Procedure:

1. Testing of Nonlinearity:

As ANNs is suitable for nonlinear time series forecasting. Hence, prior to application of ANN the nonlinearity should be check. There are several tests for checking nonlinearity. BDS (Brock-Dechert-Scheinkman) test is of the popular approach for checking nonlinearity. This test utilizes the concept of spatial correlation from chaos theory. The computational procedure is given as follows

i) Let the considered time series is

$$\{x_i\} = [x_1, x_2, x_3, ..., x_N]$$

ii) The next step is to specify a value of m (embedding dimension), embed the time series into m dimensional vectors, by taking each m successive points in the series. This transforms the series of scalars into a series of vectors with overlapping entries

$$x_1^m = (x_1, x_2, ..., x_m)$$
$$x_2^m = (x_2, x_3, ..., x_{m+1})$$
$$.$$
$$.$$
$$.$$
$$x_{N-m}^m = (x_{N-m}, x_{N-m+1}, ..., x_N)$$

iii) In the third step correlation integral is computed, which measures the spatial correlation among the points, by adding the number of pairs of points (*i, j*), where 1≤ *i* ≤ N and 1≤ *j*≤N , in the m-dimensional space which are "close" in the sense that the points are within a radius or tolerance ε of each other.

$$C_{\varepsilon,m} = \frac{1}{N_m(N_m - 1)} \sum_{i \neq j} I_{i,j;\varepsilon}$$

where $I_{i,j;\varepsilon}$= 1 if $\left\| x_i^m - x_j^m \right\| \leq \varepsilon$

= 0 otherwise

iv) If the time series is i.i.d. then $C_{\varepsilon,m} \approx [C_{\varepsilon,1}]^m$

v) The BDS test statistics is as follows

$$BDS_{\varepsilon,m} = \frac{\sqrt{N}[C_{\varepsilon,m} - (C_{\varepsilon,1})^m]}{\sqrt{V_{\varepsilon,m}}}$$

where $V_{\varepsilon,m} = 4[K^m + 2\sum_{j=1}^{m-1} K^{m-j} C_\varepsilon^{\ 2j} + (m-1)^2 C_\varepsilon^{\ 2m} - m^2 K C_\varepsilon^{\ 2m-2}]$

$$K = K_\varepsilon = \frac{6}{N_m(N_m - 1)(N_m - 2)} \sum_{i<j<N} h_{i,j,N;\varepsilon}$$

$$h_{i,j,N;\varepsilon} = \frac{[I_{i,j;\varepsilon} I_{j,N;\varepsilon} + I_{i,N;\varepsilon} I_{N,j;\varepsilon} + I_{j,i;\varepsilon} I_{i,N;\varepsilon}]}{3}$$

The choice of m and ε depends on number of data. The null hypothesis is data are independently and identically distributed (i.i.d) against the alternative hypothesis the data are not i.i.d.; this implies that the time series is non-linearly dependent. BDS test is a two-tailed test; the null hypothesis should be rejected if the BDS test statistic is greater than or less than the critical values.

2. Division of the data:

Data is divided into training and test sets. The training sample is used for ANN for model development and the test sample is utilized to evaluate the forecasting performance. Sometimes a third one called the validation sample is also utilized to avoid the over fitting problem or to determine the stopping point of the training process. It is common to use one test set for both validation and testing purposes particularly for small data sets. The literature suggests little guidance in selecting the training and testing sets. Most commonly used rule are 90% vs. 10%, 80% vs. 20% or 70% vs. 30%, etc.

3. Data Normalization:

Nonlinear activation functions such as the sigmoid function typically have the squashing role in restricting the possible output from a node to, typically, (0, 1). Hence, data normalization is done prior to training process begins.

Normalization procedure

Linear transformation to [0,1]: $X_n=(X_0-X_{min})/(X_{max}-X_{min})$

Statistical normalization: $X_n=(X_0-\text{mean}(X))/\text{var}(X)$

simple normalization: $X_n=X_0/X_{max}$

4. Selection of appropriate number of hidden nodes as well as optimum number of lagged:

There are no established theories available for the selection of *p* and *q*, hence experiments are often conducted for the determination of the optimal values of *p* and *q*.

5. Estimation of connection weights:

Estimation of connection weights are determined by learning algorithm. For time series forecasting most commonly used learning approach is gradient decent back propagation algorithm.

6. Evaluating forecasting Performance

Forecasting performance can be computed by several approaches. Some of the approaches are given below-

$$MAPE=\frac{1}{n}\sum_{t=1}^{n}\left|y_t-\hat{y}_t\right|/y_t\times 100$$

$$MSE=\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}$$

where n is the total number of forecast values. y_t is the actual value at period t and $\hat{y}_t$ is the corresponding forecast value. The model with less MAPE/MSE/RMSE is preferred for forecasting purposes.

Limitations of ANN for time series forecasting:

i) ANNs are nonlinear time series model hence, for linear time series data the approach may not be better than linear statistical model.

ii) ANNs are black-box methods. There is no exact form to describe and analyze the relationship between inputs and outputs. This causes troublesome for interpretation of results. In addition, no formal statistical test is available.

iii) ANNs are subjected to have over fitting problems owing to its large number of parameters.

iv) There are no established theories available for the selection of p and q, hence experiments are often conducted for the determination of the optimal values of p and q which is tedious.

v) ANNs usually require more data for time series forecasting.

Support Vector Machine (SVM) in time series:

Application of SVM in time series is generally utilized when the series shows non stationarity and non-linearity process. A tremendous advantage of SVM is that it is not model dependent as well as independent of stationarity and linearity. However, it may be computationally expensive during the training. The training of the data driven prediction process SVM is done by a function which is estimated utilizing the observed data. Let, a time series $y(t)$ which takes the data at time $t\{t = 0,1,2,3, \dots, N\}$.

Now, the prediction function for linear regression is defined as:

$$f(y) = (w.y) + c$$

Whereas, for non linear regression, it will be:

$$f(y) = \left(w.\emptyset(y)\right) + c$$

Where, w dentoes the weights, c represents threshold value and $\emptyset(y)$ is known as kernel function.If the observed data is linear, then equation (1) will be used. But, for non-lineadata,the mapping of $y(t)$ is done to the higher dimension feature space through some function which is denoted as $\emptyset(y)$ and eventually it is transformed into the linear process. Afer that, a linear regression will carry out in that feature space. The first and foremost objective is to find out the value of w and c which will be optimal. In SVM, there are two things viz., flatness of weights and error after the estimation which are to be minimized. The flatness of the weights is denoted by $||w||^2$ which is the eucledian norm. Firstly, one has to concentrate on minimization the $||w||^2$. Second important thing is the minimization of the error. This is also called as empirical risk. However, the overall aim is to minimize the regularized risk which is sum of empirical risk and the half of the product of the flatness of weight and a constant term which is known as regularized constant. The regularized risk can be written as-

$$R_{reg}(f) = R_{emp}(f) + \frac{\tau}{2}||w||^2$$

Where, $R_{reg}(f)$is the regularized risk, $R_{emp}(f)$ denotes the empirical risk, τ is as constant which is called as regularized constant/capacity control term and $||w||^2$ is the flatness of weights.

The regularization constant has a significant impact on a better fitting of the data and it can also be useful for the minimization of bad generalization effects. In the other words, this constant deals with the problem of over-fitting. The overfitting of the data can be redued by the proper selection of this constant value. The empirical risk can be defined as:-

$$R_{emp}(f) = \frac{1}{N}\sum_{i=0}^{N-1} L\big(y(i), \alpha(i), f(y(i), w)\big)$$

Where, $\alpha(i)$ denotes the truth data of predicted value, $L(.)$ is known as loss function and i represents the index to the time series.There are various types of loss function in literature. But, two functions viz., vapnik loss function and quadratic loss function are most popular and they are generally used. The quadratic programming problem has been made to minimize the regularised risk which is-

$$\text{Minimize, } \frac{1}{2}||w||^2 + D\sum_{i=1}^{n} L\big(\alpha(i), f(y(i), w)\big)$$

Where,

$$L\big(\alpha(i), f(y(i), w)\big) = |\alpha(i) - f(y(i), w)| - \in \text{ if } |\alpha(i) - f(y(i), w)| \geq \in$$

$$= 0; \text{ otherwise.}$$

Where, D is a constant which equals to the summation normalization factor and $\in$ represents the size of the tube.

The computation of $\in$ and D is done empirically because they are user defined. On has to choose proper value of D and $\in$. Now, dual optimization problem is formed using the lagrange multiplier which can be written as:

Maximize, $-\frac{1}{2}\sum_{i,j=1}^{N}(\beta_i - \beta_i{}^*)\left(\beta_j - \beta_j{}^*\right)\langle y(i), y(j)\rangle - \in \sum_{i=1}^{N}(\beta_i - \beta_i{}^*) + \sum_{i=1}^{N}\alpha(i)(\beta_i - \beta_i{}^*)$

Subject to, $\sum_{i-1}^{N}(\beta_i - \beta_i{}^*) = 0$; $\beta_i, \beta_i{}^* \in [0, D]$

The function $f(x)$ is defined as;

$$f(x) = \sum_{i=1}^{N}(\beta_i - \beta_i{}^*)\langle y, y(i)\rangle + C$$

KKT conditions are used to get the solution of the weights.

The significance of kernel function in non-linear support vector machine (NLSVR) is very much imporatnt for mapping the data $y(i)$into higher dimension feature space $\emptyset(\,y(i))$in which the data becomes linear. Generally notation for kernel function is given as;

$$k(y, y') = \langle\emptyset(y), \emptyset(y')\rangle;$$

There are many methods in literature to solve the quadartic programming. However, the most used method is sequential minimization optimization (SMO) algorithm.

References:

Anjoy, P., Paul, R. K., Sinha, K., Paul, A. K. and Ray, M. (2017). A hybrid wavelet based neural networks model for predicting monthly WPI of pulses in India. *Indian Journal of Agricultural Sciences*. **87 (6)**, 834-839.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2009), *Time Series Analysis: Forecasting and Control (3rd ed.), San Francisco: Holden-Day.*

Broock, W., Scheinkman, J. A., Dechert, W. D. and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Review*, **15,** 197–235.

Jha, G. K., and Sinha, K.2014.Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India. *Neural Computing and Applications*24 (3): 563-571.

Makridakis, S., Wheelwright, S.C. and Hyndman, R. J. (1998).*Forecasting: Methods and Applications (3rd ed.),* Chichester: Wiley.

Mukherjee, A., Rakshit, S., Nag, A., Ray, M.**,**Kharbikar, H. L., Kumari, S., Sarkar, S., Paul, S., Roy, S., Maity, A., Meena, V. S. and Burman, R. R. (2016). Climate Change Risk Perception, Adaptation and Mitigation Strategy: An Extension Outlook in Mountain Himalaya. In: Jaideep Kumar Bisht, Vijay Singh Meena, Pankaj Kumar Mishra and Arunava Pattanayak Edition. Conservation Agriculture (pp. 257-292). Singapore. Springer Singapore.

Ray, M., Rai, A., Ramasubramanian, V. and Singh, K. N. (2016). ARIMA-WNN hybrid model for forecasting wheat yield time series data. *Journal of the Indian Society of Agricultural Statistics,* **70(1)**, 63-70.

Ray, M., Rai, A., Singh, K. N., Ramasubramanian, V. and Kumar, A. (2017). Technology forecasting using time series intervention based trend impact analysis for wheat yield scenario in India. *Technological Forecasting & Social Change*, **118**, 128–133.

Remus, W. and O'Connor, M.(2001). *Neural Networks for Time-Series Forecasting, New york, Springer.*

Zhang, G., Patuwo, B. E. and Hu, M. Y. 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14: 35-62.

Planning of Experiments and Basic Experimental Designs

Seema Jaggi and Anindita Datta
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
seema.jaggi@icar.gov.in, anindita.datta@icar.gov.in

An experiment is usually associated with a scientific method for testing certain phenomena. An experiment facilitates the study of such phenomena under controlled conditions and thus creating controlled condition is an essential component. Scientists in the biological fields who are involved in research constantly face problems associated with planning, designing and conducting experiments. Basic familiarity and understanding of statistical methods that deal with issues of concern would be helpful in many ways. Researchers who collect data and then look for a statistical technique that would provide valid results will find that there may not be solutions to the problem and that the problem could have been avoided first by a properly designed experiment. Obviously it is important to keep in mind that we cannot draw valid conclusions from poorly planned experiments. Second, the time and cost involved in many experiments are enormous and a poorly designed experiment increases such costs in time and resources. For example, an agronomist who carries out fertilizer experiment knows the time limitation of the experiment. He knows that when seeds are to be planted and harvested. The experimenter plot must include all components of a complete design. Otherwise what is omitted from the experiment will have to be carried out in subsequent trials in the next cropping season or next year. The additional time and expenditure could be minimized by a properly planned experiment that will produce valid results as efficiently as possible. Good experimental designs are products of the technical knowledge of one's field, an understanding of statistical techniques and skill in designing experiments.

Any research endeavor may entail the phases of Conception, Design, Data collection, Analysis and Dissemination. Statistical methodologies can be used to conduct better scientific experiments if they are incorporated into entire scientific process, i.e., From inception of the problem to experimental design, data analysis and interpretation. When planning experiments we must keep in mind that large uncontrolled variations are common occurrences. Experiments are generally undertaken by researchers to compare effects of several conditions on some phenomena or in discovering an unknown effect of particular process. An experiment facilitates the study of such phenomena under controlled conditions. Therefore the creation of controlled condition is the most essential characteristic of experimentation. How we formulate our questions and hypotheses are critical to the experimental procedure that will follow. For example, a crop scientist who plants the same variety of a crop in a field may find variations in yield that are due to periodic variations across a field or to some other factors that the experimenter has no control over. The methodologies used in designing experiments will separate with confidence and accuracy a varietal difference of crops from the uncontrolled variations.

The different concepts in planning of experiment can be well explained through chapati tasting experiment.

Consider an experiment to detect the taste difference in chapati made of wheat flour of c306 and pv 18 varieties. The null hypothesis we can assume here is that there is no taste difference in chapatis made of c306 or pv18 wheat flours. After the null hypothesis is set, we have to fix the level of significance at which we can operate. The pv18 is a much higher yielding variety than c306. Hence a false rejection may not help the country to grow more pv18 and the wheat production may decrease while a false acceptance may give more production of pv18 wheat and the consumption may be less or practically nil. Thus the false acceptance or false rejection are of practically equal consequence and we agree to choose the level of significance at $\alpha = 0.05$. Now to execute the experiment, a subject is to be found with extrasensory powers who can detect the taste differences. The colours of c306 and pv18 are different and anyone, even without tasting the chapatis, can distinguish the chapatis of either kind by a mere glance. Thus the taster of the chapatis has to be blindfolded before the chapatis are given for tasting. Afterwards, the method is to be decided in which the experiment will be conducted. The experiment can be conducted in many ways and of them three methods are discussed here:

- Give the taster equal number of chapatis of either kind informing the taster about it.
- Give the taster pairs of chapatis of each kind informing the taster about it.
- Give the taster chapatis of either kind without providing him with any information. Let us use 6 chapatis in each of these methods.

Under first method of experimentation, if the null hypothesis is true, then the experimenter cannot distinguish the two kinds of chapaties and he will randomly select 3 chapatiS out of 6 chapaties given to him, as made of pvl8 wheat. In that case, all correct guesses are made if selection exactly coincides with the exactly used wheat variety and the probability for such an occurrence is:

$$1\Big/\binom{6}{3} = 1\Big/20 = 0.05$$

Under second method,the pv18 wheat variety chapaties are selected from each pair given if the null hypothesis is true. Furthermore, independent choices are made of pv18 variety chapaties from each pair. Thus the probability of making all correct guesses is

$$1/(2)^3 = 1/8 = 0.125.$$

In third method the experimenter has to make the choice for each chapati and the situation is analogous at calling heads or tails in a coin tossing experiment. The probability of making all correct guesses would then be:

$$1/2^6 = 1/64 = .016.$$

If the experimenter makes all correct guesses in third method as its probability is smaller than the selected $\alpha = 0.05$, we can reject the null hypothesis and conclude that the two wheat varieties give different tastes at chapaties. In other methods the probability of making all correct guesses does not exceed $\alpha = 0.05$ and hence with either method, we cannot reject the null hypothesis even if all correct guesses are made.

However, if 8 chapaties are used by first method and if the taster guesses all of them, we can reject the null hypothesis, at 0.05 level of significance, as the probability of making all correct guesses would then be $1\Big/\binom{8}{3} = 1\Big/56$ which is smaller than 0.05. 8 chapaties will not enable us to reject the null hypothesis even if all correct guesses are made by second method as the probability of making all correct guesses is $\left(\frac{1}{4}\right)^4 = \frac{1}{16} = 0.06$ it is easy to see that if 10 chapaties are given by second method and if all correct guesses are made, then we can reject the null hypothesis at 0.05 level of significance. Not to unduly influence the taster in making guesses, we should also present the chapaties in a random order rather than systematically presenting them for tasting.

The above discussed chapati tasting experiment brings home the following salient features of experimentation:

- All the extraneous variations in the data should be eliminated or controlled excepting the variations due to the treatments under study. One should not artificially provide circumstances for one treatment to show better results than others.
- Far a given size of the experiment, though the experiment can be done in many ways, even the best results may not turn out to be significant with some designs, while some other design can detect the treatment differences. Thus there is an imperative need the choose the right type of design, before the commencement of the experiment, lest the results may be useless.
- If for some specific reasons related to the nature .of the experiment, a particular method has to be used in experimentation, then adequate number of replications of each treatment have to be provided in order to get valid inferences.
- The treatments have to be randomly allocated to the experimental units.

The terminologies often used in planning and designing of experiments are listed below.

Treatment

Treatment refers to controllable quantitative or qualitative factors imposed at a certain level by the experimenter. For an agronomist several fertilizer concentrations applied to a particular crop or a variety of crop is a treatment. Similarly, an animal scientist looks upon several concentrations of a drug given to animal species as a treatment. In agribusiness we may look upon impact of advertising strategy on sales a treatment. To an agricultural engineer, different levels of irrigation may constitute a treatment.

Experimental Unit

An experimental unit is an entity that receives a treatment e.g., for an agronomist or horticulturist it may be a plot of a land or batch of seed, for an animal scientist it may be a group of pigs or sheep, for a scientist engaged in forestry research it may be different tree species occurring in an area, and for an agricultural engineer it may be manufactured item. Thus, an experimental unit maybe looked upon as a small subdivision of the experimental material, which receives the treatment.

Experimental Error

Differences in yields arising out of experimental units treated alike are called Experimental Error.

Controllable conditions in an experiment or experimental variable are terms as a factor. For example, a fertilizer, a new feed ration, and a fungicide are all considered as factors. Factors may be qualitative or quantitative and may take a finite number of values or type. Quantitative factors are those described by numerical values on some scale. The rates of application of fertilizer, the quantity of seed sown are examples of quantitative factors. Qualitative factors are those factors that can be distinguished from each other, but not on numerical scale e.g., type of protein in a diet, sex of an animal, genetic make up of plant etc. While choosing factors for any experiment researcher should ask the following questions, like What treatments in the experiment should be related directly to the objectives of the study? Does the experimental technique adopted require the use of additional factors? Can the experimental unit be divided naturally into groups such that the main treatment effects are different for the different groups? What additional factors should one include in the experiment to interact with the main factors and shed light on the factors of direct interest? How desirable is it to deliberately choose experimental units of different types?

Basic Principles of Design of Experiments

Given a set of treatments which can provide information regarding the objective of an experiment, a design for the experiment, defines the size and number of experimental units, the manner in which the treatments are allotted to the units and also appropriate type and grouping of the experimental units. These requirements of a design ensure validity, interpretability and accuracy of the results obtainable from an analysis of the observations.

These purposes are served by the principles of:

- Randomization
- Replication
- Local (Error) control

Randomization

After the treatments and the experimental units are decided the treatments are allotted to the experimental units at random to avoid any type of personal or subjective bias, which may be conscious or unconscious. This ensures validity of the results. It helps to have an objective comparison among the treatments. It also ensures independence of the observations, which is necessary for drawing valid inference from the observations by applying appropriate statistical techniques.

Depending on the nature of the experiment and the experimental units, there are various experimental designs and each design has its own way of randomization. Various speakers while discussing specific designs in the lectures to follow shall discuss the procedure of random allocation separately.

Replication

If a treatment is allotted to r experimental units in an experiment, it is said to be replicated r times. If in a design each of the treatments is replicated r times, the design is said to have r replications. Replication is necessary to

- Provide an estimate of the error variance which is a function of the differences among observations from experimental units under identical treatments.
- Increase the accuracy of estimates of the treatment effects.

Though, more the number of replications the better it is, so far as precision of estimates is concerned, it cannot be increased infinitely as it increases the cost of experimentation. Moreover, due to limited availability of experimental resources too many replications cannot be taken.

The number of replications is, therefore, decided keeping in view the permissible expenditure and the required degree of precision. Sensitivity of statistical methods for drawing inference also depends on the number of replications. Sometimes this criterion is used to decide the number of replications in specific experiments.

Error variance provides a measure of precision of an experiment, the less the error variance the more precision. Once a measure of error variance is available for a set of experimental units, the number of replications needed for a desired level of sensitivity can be obtained as below.

Given a set of treatments an experimenter may not be interested to know if two treatment differ in their effects by less than a certain quantity, say, d. In other words, he wants an experiment that should be able to differentiate two treatments when they differ by d or more.

The significance of the difference between two treatments is tested by t-test where

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{2s^2 / r}},$$

Here, $\bar{y}_i$, and $\bar{y}_j$ are the arithmetic means of two treatment effects each based on r replications, s^2 is measure of error variation.

Given a difference d, between two treatment effects such that any difference greater than d should be brought out as significant by using a design with r replications, the following equation provides a solution of r.

$$t = \frac{|d\ |}{\sqrt{2s^2 / r}},$$

$$r = \frac{t_0^2}{d^2} x 2s^2$$

where t_0 is the critical value of the t-distribution at the desired level of significance, that is, the value of t at 5 or 1 per cent level of significance read from the t-table. If s^2 is known or based on a very large number of observations, made available from some pilot pre-experiment investigation, then t is taken as the normal variate. If s^2 is estimated with n degree of freedom (d.f.) then t_0 corresponds to n d.f.

When the number of replication is r or more as obtained above, then all differences greater than d are expected to be brought out as significant by an experiment when it is conducted on a set of experimental units which has variability of the order of s^2. For example, in an experiment on wheat crop conducted in a seed farm in Bhopal, to study the effect of application of nitrogen and phosphorous on yield a randomized block design with three replications was adopted. There were 11 treatments two of which were (i) 60 Kg/ha of nitrogen (ii) 120 Kg/ha of nitrogen. The average yield figures for these two application of the fertilizer were 1438 and 1592 Kg/ha respectively and it is required that differences of the order of 150 Kg/ha should be brought out significant. The error mean square (s^2) was 12134.88. Assuming that the experimental error will be of the same order in future experiments and t_0 is of the order of 2.00, which is likely as the error d.f. is likely to be more than 30 as there are 11 treatments; Substituting in (1), we get:

$$r = \frac{2t_0^2 s^2}{d^2} = \frac{2x2^2 x2134.88}{150^2} = 4\,(\text{approx.})$$

Thus, an experiment with 4 replications is likely to bring out differences of the order of 150 Kg/ha as significant.

Another criterion for determining r is to take a number of replications which ensures at least 10 d.f. for the estimate of error variance in the analysis of variance of the design concerned since the sensitivity of the experiment will be very much low as the F test (which is used to draw inference in such experiments) is very much unstable below 10 d.f.

Local Control

The consideration in regard to the choice of number of replications ensure reduction of standard error of the estimates of the treatment effect because the standard error of the estimate of a treatment effect is $\sqrt{s^2/r}$, but it cannot reduce the error variance itself. It is, however, possible to devise methods for reducing the error variance. Such measures are called *error control* or local control. One such measure is to make the experimental units homogenous. Another method is to form the units into several homogenous groups, usually called blocks, allowing variation between the groups.

A considerable amount of research work has been done to divide the treatments into suitable groups of experimental units so that the treatment effect can be estimated more precisely Extensive use of combinatorial mathematics has been made for formation of such group treatments. This grouping of experiment units into different groups has led to the development of various designs useful to the experimenter. We now briefly describe the various term used in designing of an experiment

Blocking

It refers to methodologies that form the units into homogeneous or pre-experimental subject-similarity groups. It is a method to reduce the effect of variation in the experimental material on the Error of Treatment of Comparisons. For example, animal scientist may decide to group animals on age, sex, breed or some other factors that he may believe has an influence on characteristic being measured. Effective blocking removes considerable measure of variation nom the experimental error. The selection of source of variability to be used as basis of blocking, block size, block shape and orientation are crucial for blocking. The blocking factor is introduced in the experiment to increase the power of design to detect treatment effects.

The importance of good designing is inseparable from good research (results). The following examples point out the necessity for a good design that will yield good research. First, a nutrition specialist in developing country is interested in determining whether mother's milk is better than powdered milk for children under age one. The nutritionist has compared the growth of children in village A, who are all on mother's milk against the children in village B, who use powdered milk. Obviously, such a comparison ignores the health of the mothers, the sanitary-conditions of the villages, and other factors that may have contributed to the differences observed without any connection to the advantages of mother's milk or the powdered milk on the children.

A proper design would require that both mother's milk and the powdered milk be alternatively used in both villages, or some other methodology to make certain that the differences observed are attributable to the type of milk consumed and not to some uncontrollable factor. Second, a crop scientist who is comparing 2 varieties of maize, for instance, would not assign one variety to a location where such factors as sun, shade, unidirectional fertility gradient, and uneven distribution of water would either favor or handicap it over the other. If such a design were to be adopted, the researcher would have difficulty in determining whether the apparent difference in yield was due to variety differences or resulted from such factors as sun, shade, soil fertility of the field, or the distribution of water. These two examples illustrate the type of poorly designed experiments that are to be avoided.

Analysis of Variance

Analysis of Variance (ANOVA) is a technique of partitioning the overall variation in the responses into different assignable sources of variation, some of which are specifiable and others unknown. Total variance in the sample data is partitioned and is expressed as the sum of its non-negative components is a measure of the variation due to some specific independent source or factor or cause. ANOVA consists in estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to ascribable factors (causes) with the estimate due to chance factor the latter being known as experimental error or simply the error.

Total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components associated with the nature of classification of the data. The systematic procedure for achieving this is called *Analysis of Variance*. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance.

Thus, ANOVA is a statistical technique that can be used to evaluate whether there are differences between the average value, or mean, across several population groups. With this model, the *response variable is continuous* in nature, whereas the *predictor variables are categorical*. For example, in a clinical trial of hypertensive patients, ANOVA methods could be used to compare the effectiveness of three different drugs in lowering blood pressure. Alternatively, ANOVA could be used to determine whether infant birth weight is significantly different among mothers who smoked during pregnancy relative to those who did not. In a particular case, where two population means are being compared, ANOVA is equivalent to the independent two-sample *t*-test.

The fixed-effects model of ANOVA applies to situations in which the experimenter applies several treatments to the subjects of the experiment to see if the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole. In it factors are fixed and are attributable to a finite set of levels of factor eg. Sex, year, variety, fertilizer etc.

Consider for example a clinical trial where three drugs are administered on a group of men and women some of whom are married and some are unmarried. The three classifications of sex, drug and marital status that identify the source of each datum are known as factors. The individual classification of each factor is known as levels of the factors. Thus, in this example there are 3 levels of factor drug, 2 levels of factor sex and 2 levels of marital status. Here all the effects are fixed. Random effects models are used when the treatments are not fixed. This occurs when the various treatments (also known as factor levels) are sampled from a larger population. When factors are random, these are generally attributable to infinite set of levels of a factor of which a random sample are deemed to occur *eg.* research stations, clinics in Delhi, sire, etc. Suppose new inject-able insulin is to be tested using 15 different clinics of Delhi state. It is reasonable to assume that these clinics are random sample from a population of clinics from Delhi. It describe the situations where both fixed and random effects are present.

In any ANOVA model, general mean is always taken as fixed effect and error is always taken as random effect. Thus class of model can be classified on the basis of factors, other than these two factors. ANOVA can be viewed as a generalization of *t*-tests: a comparison of differences of means across more than two groups.

The ANOVA is valid under certain assumptions. These assumptions are:

- Samples have been drawn from the populations that are normally distributed.
- Observations are independent and are distributed normally with mean zero and variance σ^2.
- Effects are additive in nature.

The ANOVA is performed as one-way, two-way, three-way, etc. ANOVA when the number of factors is one, two or three respectively. In general if the number of factors is more, it is termed as multi-way ANOVA.

Completely Randomized Design

Designs are usually characterized by the nature of grouping of experimental units and the procedure of random allocation of treatments to the experimental units. In a completely randomized design the units are taken in a single group. As far as possible the units forming the group are homogeneous. This is a design in which only randomization and replication are used. There is no use of local control here.

Let there be *v* treatments in an experiment and *n* homogeneous experimental units. Let the i^{th} treatment be replicated r_i times *(i = 1,2,..., v)* such that $\sum_{i=1}^{v} r_i = n$. The treatments are allotted at random to the units.

Normally the number of replications for different treatments should be equal as it ensures equal precision of estimates of the treatment effects. The actual number of replications is, however, determined by the availability of experimental resources and the requirement of precision and sensitivity of comparisons. If the experimental material for some treatments is available in limited quantities, the numbers of their replication are reduced. If the estimates of certain treatment effects are required with more precision, the numbers of their replication are increased.

Randomization

There are several methods of random allocation of treatments to the experimental units. The *v* treatments are first numbered in any order from *1* to *v*. The *n* experimental units are also numbered suitably. One of the methods uses the random number tables. Any page of a random number table is taken. If *v* is a one-digit number, then the table is consulted digit by digit. If *v* is a two-digit number, then two-digit random numbers are consulted. All numbers greater than *v* including zero are ignored.

Let the first number chosen be n_1; then the treatment numbered n_1 is allotted to the first unit. If the second number is n_2 which may or may not be equal to n_1 then the treatment numbered n_2 is allotted to the second unit. This procedure is continued. When the i^{th} treatment number has occurred r_i times, $(i = 1,2,...,v)$ this treatment is ignored subsequently. This process terminates when all the units are exhausted.

One drawback of the above procedure is that sometimes a very large number of random numbers may have to be ignored because they are greater than *v*. It may even happen that the random number table is exhausted before the allocation is complete. To avoid this difficulty the following procedure is adopted. We have described the procedure by taking *v* to be a two-digit number.

Let *P* be the highest two-digit number divisible by *v*. Then all numbers greater than *P* and zero are ignored. If a selected random number is less than *v*, then it is used as such. If it is greater than or equal to *v*, then it is divided by *v* and the remainder is taken to the random number. When a number is completely divisible by *v*, then the random number is *v*. If *v* is an *n*-digit number, then *P* is taken to be the highest *n*-digit number divisible by *v*. The rest of the procedure is the same as above.

Analysis

This design provides a one-way classified data according to levels of a single factor. For its analysis the following model is taken:

$$y_{ij} = \mu + t_i + e_{ij}\ , \qquad i = 1,\cdots,v;\, j = 1,\cdots r_i,$$

where y_{ij} is the random variable corresponding to the observation y_{ij} obtained from the *j*th replicate of the *i*th treatment, μ is the general mean, t_i is the fixed effect of the *i*th treatment and e_{ij} is the error component which is a random variable assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Let $\sum y_{ij} = T_i \quad (i = 1,2,...,v)$ be the total of observations from *i*th treatment. Let further $\sum_i T_i = G$. Correction factor (*C.F.*) = G^2/n.

Sum of squares due to treatments $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$

Total sum of squares = $\sum_{i=1}^{v}\sum_{j=1}^{r_i} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Treatments	$v - 1$	SST $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$	$MST = SST / (v - 1)$	MST/MSE
Error	$n - v$	SSE = by subtraction	MSE = SSE / (n - v)	
Total	$n - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test where F is the ratio *MST / MSE* with *(v - 1)* and *(n - v)* degrees of freedom.

3. Randomized Complete Block Design

It has been seen that when the experimental units are homogeneous then a CRD should be adopted. In any experiment, however, besides treatments the experimental material is a major source of variability in the data. When experiments require a large number of experimental units, the experimental units may not be homogeneous, and in such situations CRD can not be recommended. When the experimental units are heterogeneous, a part of the variability can be accounted for by grouping the experimental units in such a way that experimental units within each group are as homogeneous as possible. The treatments are then allotted randomly to the experimental units within each group (or blocks). The principle of first forming homogeneous groups of the experimental units and then allotting at random each treatment once in each group is known as local control. This results in an increase in precision of estimates of the treatment contrasts, due to the fact that error variance that is a function of comparisons within blocks, is smaller because of homogeneous blocks. This type of allocation makes it possible to eliminate from error variance a portion of variation attributable to block differences. If, however, variation between the blocks is not significantly large, this type of grouping of the units does not lead to any advantage; rather some degrees of freedom of the error variance is lost without any consequent decrease in the error variance. In such situations it is not desirable to adopt randomized complete block designs in preference to completely randomized designs.

If the number of experimental units within each group is same as the number of treatments and if every treatment appears precisely once in each group then such an arrangement is called a ***randomized complete block design.***

Suppose the experimenter wants to study v treatments. Each of the treatments is replicated r times (the number of blocks) in the design. The total number of experimental units is, therefore, vr. These units are arranged into r groups of size v each. The error control measure in this design consists of making the units in each of these groups homogeneous.

The number of blocks in the design is the same as the number of replications. The v treatments are allotted at random to the v plots in each block. This type of homogeneous grouping of the experimental units and the random allocation of the treatments separately in each block are the two main characteristic features of randomized block designs. The availability of resources and considerations of cost and precision determine actual number of replications in the design.

Analysis

The data collected from experiments with randomized block designs form a two-way classification, that is, classified according to the levels of two factors, *viz.,* blocks and treatments. There are *vr* cells in the two-way table with one observation in each cell. The data are orthogonal and therefore the design is called an *orthogonal design.* We take the following model:

$$y_{ij} = \mu + t_i + b_j + e_{ij}\ , \qquad \begin{pmatrix} i = 1,2,...,v; \\ j = 1,2,...,r \end{pmatrix},$$

where y_{ij} denotes the observation from i^{th} treatment in j^{th} block. The fixed effects μ, t_i, b_j denote respectively the general mean, effect of the i^{th} treatment and effect of the j^{th} block. The random variable e_{ij} is the error component associated with y_{ij}. These are assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Following the method of analysis of variance for finding sums of squares due to blocks, treatments and error for the two-way classification, the different sums of squares are obtained as follows: Let $\sum_j y_{ij} = T_i \ \ (i = 1,2,...,v)$ = total of observations from i^{th} treatment and $\sum_j y_{ij} = B_j \quad j = 1,\cdots,r$ = total of observations from j^{th} block.

These are the marginal totals of the two-way data table. Let further, $\sum_i T_i = \sum_j B_j = G.$

Correction factor (*C.F.*) = G^2/rv, Sum of squares due to treatments $= \sum_i \frac{T_i^2}{r} - C.F.$,

Sum of squares due to blocks $= \sum_j \frac{B_j^2}{v} - C.F.$, Total sum of squares $= \sum_{ij} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Blocks	$r - 1$	$SSB = \sum_j \frac{B_j^2}{v} - C.F.$	$MSB = SSB / (r - 1)$	MSB/MSE
Treatments	$v - 1$	$SST = \sum_i \frac{T_i^2}{r} - C.F.$	$MST = SST / (v - 1)$	MST/MSE
Error	$(r - 1)(v - 1)$	SSE = *by subtraction*	$MSE =$ $SSE / (v - 1)(r - 1)$	
Total	$vr - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test, where F is the ratio MST / MSE with $(v - 1)$ and $(v - 1)(r - 1)$ degrees of freedom. We may then be interested to either compare the treatments in pairs or evaluate special contrasts depending upon the objectives of the experiment. This is done as follows:

The critical difference for testing the significance of the difference of two treatment effects, say $t_i - t_j$ is $C.D. = t_{(v-1)(r-1),\alpha/2}\sqrt{2MSE / r}$, where $t_{(v-1)(r-1),\alpha/2}$ is the value of Student's t at the level of significance α and degree of freedom $(v - 1)(r - 1)$. If the difference of any two-treatment means is greater than the C.D. value, the corresponding treatment effects are significantly different.

4. Latin Square Design

Latin square designs are normally used in experiments where it is required to remove the heterogeneity of experimental material in two directions. These designs require that the number of replications equal the number of ***treatments*** or ***varieties***.

Definition 1. A Latin square arrangement is an arrangement of v symbols in v^2 cells arranged in v rows and v columns, such that every symbol occurs precisely once in each row and precisely once in each column. The term v is known as the **order** of the Latin square.

If the symbols are taken as A, B, C, D, a Latin square arrangement of order 4 is as follows:

$$\begin{array}{cccc} A & B & C & D \\ B & C & D & A \\ C & D & A & B \\ D & A & B & C \end{array}$$

A Latin square is said to be in the ***standard form*** if the symbols in the first row and first column are in natural order, and it is said to be in the ***semi-standard form*** if the symbols of the first row are in natural order. Some authors denote both of these concepts by the term ***standard form***. However, there is a need to distinguish between these two concepts. The standard form is used for randomizing the Latin-square designs, and the semi-standard form is needed for studying the properties of the orthogonal Latin squares.

Definition 2. If in two Latin squares of the same order, when superimposed on one another, every ordered pair of symbols occurs exactly once, the two Latin squares are said to be **orthogonal**. If the symbols of one Latin square are denoted by Latin letters and the symbols of the other are denoted by Greek letters, the pair of orthogonal Latin squares is also called a **graeco-latin square**.

Definition 3. If in a set of Latin squares every pair is orthogonal, the set is called a set of **mutually orthogonal latin squares (MOLS)**. It is also called a **hypergraeco latin square.**

The following is an example of graeco latin square:

$$\begin{array}{cccc} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{array} \qquad \begin{array}{cccc} \alpha & \gamma & \delta & \beta \\ \beta & \delta & \gamma & \alpha \\ \gamma & \alpha & \beta & \delta \\ \delta & \beta & \alpha & \gamma \end{array}$$

$$\begin{array}{cccc} A\alpha & B\gamma & C\delta & D\beta \\ B\beta & A\delta & D\gamma & C\alpha \\ C\gamma & D\alpha & A\beta & B\delta \\ D\delta & C\beta & B\alpha & A\gamma \end{array}$$

We can verify that in the above arrangement every pair of ordered Latin and Greek symbols occurs exactly once, and hence the two latin squares under consideration constitute a graecolatin square.

It is well known that the maximum number of MOLS possible of order v is $v - 1$. A set of $v - 1$ MOLS is known as a complete set of MOLS. Complete sets of MOLS of order v exist when v is a **prime or prime power.**

Randomization

According to the definition of a Latin square design, treatments can be allocated to the v^2 experimental units (may be animal or plots) in a number of ways. There are, therefore, a number of Latin squares of a given order. The purpose of randomization is to select one of these squares at random. The following is one of the methods of random selection of Latin squares.

Let a $v \times v$ Latin square arrangement be first written by denoting treatments by Latin letters *A, B, C, etc.* or by numbers *1, 2, 3, etc.* Such arrangements are readily available in the ***Tables for Statisticians and Biometricians*** (Fisher and Yates, 1974). One of these squares of any order can be written systematically as shown below for a 5×5 Latin square:

$$\begin{array}{ccccc}
A & B & C & D & E \\
B & C & D & E & A \\
C & D & E & A & B \\
D & E & A & B & C \\
E & A & B & C & D
\end{array}$$

For the purpose of randomization rows and columns of the Latin square are rearranged randomly. There is no randomization possible within the rows and/or columns. For example, the following is a row randomized square of the above 5×5 Latin square;

$$\begin{array}{ccccc}
A & B & C & D & E \\
B & C & D & E & A \\
E & A & B & C & D \\
D & E & A & B & C \\
C & D & E & A & B
\end{array}$$

Next, the columns of the above row randomized square have been rearranged randomly to give the following random square:

$$\begin{array}{ccccc}
E & B & C & A & D \\
A & C & D & B & E \\
D & A & B & E & C \\
C & E & A & D & B \\
B & D & E & C & A
\end{array}$$

As a result of row and column randomization, but not the randomization of the individual units, the whole arrangement remains a Latin square.

Analysis of Latin Square Designs

In Latin square designs there are three factors. These are the factors *P, Q,* and treatments. The data collected from this design are, therefore, analyzed as a three-way classified data. Actually, there should have been v^3 observations as there are three factors each at v levels. But because of the particular allocation of treatments to the cells, there is only one observation per cell instead of *v* in the usual three way classified orthogonal data. As a result we can obtain only the sums of squares due to each of the three factors and error sum of squares. None of the interaction sums of squares of the factors can be obtained. Accordingly, we take the model

$$Y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$$

where y_{ijs} denotes the observation in the i^{th} row, j^{th} column and under the s^{th} treatment; $\mu, r_i, c_j, t_s (i, j, s = 1,2,...,v)$ are fixed effects denoting in order the general mean, the row, the column and the treatment effects. The e_{ijs} is the error component, assumed to be independently and normally distributed with zero mean and a constant variance, σ^2.

The analysis is conducted by following a similar procedure as described for the analysis of two-way classified data. The different sums of squares are obtained as below: Let the data be arranged first in a *row × column* table such that y_{ij} denotes the observation of (*i*, *j)*th cell of table.

Let $R_i = \sum_j y_{ij} = i^{th}\ row\ total\ (i = 1,2,...,v)$, $C_j = \sum_i y_{ij} = j^{th}\ column\ total\ (j = 1,2,...,v)$, T_s = sum of those observations which come from s^{th} treatment *(s= 1,2,...,v),* $G = \sum_i R_i = grand\ total$. Correction factor, $C.F. = \frac{G^2}{v^2}$. Treatment sum of squares = $\sum_s \frac{T_s^2}{v} - C.F.$, Row sum of squares = $\sum_i \frac{R_i^2}{v} - C.F.$, Column sum of squares = $\sum_j \frac{C_j^2}{v} - C.F.$

Analysis of Variance of *v × v* Latin Square Design

Sources of Variation	D.F.	S.S.	M.S.	F
Rows	*v -1*	$\sum_i \frac{R_i^2}{v} - C.F.$		
Columns	*v - 1*	$\sum_j \frac{C_j^2}{v} - C.F.$		
Treatments	*v - 1*	$\sum_s \frac{T_s^2}{v} - C.F.$	s_t^2	s_t^2 / s_e^2
Error	*(v - 1)(v - 2)*	By subtraction	s_e^2	
Total	v^2*-1*	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis of equal treatment effects is tested by *F*-test, where *F* is the ratio of treatment mean squares to error mean squares. If *F* is not significant, treatment effects do not differ significantly among themselves. If *F* is significant, further studies to test the significance of any treatment contrast can be made in exactly the same way as discussed for randomized block designs.

ANCOVA

Anindita Datta
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
anindita.datta@icar.gov.in

Introduction

The meaning of ANVOVA is Analysis of Covariance. It is a general linear model with one continuous outcome variable (quantitative) and one or more factor variables (qualitative). ANCOVA is a merger of ANOVA and regression for continuous variables. ANCOVA tests whether certain factors have an effect on the outcome variable after removing the variance for which quantitative predictors (covariates) account. The inclusion of covariates can increase statistical power because it accounts for some of the variability.

It is well known that in designed experiments the ability to detect existing differences among treatments increases as the size of the experimental error decreases, a good experiment attempts to incorporate all possible means of minimizing the experimental error. Besides proper experimentation, a proper data analysis also helps in controlling experimental error. In situations where blocking alone may not be able to achieve adequate control of experimental error, proper choice of data analysis may help a great deal. By measuring one or more *covariates* - the characters whose functional relationships to the character of primary interest are known - the Analysis of Covariance (ANCOVA) can reduce the variability among experimental units by adjusting their values to a common value of the covariates. For example, in an animal feeding trial, the initial body weight of the animals usually differs. Using this initial body weight as a covariate, the final weights recorded after the animals have been subjected to various physiological feeds (treatments) can be adjusted to the values that would have been obtained had there been no variation in the initial body weights of the animals at the start of the experiment. An another example, in a field experiment where rodents have (partially) damaged some of the plots, covariance analysis with rodent damage as a covariate could be useful in adjusting plot yields to the levels that they should have been had there been no rodent damage in any plot.

ANCOVA requires measurement of the character of primary interest plus the measurement of one or more variables known as *covariates*. It also requires that the functional relationship of the covariates with the character of primary interest is known beforehand. Generally a linear relationship is assumed, though other type of relationships could also be assumed.

Consider the case of a variety trial in which weed incidence is used as a covariate. With a known functional relationship between weed incidence and grain yield, the character of primary interest, the covariance analysis can adjust grain yield in each plot to a common level of weed incidence. With this adjustment, the variation in yield due to weed incidence is quantified and effectively separated from that due to varietal difference.

ANCOVA can be applied to any number of covariates and to any type of functional relationship between variables *viz.* quadratic, inverse polynomial, etc. Here we illustrate the use of covariance analysis with the help of a single covariate that is linearly related with the character of primary interest. It is expected that this simplification shall not unduly reduce the applicability of the technique, as a single covariate that is linearly related with the primary variable is adequate for most of the experimental situations in agricultural research.

Uses of Covariance Analysis in Agricultural Research

There are several important uses of covariance analysis in agricultural research. Some of the most important ones are:

1. To control experimental error and to adjust treatment means.
2. To aid in the interpretation of experimental results.
3. To estimate missing data.

Error Control and Adjustment of Treatment Means

It is now well realized that the size of experimental error is closely related to the variability between experimental units. It is also known that proper blocking can reduce experimental error by maximizing the differences between the blocks and thus minimizing differences within blocks. Blocking, however, can not cope with certain types of variability such as spotty soil heterogeneity and unpredictable insect incidence. In both instances, heterogeneity between experimental plots does not follow a definite pattern, which causes difficulty in getting maximum differences between blocks. Indeed, blocking is ineffective in the case of nonuniform insect incidences because blocking must be done before the incidence occurs. Furthermore, even though it is true that a researcher may have some information on the probable path or direction of insect movement, unless the direction of insect movement coincides with the soil fertility gradient, the choice of whether soil heterogeneity or insect incidence should be the criterion for blocking is difficult. The choice is especially difficult if both sources of variation have about the same importance.

Use of covariance analysis should be considered in experiments in which blocking couldn't adequately reduce the experimental error. By measuring an additional variable (*e.g.,* covariate X) that is known to be linearly related to the primary variable Y, the source of variation associated with the covariate can be deducted from experimental error. This adjusts the primary variable Y linearly upward or downward, depending on the relative size of its respective covariate. The adjustment accomplishes two important improvements:

1. The treatment mean is adjusted to a value that it would have had; had there been no differences in the values of the covariate.
2. The experimental error is reduced and the precision for comparing treatment means is increased.

Although blocking and covariance techniques are both used to reduce experimental error, the differences between the two techniques are such that they are usually not interchangeable. The ANCOVA can be used only when the covariate representing the heterogeneity among the experimental units can be measured quantitatively. However, that is not a necessary condition for blocking. In addition, because blocking is done before the start of the experiment, it can be used only to cope with sources of variation that are known or predictable. ANCOVA, on the other hand, can take care of unexpected sources of variation that occur during the experiment. Thus, ANCOVA is useful, as a supplementary procedure to take care of sources of variation that cannot be accounted for by blocking.

When covariance analysis is used for error control and adjustment of treatment means, the covariate must not be affected by the treatments being tested. Otherwise, the adjustment removes both the variation due to experimental error and that due to treatment effects. A good example of covariates that are free of treatment effects are those that are measured before the treatments are applied, such as soil analysis and residual effects of treatments applied in the past experiments. In other cases, care must be exercised to ensure that the covariates defined are not affected by the treatments being tested. This technique can be illustrated through the following example:

Example 1: A trial was designed to evaluate 15 rice varieties grown in soil with a toxic level of iron. The experiment was in a RCB design with three replications. Guard rows of a susceptible check variety were planted on two sides of each experimental plot. Scores for tolerance for iron toxicity were collected from each experimental plot as well as from guard rows. For each experimental plot, the score of susceptible check (averaged over two guard rows) constitutes the value of the covariate for that plot. Data on the tolerance score of each variety (Y variable) and on the score of the corresponding susceptible check (X variable) are shown below:

Scores of tolerance for iron toxicity (Y) of 15 rice varieties and those of the corresponding guard rows of a susceptible check variety (X) in a RCB trial

Variety Number	Replication-I		Replication-II		Replication-III	
	X	Y	X	Y	X	Y
1.	15	22	16	13	16	14
2.	16	14	15	23	15	23
3.	15	24	15	24	15	23
4.	16	13	15	23	15	23
5.	17	17	17	16	16	16
6.	16	14	15	23	15	23
7.	16	13	15	23	16	13
8.	16	16	17	17	16	16
9.	17	14	15	23	15	24
10.	17	17	17	17	15	26
11.	16	15	15	24	15	25
12.	16	15	15	23	15	23
13.	15	24	15	24	16	15
14.	15	25	15	24	15	23
15.	15	24	15	25	16	16

The usual analysis of variance without using the covariate (X variable) is as follows:

Source	**DF**	**SS**	**Mean Square**	**F Value**	**Pr > F**
Replication	2	104.0444	52.0222	2.85	0.0745
Treatment	14	265.9111	18.9937	1.04	0.4448
Error	28	510.6222	18.2365		
Total	**44**	**880.5778**			

R-Square	**C.V.**	**Root MSE**	**Y - Mean**
0.4201	21.5436	4.2704	19.82222

Using the covariate, the analysis is the following:

Source	**DF**	**S.S.**	**M.S.**	**F-Value**	**Pr > F**
Replication	2	22.4802	11.2402	2.71	0.0844
Treatment	14	152.5606	10.8972	2.63	0.0151
Covariate X	1	398.7516	398.7516	96.24	0.0001
Error	27	111.8707	4.1434		

R-Square	**C.V.**	**Root MSE**	**Y Mean**
0.8730	10.2689	2.0355	19.8222

It is interesting to note that the use of a covariate has resulted into a considerable reduction in the error mean square and hence the CV has also reduced drastically. This has helped in catching the small differences among the treatment effects as significant. This was not possible when the covariate was not used. The covariance analysis will thus result into a more precise comparison of treatment effects.

The probability of significance of pairwise comparisons among the least square estimates of the treatment effects are given below:

```
                        Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

i/j    1       2       3       4       5       6       7       8       9

1      .     0.3370  0.0666  0.4431  0.0019  0.3370  1.0000  0.0252  0.0232

2   0.3370      .    0.3370  0.8425  0.0237  1.0000  0.3370  0.1834  0.1697

3   0.0666  0.3370      .    0.2497  0.1620  0.3370  0.0666  0.6757  0.6751

4   0.4431  0.8425  0.2497      .    0.0157  0.8425  0.4431  0.1320  0.1191

5   0.0019  0.0237  0.1620  0.0157      .    0.0237  0.0019  0.2361  0.2493

6   0.3370  1.0000  0.3370  0.8425  0.0237      .    0.3370  0.1834  0.1697

7  1.0000  0.3370  0.0666  0.4431  0.0019  0.3370      .    0.0252  0.0232

8   0.0252  0.1834  0.6757  0.1320  0.2361  0.1834  0.0252      .    0.9727

9   0.0232  0.1697  0.6751  0.1191  0.2493  0.1697  0.0232  0.9727      .

10  0.0001  0.0019  0.0237  0.0012  0.3370  0.0019  0.0001  0.0361  0.0385

11  0.0874  0.4294  0.8575  0.3249  0.1046  0.4294  0.0874  0.5445  0.5439

12  0.2497  0.8425  0.4431  0.6915  0.0351  0.8425  0.2497  0.2493  0.2361

13  0.1270  0.5524  0.7066  0.4294  0.0739  0.5524  0.1270  0.4298  0.4229

14  0.0446  0.2497  0.8425  0.1803  0.2158  0.2497  0.0446  0.8096  0.8204

15  0.0589  0.3249  0.9860  0.2393  0.1452  0.3249  0.0589  0.6736  0.6809
```

```
 Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

  i/j     10      11     12        13        14        15

   1    0.0001  0.0874  0.2497    0.1270    0.0446    0.0589

   2    0.0019  0.4294  0.8425    0.5524    0.2497    0.3249

   3    0.0237  0.8575  0.4431    0.7066    0.8425    0.9860

   4    0.0012  0.3249  0.6915    0.4294    0.1803    0.2393

   5    0.3370  0.1046  0.0351    0.0739    0.2158    0.1452

   6    0.0019  0.4294  0.8425    0.5524    0.2497    0.3249

   7    0.0001  0.0874  0.2497    0.1270    0.0446    0.0589
```

8	0.0361	0.5445	0.2493	0.4298	0.8096	0.6736
9	0.0385	0.5439	0.2361	0.4229	0.8204	0.6809
10	.	0.0124	0.0031	0.0079	0.0351	0.0191
11	0.0124	.	0.5524	0.8425	0.7066	0.8425
12	0.0031	0.5524	.	0.6915	0.3370	0.4294
13	0.0079	0.8425	0.6915	.	0.5671	0.6915
14	0.0351	0.7066	0.3370	0.5671	.	0.8575
15	0.0191	0.8425	0.4294	0.6915	0.8575	.

References

Cochran, W. G., and Cox, G. M. (1957). *Experimental Design,* 2nd edition. New York: Wiley.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York.

FACTORIAL EXPERIMENTS

Seema Jaggi and Anindita Datta
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
seema.jaggi@icar.gov.in, anindita.datta@icar.gov.in

1. Introduction

Factorial Experiments are experiments that investigate the effects of two or more factors or input parameters on the output response of a process. Factorial experiment design, or simply factorial design, is a systematic method for formulating the steps needed to successfully implement a factorial experiment. Estimating the effects of various factors on the output of a process with a minimal number of observations is crucial to being able to optimize the output of the process.

In a factorial experiment, the effects of varying the levels of the various factors affecting the process output are investigated. Each complete trial or replication of the experiment takes into account all the possible combinations of the varying levels of these factors. Effective factorial design ensures that the least number of experiment runs are conducted to generate the maximum amount of information about how input variables affect the output of a process.

For example, an experiment on rooting of cuttings involving two factors, each at two levels, such as two hormones at two doses, is referred to as a 2 x 2 or a 2^2 factorial experiment. Its treatments consist of the following four possible combinations of the two levels in each of the two factors.

Treatment number	**Treatment Combination**	
	Hormone	**Dose (ppm)**
1	NAA	10
2	NAA	20
3	IBA	10
4	IBA	20

The total number of treatments in a factorial experiment is the product of the number of levels of each factor; in the 2^2 factorial example, the number of treatments is 2 x 2 = 4, in the 2^3 factorial, the number of treatments is 2 x 2 x 2 = 8. The number of treatments increases rapidly with an increase in the number of factors or an increase in the levels in each factor. For a factorial experiment involving 5 clones, 4 espacements, and 3 weed-control methods, the total number of treatments would be 5 x 4 x 3 = 60. Thus, indiscriminate use of factorial experiments has to be avoided because of their large size, complexity, and cost. Furthermore, it is not wise to commit oneself to a large experiment at the beginning of the investigation when several small preliminary experiments may offer promising results. For example, a tree breeder has collected 30 new clones from a neighbouring country and wants to assess their reaction to the local environment. Because the environment is expected to vary in terms of soil fertility, moisture levels, and so on, the ideal experiment would be one that tests the 30 clones in a factorial experiment involving such other variable factors as fertilizer, moisture level, and population density. Such an experiment, however, becomes extremely large as factors other than clones are added. Even if only one factor, say nitrogen or fertilizer with three levels were included, the number of treatments would increase from 30 to 90. Such a large experiment would mean difficulties in financing, in obtaining an adequate experimental area, in controlling soil heterogeneity, and so on. Thus, the more practical approach would be to test the 30 clones first in a single-factor experiment, and then use the results to select a few clones for further studies in more detail. For example, the initial single-factor experiment may show that only five clones are outstanding enough to warrant further testing. These five clones could then be put into a factorial experiment with three levels of nitrogen, resulting in an experiment with 15 treatments rather than the 90 treatments needed with a factorial experiment with 30 clones.

The amount of change produced in the process output for a change in the 'level' of a given factor is referred to as the 'main effect' of that factor. Table 1 shows an example of a simple factorial experiment involving two factors with two levels each. The two levels of each factor may be denoted as 'low' and 'high', which are usually symbolized by '-' and '+' in factorial designs, respectively.

Table 1. A Simple 2-Factorial Experiment

	A (-)	**A (+)**
B (-)	20	40
B (+)	30	52

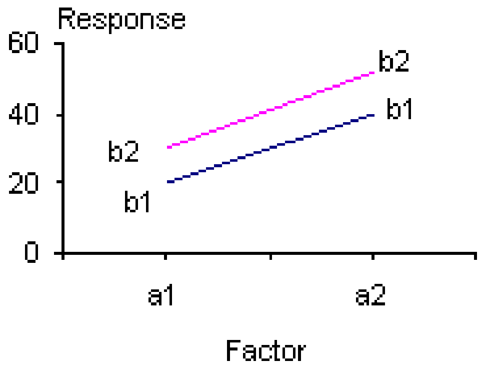
The main effect of a factor is basically the 'average' change in the output response as that factor goes from '-' to '+'. Mathematically, this is the average of two numbers: 1) the change in output when the factor goes from low to high level as the other factor stays low, and 2) the change in output when the factor goes from low to high level as the other factor stays high.

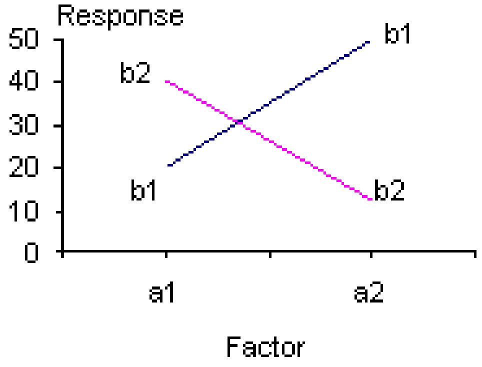
In the example in Table 1, the output of the process is just 20 (lowest output) when both A and B are at their '-' level, while the output is maximum at 52 when both A and B are at their '+' level. The main effect of A is the average of the change in output response when B stays '-' as A goes from '-' to '+', or (40-20) = 20, and the change in output response when B stays '+' as A goes from '-' to '+', or (52-30) = 22. The main effect of A, therefore, is equal to 21.

Similarly, the main effect of B is the average change in output as it goes from '-' to '+' , i.e., the average of 10 and 12, or 11. Thus, the main effect of B in this process is 11. Here, one can see that the factor A exerts a greater influence on the output of process, having a main effect of 21 versus factor B's main effect of only 11. It must be noted that aside from 'main effects', factors can likewise result in 'interaction effects.' Interaction effects are changes in the process output caused by two or more factors that are interacting with each other. Large interactive effects can make the main effects insignificant, such that it becomes more important to pay attention to the interaction of the involved factors than to investigate them individually. In Table 1, as effects of A (B) is not same at all the levels of B (A) hence, A and B are interacting.

Thus, **interaction** is the failure of the differences in response to changes in levels of one factor, to retain the same order and magnitude of performance through out all the levels of other factors OR the factors are said to interact if the effect of one factor changes as the levels of other factor(s) changes.

Graphical representation of lack of interaction between factors and interaction between factors are shown below. In case of two parallel lines, the factors are non-interacting.





If interactions exist which is fairly common, we should plan our experiments in such a way that they can be estimated and tested. It is clear that we cannot do this if we vary only one factor at a time. For this purpose, we must use multilevel, multifactor experiments.

The running of factorial combinations and the mathematical interpretation of the output responses of the process to such combinations is the essence of factorial experiments. It allows to understand which factors affect the process most so that improvements (or corrective actions) may be geared towards these.

We may define factorial experiments as experiments in which the effects (main effects and interactions) of more then one factor are studied together. In general if there are 'n' factors, say, F_1, F_2,..., F_n and i^{th} factor has s_i levels, i=1,...,n, then total number of treatment combinations is $\prod_{i=1}^{n} s_i$. Factorial experiments are of two types.

Experiments in which the number of levels of all the factors are same i.e all s_i's are equal are called **symmetrical factorial experiments** and the experiments in at least two of the s_i's are different are called as **asymmetrical factorial experiments**. Factorial experiments provide an opportunity to study not only the individual effects of each factor but also there interactions. They have the further advantage of economising on experimental resources. When the experiments are conducted factor by factor much more resources are required for the same precision than when they are tried in factorial experiments.

2. Experiments with Factors Each at Two Levels

The simplest of the symmetrical factorial experiments are the experiments with each of the factors at 2 levels. If there are 'n' factors each at 2 levels, it is called as a 2^n factorial where the power stands for the number of factors and the base the level of each factor. Simplest of the symmetrical factorial experiments is the 2^2 factorial experiment i.e. 2 factors say A and B each at two levels say 0 (low) and 1 (high). There will be 4 treatment combinations which can be written as

00 = $a_0 b_0$ = 1; A and B both at first (low) levels

10 = $a_1 b_0$ = a ; A at second (high) level and B at first (low) level

01 = $a_0 b_1$ = b ; A at first level (low) and B at second (high) level

11 = $a_1 b_1$ = ab; A and B both at second (high) level.

In a 2^2 factorial experiment wherein r replicates were run for each combination treatment, the main and interactive effects of A and B on the output may be mathematically expressed as follows:

A = [ab + a - b - (1)] / 2r; (main effect of factor A)

B = [ab + b - a - (1)] / 2r; (main effect of factor B)

AB = [ab + (1) - a - b] / 2r; (interactive effect of factors A and B)

where r is the number of replicates per treatment combination; a is the total of the outputs of each of the r replicates of the treatment combination a (A is 'high and B is 'low); b is the total output for the n replicates of the treatment combination b (B is 'high' and A is 'low); ab is the total output for the r replicates of the treatment combination ab (both A and B are 'high'); and (1) is the total output for the r replicates of the treatment combination (1) (both A and B are 'low').

Had the two factors been independent, then [ab + (1) - a - b] / 2n will be of the order of zero. If not then this will give an estimate of interdependence of the two factors and it is called the interaction between A and B. It is easy to verify that the interaction of the factor B with factor A is BA which will be same as the interaction AB and hence the interaction does not depend on the order of the factors. It is also easy to verify that the main effect of factor B, a contrast of the treatment totals is orthogonal to each of A and AB.

Table 2. Two-level 2-Factor Full-Factorial

RUN	Comb.	M	A	B	AB
1	(1)	+	-	-	+
2	a	+	+	-	-
3	b	+	-	+	-
4 = 2^2	ab	+	+	+	+

Consider the case of 3 factors A, B, C each at two levels (0 and 1) i.e. 2^3 factorial experiment. There will be 8 treatment combinations which are written as

000 = $a_0\, b_0\, c_0$ = (1); A, B and C all three at first level

100 = $a_1\, b_0\, c_0$ = a ; A at second level and B and C at first level

010 = $a_0\, b_1\, c_0$ = b ; A and C both at first level and B at second level

110 = $a_1\, b_1\, c_0$ = ab; A and B both at second level and C is at first level.

001 = $a_0\, b_0\, c_1$ = c ; A and B both at first level and C at second level.

101 = $a_1\, b_0\, c_1$ = ac; A and C at second level, B at first level

011 = $a_0\, b_1\, c_1$ = bc; A is at first level and B and C both at second level

111 = $a_1\, b_1\, c_1$ = abc; A, B and C all the three at second level

In a three factor experiment there are three main effects A, B, C; 3 first order or two factor interactions AB, AC, BC; and one second order or three factor interaction ABC.

Table 3. Two-level 3-Factor Full-Factorial Experiment Pattern

RUN	Comb.	M	A	B	AB	C	AC	BC	ABC
1	(1)	+	-	-	+	-	+	+	-
2	A	+	+	-	-	-	-	+	+
3	B	+	-	+	-	-	+	-	+
4	Ab	+	+	+	+	-	-	-	-
5	C	+	-	-	+	+	-	-	+
6	Ac	+	+	-	-	+	+	-	-
7	Bc	+	-	+	-	+	-	+	-
8 = 2^3	Abc	+	+	+	+	+	+	+	+

Main effect A = $\frac{1}{4}$ {[abc] -[bc] +[ac] -[c] + [ab] -[b] + [a] -[1]}

$$= \frac{1}{4}(a-1)\,(b+1)\,(c+1)$$

$$AB = \frac{1}{4}\ [(abc)-(bc) -(ac) +c) - (ab) - (b) - (a)+ (1)\ ]$$

$$ABC = \frac{1}{4}\ [\ (abc) - (bc) - (ac) + (c) - (ab) + (b) + (a) - (1)\ ]$$

or equivalently,

$$AB \quad = \ \frac{1}{4}\ (a\text{-}1)\ (b\text{-}1)\ (c+1)$$

$$ABC = \frac{1}{4}\ (a\text{-}1)\ (b\text{-}1)\ (c\text{-}1)$$

The method of representing the main effect or interaction as above is due to Yates and is very useful and quite straightforward. For example, if the design is 2^4 then

$$A \ = (1/2^3)\ \ [\ (a\text{-}1)\ (b+1)\ (c+1)\ (d+1)\ ]$$

$$ABC \ = \ (1/2^3)\ \ [\ (a\text{-}1)\ (b\text{-}1)\ (c\text{-}1)\ (d+1)]$$

In case of a 2^n factorial experiment, there will be 2^n (=v) treatment combinations with 'n' main effects, $\binom{n}{2}$ first order or two factor interactions, $\binom{n}{3}$ second order or three factor interactions, $\binom{n}{4}$ third order or four factor interactions and so on , $\binom{n}{r}$, (r-1)th order or r factor interactions and $\binom{n}{n}$ (n-1)th order or n factor interaction. Using these v treatment combinations, the experiment may be laid out using any of the suitable experimental designs viz. completely randomised design or block designs or row-column designs, etc.

Steps for Analysis

1. The Sum of Squares (S.S.) due to treatments, replications [in case randomised block design is used], due to rows and columns (in case a row-column design has been used), total S.S. and error S.S. is obtained as per established procedures. No replication S.S. is required in case of a completely randomised design.
2. The treatment sum of squares is divided into different components viz. main effects and interactions each with single d.f. The S.S. due to these factorial effects is obtained by dividing the squares of the factorial effect total by $r.2^n$. For obtaining 2^n-1 factorial effects in a 2^n factorial experiment, the 'n' main effects is obtained by giving the positive signs to those treatment totals where the particular factor is at second level and minus to others and dividing the value so obtained by $r.2^{n-1}$, where r is the number of replications of the treatment combinations. All

interactions can be obtained by multiplying the corresponding coefficients of main effects.

For a 2^2 factorial experiment, the S.S. due to a main effect or the interaction effect is obtained by dividing the square of the effect total by 4r. Thus,

S.S. due to main effect of A = $[A]^2/4r$, with 1 d.f.

S.S. due to main effect of B = $[B]^2/4r$, with 1 d.f

S.S. due to interaction AB = $[AB]^2/4r$, with 1 d.f.

3. Mean squares (M.S) is obtained by dividing each S.S. by corresponding degrees of freedom.
4. After obtaining the different S.S.'s, the usual Analysis of variance (ANOVA) table is prepared and the different effects are tested against error mean square and conclusions drawn.
5. Standard errors (S.E.'s) for main effects and two factor interactions:

S.E of difference between main effect means = $\sqrt{\dfrac{2MSE}{r.2^{n-1}}}$

S.E of difference between A means at same level of B=S.E of difference between B means at same level of A= $\sqrt{\dfrac{2MSE}{r.2^{n-2}}}$

In general,

S.E. for difference between means in case of a r-factor interaction = $\sqrt{\dfrac{2MSE}{r.2^{n-r}}}$

The critical differences are obtained by multiplying the S.E. by the student's t value at α% level of significance at error degrees of freedom.

The ANOVA for a 2^2 factorial experiment with r replications conducted using a RCBD is as follows:

ANOVA

Sources of Variation	DF	S.S.	M.S.	F
Between Replications	r-1	SSR	MSR=SSR/(r-1)	MSR/MSE
Between treatments	2^2-1=3	SST	MST=SST/3	MST/MSE
A	1	SSA=$[A]^2/4r$	MSA=SSA	MSA/MSE
B	1	SSB=$[B]^2/4r$	MSB=SSB	MSB/MSE
AB	1	SSAB=$[AB]^2/4r$	MSAB=SSAB	MSAB/MSE
Error	3(r-1)	SSE	MSE=SSE/3(r-1)	
Total	4r-1	TSS		

ANOVA for a 2^3-factorial experiment conducted in RCBD with r replications is given by

ANOVA

Sources of Variation	**DF**	**SS**	**MS**	**F**
Between Replications	r-1	SSR	MSR=SSR/(r-1)	MSR/MSE
Between treatments	2^3 -1=7	SST	MST=SST/7	MST/MSE
A	1	SSA	MSA=SSA	MSA/MSE
B	1	SSB	MSB=SSB	MSB/MSE
C	1	SSC	MSC=SSC	MSC/MSE
AB	1	SSAB	MSAB=SSAB	MSAB/MSE
AC	1	SSAC	MSAC=SSAC	MSAC/MSE
BC	1	SSBC	MSBC=SSBC	MSBC/MSE
ABC	1	SSABC	MSABC=SSABC	MSABC/MSE
Error	(r-1)(2^3-1) =7(r-1)	SSE	MSE=SSE/7(r-1)	
Total	r.2^3-1=8r-1	TSS		

Similarly ANOVA table for a 2^n factorial experiment can be made.

3. Experiments with Factors Each at Three Levels

When factors are taken at three levels instead of two, the scope of an experiment increases. It becomes more informative. A study to investigate if the change is linear or quadratic is possible when the factors are at three levels. The more the number of levels, the better, yet the number of the levels of the factors cannot be increased too much as the size of the experiment increases too rapidly with them. Consider two factors A and B, each at three levels say 0, 1 and 2 (3^2-factorial experiment). The treatment combinations are

00	= a_0b_0	= 1	; A and B both at first levels
10	= a_1b_0	= a	; A is at second level and B is at first level
20	= a_2b_0	= a^2	; A is at third level and b is at first level
01	= a_0b_1	= b	; A is at first level and B is at second level
11	= a_1b_1	= ab	; A and B both at second level
21	= a_2b_1	= a^2b	; A is at third level and B is at second level
02	= a_0b_2	= b^2	; A is at first level and B is at third level
12	= a_1b_2	= ab^2	; A is at second level and B is at third level
22	= a_2b_2	= a^2b^2	; A and B both at third level

Any standard design can be adopted for the experiment.

The main effects A, B can respectively be divided into linear and quadratic components each with 1 d.f. as A_L, A_Q, B_L and B_Q. Accordingly AB can be partitioned into four components as $A_L B_L$, $A_L B_Q$, $A_Q B_L$, $A_Q B_Q$.

The coefficients of the treatment combinations to obtain the above effects are given as

Treatment Totals→ Factorial Effects ↓	**[1]**	**[a]**	**[a^2]**	**[b]**	**[ab]**	**[a^{2}b]**	**[b^2]**	**[ab^2]**	**[a^2 b^2]**	**Divisor**
M	+1	+1	+1	+1	+1	+1	+1	+1	+1	$9r=r\times3^2$
A_L	-1	0	+1	-1	0	+1	-1	0	+1	$6r=r\times2\times 3$
A_Q	+1	-2	+1	+1	-2	+1	+1	-2	+1	$18r=6\times3$
B_L	-1	-1	-1	0	0	0	+1	+1	+1	$6r=r\times2\times3$
A_L B_L	+1	0	-1	0	0	0	-1	0	+1	$4r=r\times2\times2$
A_Q BL	-1	+2	-1	0	0	0	+1	-2	+1	$12r=r\times6\times2$
B_Q	+1	+1	+1	-2	-2	-2	+1	+1	+1	$18r=r\times3\times6$
A_L B_Q	-1	0	+1	+2	0	-2	-1	0	+1	$12r=r\times2\times6$
A_Q B_Q	+1	-2	+1	-2	+4	-2	+1	-2	+1	$36r=r\times6\times6$

The rule to write down the coefficients of the linear (quadratic) main effects is to give a coefficient as +1 (+1) to those treatment combinations containing the third level of the corresponding factor, coefficient as 0(-2) to the treatment combinations containing the second level of the corresponding factor and coefficient as -1(+1) to those treatment combinations containing the first level of the corresponding factor. The coefficients of the treatment combinations for two factor interactions are obtained by multiplying the corresponding coefficients of two main effects. The various factorial effect totals are given as

$$[A_L] = +1[a^2b^2]+0[ab^2] -1[b^2]+1[a^2b]+0[ab] -1[b]+1[a^2]+0[a] -1[1]$$

$$[A_Q] = +1[a^2b^2] -2[ab^2]+1[b^2]+1[a^2b] -2[ab]+1[b]+1[a^2] -2[a]+1[1]$$

$$[B_L] = +1[a^2b^2]+1[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2] -1[a] -1[1]$$

$$[A_LB_L]= +1[a^2b^2]+0[ab^2] -1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2]+0[a] -1[1]$$

$$[A_QB_L]= +1[a^2b^2] -2[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2]+2[a] -1[1]$$

$$[B_Q] = +1[a^2b^2]+1[ab^2]+1[b^2] -2[a^2b] -2[ab] -2[b] -1[a^2] -1[a] -1[1]$$

$$[A_LB_Q]= +1[a^2b^2]+0[ab^2] -1[b^2] -2[a^2b]+0[ab]+2[b]+1[a^2]+0[a] -1[1]$$

$[A_QB_Q] = +1[a^2b^2] -2[ab^2]+1[b^2] -2[a^2b]+4[ab] -2[b]+1[a^2] -2[a]+1[1]$

Factorial effects are given by

$A_L = [A_L]/r.3$ $A_Q= [A_Q]/r.3$ $B_L = [B_L]/r.3$ $A_LB_L = [A_LB_L]/r.3$

$A_QB_L = [A_QB_L]/r.3$ $B_Q = [B_Q]/r.3$ $A_LB_Q = [A_LB_Q]/r.3$ $A_QB_Q = [A_QB_Q]/r.3$

The sum of squares due to various factorial effects is given by

$$SSA_L = \frac{[A_L]^2}{r.2.3}; \qquad SSA_q = \frac{[A_Q]^2}{r.6.3}; \qquad SSB_L = \frac{[B_L]^2}{r.3.2};$$

$$SSA_LB_L = \frac{[A_LB_L]^2}{r.2.2};$$

$$SSA_QB_L = \frac{[A_QB_L]^2}{r.6.2}; \; SSB_Q = \frac{[B_Q]^2}{r.3.6}; \qquad SSA_LB_Q = \frac{[A_LB_Q]^2}{r..2.6};$$

$$SSA_QB_Q = \frac{[A_QB_Q]^2}{r.6.6};$$

If a RBD is used with r-replications then the outline of analysis of variance is

ANOVA

Sources of Variation	**D.f**	**SS**	**MS**
Between Replications	r-1	SSR	MSR=SSR/(r-1)
Between treatments	3^2-1=8	SST	MST=SST/8
A	2	SSA	MSA=SSA/2
A_L	1	SSA_L	$MSA_L= SSA_L$
A_Q	1	SSA_Q	$MSA_Q=SSA_Q$
B	2	SSB	MSB=SSB/2
B_L	1	SSB_L	$MSB_L= SSB_L$
B_Q	1	SSB_Q	$MSB_Q=SSB_Q$
AB	4	SSAB	MSAB=SSAB/2
A_LB_L	1	SSA_LB_L	$MSA_LB_L=SSA_LB_L$
A_QB_L	1	SSA_QB_L	$MSA_QB_L=SSA_QB_L$
A_LB_Q	1	SSA_LB_Q	$MSA_LB_Q=SSA_LB_Q$
A_QB_Q	1	SSA_QB_Q	$MSA_QB_Q=SSA_QB_Q$
Error	(r-1)(3^2-`1) =8(r-1)	SSE	MSE=SSE/8(r-1)
Total	r.3^2-1=9r-1	TSS	

In general, for n factors each at 3 levels, the sum of squares due to any linear (quadratic) main effect is obtained by dividing the square of the linear (quadratic) main effect total by r.2.3^{n-1}(r.6.3^{n-1}). Sum of squares due to a ‘p’ factor interaction is given by taking the square of the total of the particular interaction component divided by r.(a_1 a_2 ...a_p). 3^{n-p}, where a_1, a_2,...,a_p are taken as 2 or 6 depending upon the linear or quadratic effect of particular factor.

4. Confounding in Factorial Experiments

When the number of factors and/or levels of the factors increase, the number of treatment combinations increase very rapidly and it is not possible to accommodate all these treatment combinations in a single homogeneous block. For example, a 2^5 factorial would have 32 treatment combinations and blocks of 32 plots are quite big to ensure homogeneity within them. A new technique is therefore necessary for designing experiments with a large number of treatments. One such device is to take blocks of size less than the number of treatments and have more than one block per replication. The treatment combinations are then divided into as many groups as the number of blocks per replication. The different groups of treatments are allocated to the blocks.

There are many ways of grouping the treatments into as many groups as the number of blocks per replication. It is known that for obtaining the interaction contrast in a factorial experiment where each factor is at two levels, the treatment combinations are divided into two groups. Such two groups representing a suitable interaction can be taken to form the contrasts of two blocks each containing half the total number of treatments. In such case the contrast of the interaction and the contrast between the two block totals are given by the same function. They are, therefore, mixed up and can not be separated. In other words, the interaction has been confounded with the blocks. Evidently the interaction confounded has been lost but the other interactions and main effects can now be estimated with better precision because of reduced block size. This device of reducing the block size by taking one or more interaction contrasts identical with block contrasts is known as **confounding**. Preferably only higher order interactions, that is, interactions with three or more factors are confounded, because their loss is immaterial. As an experimenter is generally interested in main effects and two factor interactions, these should not be confounded as far as possible.

When there are two or more replications, if the same set of interactions are confounded in all the replications, confounding is called **complete** and if different sets of interaction are confounded in different replications, confounding is called **partial**. In complete confounding all the information on confounded interactions are lost. But in partial confounding, the confounded interactions can be recovered from those replications in which they are not confounded.

Advantages of Confounding

It reduces the experimental error considerably by stratifying the experimental material into homogeneous subsets or subgroups. The removal of the variation among incomplete blocks (freed from treatments) within replicates results in smaller error mean square as compared with a RBD, thus making the comparisons among some treatment effects more precise.

Disadvantages of Confounding

- In the confounding scheme, the increased precision is obtained at the cost of sacrifice of information (partial or complete) on certain relatively unimportant interactions.
- The confounded contrasts are replicated fewer times than are the other contrasts and as such there is loss of information on them and they can be estimated with a lower degree of precision as the number of replications for them is reduced.
- An indiscriminate use of confounding may result is complete or partial loss of information on the contrasts or comparisons of greatest importance. As such the experimenter should confound only those treatment combinations or contrasts which are of relatively less or of importance at all.
- The algebraic calculations are usually more difficult and the statistical analysis is complex, especially when some of the units (observations) are missing.

Confounding in 2^3 Experiment

Although 2^3 is a factorial with small number of treatment combinations but for illustration purpose, this example has been considered. Let the three factors be A, B, C each at two levels.

Factorial Effects → **Treat. Combinations ↓**	A	B	C	AB	AC	BC	ABC
(1)	-	-	-	+	+	+	-
(a)	+	-	-	-	-	+	+
(b)	-	+	-	-	+	-	-
(ab)	+	+	-	+	-	-	-
(c)	-	-	+	+	-	-	+
(ac)	+	-	+	-	+	-	-
(bc)	-	+	+	-	-	+	-
(abc)	+	+	+	+	+	+	+

The various factorial effects are as follows:

A = (abc) + (ac) + (ab) + (a) - (bc) - (c) - (b) - (1)

B = (abc) + (bc) + (ab) + (b) - (ac) - (c) - (a) - (1)

C = (abc) + (bc) + (ac) + (c) - (ab) - (b) - (a) - (1)

AB = (abc) + (c) + (ab) + (1) - (bc) - (ac) - (b) - (a)

AC = (abc) + (ac) + (b) + (1) - (bc) - (c) - (ab) - (a)

BC = (abc) + (bc) + (a) + (1) - (ac) - (c) - (ab) - (b)

ABC = (abc) + (c) + (b) + (a) - (bc) - (ac) - (ab) - (1)

Let the highest order interaction ABC be confounded and we decide to use two blocks of 4 units (plots) each per replicate.

Thus in order to confound the interaction ABC with blocks all the treatment combinations with positive sign are allocated at random in one block and those with negative signs in the other block. Thus the following arrangement gives ABC confounded with blocks and hence we loose information on ABC.

Replication I

Block 1:	(1)	(ab)	(ac)	(bc)
Block 2 :	(a)	(b)	(c)	(abc)

It can be observed that the contrast estimating ABC is identical to the contrast estimating block effects.

The other six factorial effects viz. A, B, C, AB, AC, BC each contain two treatments in block 1 (or 2) with the positive signs and two with negative sign so that they are orthogonal with block totals and hence these differences are not influenced among blocks and can thus be estimated and tested as usual without any difficulty. Whereas for confounded interaction, all the treatments in one group are with positive sign and in the other with negative signs.

Similarly if AB is to be confounded, then the two blocks will consists of

Block 1	(abc)	(c)	(ab)	(1)
Block 2	(bc)	(ac)	(b)	(a)

Here AB is confounded with block effects and cannot be estimated independently whereas all other effects A, B, C, AC, Bc and ABC can be estimated independently.

When an interaction is confounded in one replicate and not in another, the experiment is said to be partially confounded. Consider again 2^3 experiment with each replicate divided into two blocks of 4 units each. It is not necessary to confound the same interaction in all the replicates and several factorial effects may be confounded in one single experiment. For example, the following plan confounds the interaction ABC, AB, BC and AC in replications I, II, III and IV respectively.

Rep. I		**Rep. II**		**Rep. III**		**Rep. IV**	
Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**	**Block 7**	**Block 8**
(abc)	(ab)	(abc)	(ac)	(abc)	(ab)	(abc)	(ab)
(a)	(ac)	(c)	(bc)	(bc)	(ac)	(ac)	(bc)
(b)	(bc)	(ab)	(a)	(a)	(b)	(b)	(a)
(c)	(1)	(1)	(b)	(1)	(c)	(1)	(c)

In the above arrangement, the main effects A, B and C are orthogonal with block totals and are entirely free from block effects. The interaction ABC is completely confounded with blocks in replicate 1, but in the other three replications the ABC is orthogonal with blocks and consequently an estimate of ABC may be obtained from replicates II, III and IV. Similarly it is possible to recover information on the other confounded interactions AB (from I, III, IV), BC (from I, II, IV) and AC (from I, II, III). Since the partially confounded interactions are estimated from only a portion of the observations, they are determined with a lower degree of precision than the other effects.

For carrying out the statistical analysis, the various factorial effects and their S.S. are estimated in the usual manner with the modification that for **completely confounded** interactions neither the S.S due to confounded interaction is computed nor it is included in the ANOVA table. The confounded component is contained in the (2p-1) degrees of freedom (D.f.) (in case of p replicates) due to blocks. The partitioning of the d.f for a 2^3 completely confounded factorial is as follows.

Source of Variation	D.f
Blocks	2p-1
A	1
B	1
C	1
AB	1
AC	1
BC	1
Error	6(p-1)
Total	8p-1

In general for a 2^n completely confounded factorial in p replications, the different d.f's are given as follows

Source of Variation	D.f
Replication	p-1
Blocks within replication	$p(2^{n-r}-1)$
Treatments	$2^n-1-(2^{n-r}-1)$
Error	By subtraction
Total	$p2^n-1$

The treatment d.f has been reduced by $2^{n-r}-1$ as this is the total d.f confounded per block.

In case of partial confounding, we can estimate the effects confounded in one replication from the other replication in which it is not confounded. In (2^n, 2^r) factorial experiment with p replications, following is the splitting of d.f's.

Source of Variation	D.f
Replication	p-1
Blocks within replication	$p(2^{n-r}-1)$
Treatments	2^n-1
Error	By subtraction
Total	$p2^n-1$

The S.S. for confounded effects are to be obtained from those replications only in which the given effect is not confounded.

5. Fractional Factorial

In a factorial experiment, as the number of factors to be tested increases, the complete set of factorial treatments may become too large to be tested simultaneously in a single experiment. A logical alternative is an experimental design that allows testing of only a fraction of the total number of treatments. A design uniquely suited for experiments involving large number of factors is the fractional factorial. It provides a systematic way of selecting and testing only a fraction of the complete set of factorial treatment combinations. In exchange, however, there is loss of information on some pre-selected effects. Although this information loss may be serious in experiments with one or two factors, such a loss becomes more tolerable with large number of factors. The number of interaction effects increases rapidly with the number of factors involved, which allows flexibility in the choice of the particular effects to be sacrificed. In fact, in cases where some specific effects are known beforehand to be small or unimportant, use of the fractional factorial results in minimal loss of information.

In practice, the effects that are most commonly sacrificed by use of the fractional factorial are high order interactions - the four-factor or five-factor interactions and at times, even the three-factor interaction. In almost all cases, unless the researcher has prior information to indicate otherwise one should select a set of treatments to be tested so that all main effects and two-factor interactions can be estimated.

In forestry research, the fractional factorial is to be used in exploratory trials where the main objective is to examine the interactions between factors. For such trials, the most appropriate fractional factorials are those that sacrifice only those interactions that involve more than two factors.

With the fractional factorial, the number of effects that can be measured decreases rapidly with the reduction in the number of treatments to be tested. Thus, when the number of effects to be measured is large, the number of treatments to be tested, even with the use of fractional factorial, may still be too large. In such cases, further reduction in the size of the experiment can be achieved by reducing the number of replications. Although the use of fractional factorial without replication is uncommon in forestry experiments, when fractional factorial is applied to exploratory trials, the number of replications required can be reduced to the minimum.

Another desirable feature of fractional factorial is that it allows reduced block size by not requiring a block to contain all treatments to be tested. In this way, the homogeneity of experimental units within the same block can be improved. A reduction in block size is, however, accompanied by loss of information in addition to that already lost through the reduction in number of treatments.

DATA MINING: AN OVERVIEW

Shashi Dahiya
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
shashi.dahiya@icar.gov.in

Introduction

Rapid advances in data collection and storage technology have enables organizations to accumulate vast amounts of data. However, extracting useful information has proven extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small. In other situations, the questions that need to be answered cannot be addressed using existing data analysis techniques, and thus, new methods need to be developed.

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It has also opened up exiting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways. Data Mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation, such as predicting whether a newly arrived customer will spend more than Rs.1000 at a department store.

Data mining, or knowledge discovery, has become an indispensable technology for businesses and researchers in many fields. Drawing on work in such areas as statistics, machine learning, pattern recognition, databases, and high performance computing, data mining extracts useful information from the large data sets now available to industry and science.

Knowledge Discovery in Database

The transformation of data into knowledge has been using mostly manual methods for data analysis and interpretation, which makes the process of pattern extraction of databases too expensive, slow and highly subjective, as well as unthinkable if the volume of data is huge. The interest in automating the analysis process of great volumes of data has been fomenting several research projects in an emergent field called *Knowledge Discovery in Databases* (KDD). KDD is the process of knowledge extraction from great masses of data with the goal of obtaining meaning and consequently understanding of the data, as well as to acquire new knowledge. This process is very complex because it consists of a technology composed of a group of mathematical and technical models of software that are used to find patterns and regularities in the data.

Knowledge discovery in databases (KDD) has been defined as the *process of discovering valid, novel, and potentially useful patterns from data*. Let us examine these terms in more details:

- Data is a set of facts *F* (e.g. cases in databases).
- Pattern is an expression *E* in a language L describing facts in a subset $F_E$of *F*. *E* is called a pattern if it simpler than the enumeration of all facts in F_E.
- Process: Usually in KDD is a multi step process, which involves data preparation, search for patterns, knowledge evaluation, and refinement involving iteration after modification. The process is assumed to be non-trivial-that is, to have some degree of search autonomy.
- Validity: The discovered patterns should be valid on new data with some degree of certainty.
- Novel: The patterns are novel (at least to the system). Novelty can be measured with respect to changes in data (by comparing current values to previous or expected values) or knowledge (how a new finding is related to old ones). In general, it can be measured by a function *N (E, F)*, which can be a Boolean function or a measure of degree of novelty or unexpectedness.
- Potentially useful: The patterns should potentially lead to some useful actions, as measured by some utility function. Such a function U maps expressions in L to a partially or totally ordered measure space M_U: hence u=*U (E,F)*.
- Ultimately Understandable: A goal of KDD is to make patterns understandable to humans in order to facilitate a better understanding of the underlying data. While this is difficult to measure precisely, one frequent substitute is the simplicity measure. Several measure of simplicity exist, and they range form the purely syntactic to the semantic. It is assumed that this is measured, if possible, by a function S mapping expressions E in L to a partially or totally ordered space M_S: hence, *s= S (E, F)*.

An important notion, called interestingness, is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity. Some KDD systems have an explicit interestingness function $i = I(E, F, C, N, U, S)$ which maps expressions in L to a measure space M_I. Other systems define interestingness indirectly via an ordering of the discovered patterns.

Based on the notions given above, we can now make an attempt to define knowledge.

Knowledge: A pattern $E \in$ is called knowledge if for some user-specified threshold i $\in M_I$, $I(E, F, C, N, U, S) > i$.

This definition of knowledge is purely user-oriented and determined by whatever functions and thresholds the user chooses.

To extract knowledge from databases, it is essential that the *Expert* follows some steps or basic stages in order to find a path from the raw data to the desired knowledge. The KDD process organizes these stages in a sequential and iterative form. In this way, it would be interesting if the obtained results of these steps were analyzed in a more interactive and friendly way, seeking a better evaluation of these results. The process of knowledge extraction from databases combines methods and statistical tools, machine learning and databases to find a mathematical and/or logical description, which can be eventually complex, of patterns and regularities in data. The knowledge extraction from a large amount of data should be seen as an interactive and iterative process, and not as a system of automatic analysis.

The interactivity of the KDD process refers to the greater understanding, on the part of the users of the process, of the application domain. This understanding involves the selection of a representative data subset, appropriate pattern classes and good approaches to evaluating the knowledge. For a better understanding the functions of the users that use the KDD process can be divided in three classes:

(a) *Domain Expert*, who should possess a large understanding of the application domain;

(b) *Analyst*, who executes the KDD process and, therefore, he should have a lot of knowledge of the stages that compose this process and

(c) *Final User*, who does not need to have much knowledge of the domain, the *Final User* uses knowledge extracted from the KDD process to aid him in a decision-making process.

KDD Process: Knowledge discovery from data can be understood as a process that contains, at least, the steps of application domain understanding, selection and preprocessing of data, Data Mining, knowledge evaluation and consolidation and use of the knowledge. The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user. Practical view of the KDD process emphasizing the interactive nature of the process outlines the following basic steps:

- **Data Selection**: Where data relevant to the analysis task are retrieved from the database.
- **Data Preprocessing**: To remove noise and inconsistent data which is called cleaning and integration of data that is combining multiple data sources.
- **Data Transformation**: Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining**: An essential process where intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation**: To identify the truly interesting patterns representing knowledge based on some interestingness measures.
- **Knowledge Presentation**: Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

The several steps of KDD have been shown in the following figure.

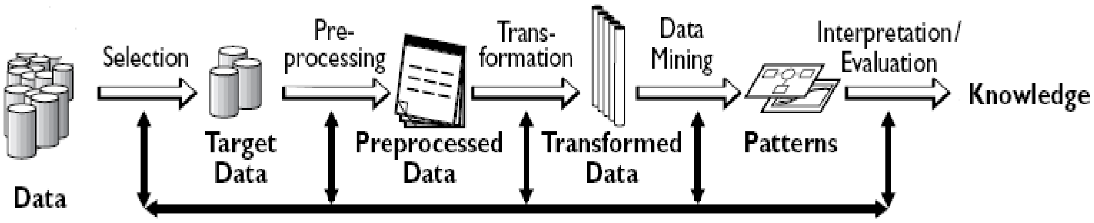


Figure: Various Steps of KDD process

The KDD process begins with the understanding of the application domain, considering aspects such as the objectives of the application and the data sources. Next, a representative sample (e.g. using statistical techniques) is removed from database, preprocessed and submitted to the methods and tools of the Data Mining stage with the objective of finding patterns/models (knowledge) in the data. This knowledge is then evaluated as to its quality and/or usefulness, so that it can be used to support a decision-making process.

The data mining component of the KDD process is mainly concerned with means by which patterns are extracted and enumerated from the data. Knowledge discovery involves the evaluation and possibly interpretation of the patterns to make the decision of what constitutes knowledge and what does not. It also includes of encoding schemes, preprocessing, sampling and projections of the data prior to the data mining step.

Data Mining

Generally, Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data Mining can be defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” and "the science of extracting useful information from large data sets or databases". Although it is usually used in relation to analysis of data, data mining, like artificial intelligence, is an umbrella term and is used with varied meaning in a range of wide contexts. It is usually associated with a business or other organization's need to identify trends.

Data Mining involves the process of analyzing data to show patterns or relationships; sorting through large amounts of data; and picking out pieces of relative information or patterns that occur e.g., picking out statistical information from some data.

The Data-Mining Communities: As data-mining has become recognized as a powerful tool, several different communities have laid claim to the subject:

1. Statistics.
2. AI, where it is called \machine learning."
3. Researchers in clustering algorithms.
4. Visualization researchers.
5. Databases.

In a sense, data mining can be thought of as algorithms for executing very complex queries on non-main-memory data.

Motivating Challenges

Traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by new data sets. The following are some of the specific challenges that motivated the development of data mining:

- **Scalability:** Because of advances in data generation and collection datasets with sizes of gigabytes, terabytes, or even petabytes are becoming common. If data mining algorithms are to handle these massive datasets, then they must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. For instance, out-of-core algorithms may be necessary when processing data sets that cannot fit into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

- **High Dimensionality:** It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. It the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken. Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data. Also, for some data analysis algorithms, the computational complexity increase rapidly as the dimensionality (the number of features) increases.
- **Heterogeneous and Complex Data:** Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. Recent years have also seen the emergence of more complex data objects. Examples of such non-traditional types of data include collections of Web pages containing semi-structured text and hyper lines; DNA data with sequential and three-dimensional structure; and climate data that consists of time series measurements (temperature, pressure, etc.) at various locations on the Earth’s surface. Techniques developed for mining such complex objects should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity, and parent-child relationships between the elements in semi-structures text and XML documents.
- **Data Ownership and Distribution:** Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple

entities. This requires the development of distributed data mining techniques. Among the key challenges faced distributed data mining algorithms include (1) how to reduce the amount of communication needed to perform the distributed computation, (2) how to effectively consolidate the data mining results obtained from multiple sources, and (3) how to address data security issues.

- **Non-Traditional Analysis:** The traditional statistical approach is based on a hypothesize-the test paradigm. In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analysed with respect to the hypothesis. Unfortunately, this process is extremely labor-intensive. Current data analysis tasks often require the generation and evaluated of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Furthermore, the data sets analyzed in data mining are typically nor the result of a carefully designed experiments and often represent opportunistic samples of the data, rather than random samples. Also, the data sets frequently involve non-traditional types of data and data distributions.

Data Preprocessing

Data preprocessing is a broad area and consists of a number of different strategies and techniques that are interrelated in complex ways. We will present some of the most important ideas and approaches, and try to point the interrelationships among them. The preprocessing techniques fall into two categories: selecting data objects and attributes for the analysis or creating/ changing the attributes. In both cases the goal is to improve the data mining analysis with respect to time, cost, and quality. Specifically, following are the important preprocessing techniques:

- **Aggregation:** Sometimes “less is more” and this is the case with aggregation, the combining of two or more objects into a single object.. Consider a dataset consisting of transactions (data objects) recording the daily sales of products in various store locations for different days over the course of a year. One way of aggregate the transactions of this data set is to replace all the transactions of a single store with a single storewide transaction. This reduces the hundreds or thousands of transactions that occur daily at a specific store to a single daily transaction, and the number of data objects is reduced to the number of stores.

An obvious issue is how an aggregate transaction is created; i.e. how the values of each attribute are combined across all the records corresponding to a particular location to create the aggregate transaction that represents the sales of a single store or date. Quantitative attributes, such as price, are typically aggregated by taking a sum or an average. A qualitative attribute, such as item, can either be omitted or summarized as the set of all the items that were sold at that location.

- **Sampling:** Sampling is a commonly used approach for selectinga subset of the data objects to be analyzed. In statistics, it has long been used for both the preliminary investigation of the data and the final data analysis. Sampling can also be very useful in data mining. However, the motivations for sampling in statistics and data mining are often different. Statisticians use sampling because obtaining the entire set of data of interest is too expensive or time consuming to process all the data. In some cases, using a sampling algorithm can reduce the data size to the point where a better, but more expensive algorithm can be used.
- **Dimensionality reduction:** Datasets can have a large number of feature. Consider set documents, where each documents is represented by a vector whose components are the frequencies with which each word occurs in the document. In such cases, there are typically thousands or tens of thousands of attributes (components), one for each word in the vocabulary. As another example, consider a set of time series consisting of the daily closing price of various stocks over a period of 30 days. In this case, the attributes, which are the prices on specific days again number in the thousands.

There is variety of benefits to dimensionality reduction. A key benefit is that many data mining algorithms work better if the dimensionality — the number of attributes in the data—is lower. This is partly because the dimensionality reduction can eliminateirrelevant features and reduce noise and partly because of the curse of dimensionality. Another benefit of dimensionality reduction is that a reduction of dimensionality can lead to a more understandable model because the model may involve fewer attributes. Also, dimensionality reduction may allow the data to be more easily visualized. Even if dimensionality reduction doesn’t reduce the data to two or three dimensions, data is often visualized by looking at pairs or triplets of attributes, and the number of such combinations is greatly reduced.

Finally, the amount of time and memory required by the data mining algorithms is reduced with a reduction in dimensionality.

- **Feature subset selection:** The term dimensionality reduction is often those techniques that reduce the dimensionality of data set by creating new attributes that are a combination of the old attributes. The reduction of dimensionality by selecting new attributes that are a subset of the old is known as feature subset selection or feature selection. While it might seem that such as approach would lose information, this is not the case if redundant and irrelevant features are present. Redundant features duplicate much or all the information contained in one or more other attributes. For example, the purchase price of a product and ge amount of sales tax paid contain much of the same information. Irrelevant features contain almost no useful information for the data mining task at hand. For instance, student’s ID numbers are irrelevant to the task of predicting student’s grade point averages. Redundant and irrelevant features can reduce classification accuracy and the quality of the clusters that are found.
- **Feature creation:** It is frequently possible to create, from the original attributes, a new set of attributes that captures the important information in a data set much more effectively. Furthermore, the number of new attributes can be smaller than the number of original attributes, allowing us to reap all the benefits of dimensionality reduction. Three related methodologies for creating new attributes are: feature extraction, mapping the data to a new space, and feature construction.
- **Discretization and Binarization:** Some data mining algorithms, especially certain classification algorithms, require that the data be in the form of categorical attributes. Algorithms that fine association patterns require that the data be in the form of binary attributes. Thus, it is often necessary to transform a continuous attribute into a categorical attribute (discretization), and both continuous and discrete attributes may need to be transformed into one or more binary attributes (binarization). Additionally, if a categorical attribute has a large number of values (categories), or some values occur infrequently,

then it may be beneficial for certain data mining tasks to reduce the number of categories by combining some of the values.

- **Variable transformation:** A variable transformation refers to a transformation that is applied to all the values of a variable. In other words, for each subject, the transformation is applied to the value of the variable for that object. For example, if only the magnitude of a variable is important, then the values of the variable can be transformed by taking the absolute value.

What kinds of Data can be Mined?

Data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data, and transactional data. Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the world wide web data). Techniques for mining of these kinds of data may be different. Data mining will certainly continue to embrace new data types as they emerge.

Tasks in Classical Data Mining

The two “high-level” primary goals of data mining in practice tend to be prediction and description. Data Mining tasks are generally divided into two major categories:

Predictive Tasks: the objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the *target* or *dependent variable*, while the attributes used for making the prediction are known as the *explanatory* or *independent variables*.

Descriptive Tasks: Here, the objective is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often explanatory in nature and frequently require post processing techniques to validate and explain and results.

The relative importance of prediction and description for particular data mining applications can vary considerably. However, in context of KDD, description tends to be more important than prediction.

Discovering patterns and rules: Other data mining applications are concerned with pattern detection. One example is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest. Another use is in astronomy, where detection of unusual stars or galaxies may lead to the discovery of previously unknown phenomenon. Yet another is the task of finding combinations of items that occur frequently in transaction databases (e.g., grocery products that are often purchased together). This problem has been the focus of much attention in data mining and has been addressed using algorithmic techniques based on association rules.

A significant challenge here, one that statisticians have traditionally dealt with in the context of outlier detection, is deciding what constitutes truly unusual behavior in the context of normal variability. In high dimensions, this can be particularly difficult. Background knowledge and human interpretation can be invaluable.

To achieve the goals of prediction and description, following data mining tasks are carried out.

- Classification
- Association Rule Mining
- Clustering
- Evolution Analysis
- Outlier Detection
- Dependency Modeling
- Change and Deviation Detection

1. Classification: Classification, which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications. Examples include, detecting spam email messages based upon the message header and content, categorizing cells as malignant or benign based upon the results of MRI scans, and classifying galaxies based upon their shapes.

The input data for a classification task is a collection of records. Each record, also known as an instance or example, is categorized by a tuple (x, y), where x is the attribute set and y is a special attribute, designated as the class label (also known as category or the target attribute). The attributes set in a dataset for classification can be either discrete or continuous but the class label must be a discrete attribute. This is the key characteristic that distinguishes classification from regression, a predictive modeling task in which y is a continuous attribute.

Definition (classification): Classification is the task of learning a target function *f* that maps each attribute set x to one of the predefined class labels y.

The target function is also known informally as a classification model. A classification model is useful for the following purposes.

Descriptive Modeling: A classification model can serve as an explanatory tool to distinguish between objects of different classes. For example, it would be useful-for both biologists and others-to have a descriptive model that summarizes that data shown… and explains what features define a vertebrate as a mammal, reptile, bird, fish, and amphibian.

Predictive Modeling: A classification model can also be used to predict the class label of unknown records. A classification model can be treated as a black box that automatically assigns a class label when presented with the attribute set of an unknown record.

Classification techniques are most suited for predicting or describing data sets with binary or nominal categories. They are less effective for ordinal categories (e.g., to classify a person as a member of high, medium or low income group) because they do not consider the implicit order among the categories. Other forms of relationships, such as subclass-super class relationships among categories (e.g., humans and apes are primates, which in turn is a subclass of mammals) are also ignored.

The classifier-training algorithm uses pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

2. Association Rule Mining: Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in 1993.It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared later. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub-problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence. Suppose one of the large item sets is L_k, $L_k = \{I_1, I_2, \dots, I_k\}$, association rules with this item sets are generated in the following way: the first rule is $\{I_1, I_2, \dots, I_{k-1}\}\Rightarrow\{I_k\}$, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem. The first sub-problem can be further divided into two sub-problems: candidate large item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item- sets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only “interesting” rules, generating only “non redundant” rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.

Methods for association rule mining:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

3. Clustering: Clustering or cluster analysis divides the data into groups (clusters) that are meaningful, useful or both. If meaningful groups are the goal, then the clusters should capture the natural structure of the data. In some cases, however, cluster analysis is only a useful starting point for other purposes, such as data summarization. Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering. There are various clustering methods:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

4. Evolution Analysis: Data evolution analysis describes and models regularities or trends for objects whose behaviors changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct feature of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

5. Outlier Detection: A database may contain data objects that do not comply with the general behavior or model of the data. Theses data objects are outliers. Most data mining methods discard outliers as noise as exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

6. Dependency modeling: Dependency modeling consists of finding a model that describes significant dependencies between variables. Dependency models exist at two levels: (1) the *structural level* of the model specifies (often in graphic form) which variables are locally dependent on each other and (2) the *quantitative level* of the model specifies the strengths of the dependencies using some numeric scale. For example, probabilistic dependency networks use conditional independence to specify the structural aspect of the model and probabilities or correlations to specify the strengths of the dependencies. Probabilistic dependency networks are increasingly finding applications in areas as diverse as the development of probabilistic medical expert systems from databases, information retrieval, and modeling of the human genome.

7. Change and deviation detection: Change and deviation detection focuses on discovering the most significant changes in the data from previously measured or normative values.

Components of Data Mining Algorithms

The data mining algorithms that address various data mining tasks have four basic components:

1. **Model or Pattern Structure:** Determining the underlying structure of functional forms that we seek from the data.
2. **Score Function:** Score functions are for judging the quality of a fitted model. Score Functions quantify how well a model or parameter structure fits a given data set. In an ideal world the choice of score function would precisely reflect the utility (i.e., the true expected benefit) of a particular predictive model. In practice, however, it is often difficult to specify precisely the true utility of a model's predictions. Hence, simple, "generic" score functions, such as least squares and classification accuracy are commonly used.
3. **Optimization and Search Method:** Optimizing the score function and searching over different model and pattern structures. The score function is a measure of how ell aspects of the data match proposed models or patterns. Usually, these models or patters are described in terms of a structure, sometimes with unknown parameter values. The goal of optimization and search is to determine the structure and the parameter values that achieve a minimum (or maximum, depending on the context) value of the score function. The task of finding the "best" values of parameters in models is typically cast as an optimization (for estimation) problem. The task of finding interesting patterns (such as rules) from a large family of potential patterns is typically cast as a combinatorial search problem, and is, often accomplished using heuristic search techniques. In linear regression, a prediction rule is usually found by minimizing a least squares score function (the sum of squared errors between the prediction from a model and the observed values of the predicted variable). Such a score function is amenable to mathematical manipulation, and the model that minimizes it can be found algebraically. In contrast, a score function such as misclassification rate in supervised classification is difficult to minimize analytically.
4. **Data Management Strategy:** Handling the data access efficiently during the search/optimization. The final component in any data mining algorithm is the data management strategy: the ways in which the data stored, indexed, and accessed. Most well-known data analysis algorithms in statistics and machine

learning have been developed under the assumption that all individual data points can be accessed quickly and efficiently in random-access memory(RAM), while main memory technology has improved rapidly, there have been equally rapid improvements in secondary (disk) and tertiary tape) storage technologies, to the extent that many massive data sets still reside largely on disk or tape and will not fit in available RAM. Thus, there will probably be a price to pay for accessing massive data sets, since not all data points can be simultaneously close to the main processor.

Some Challenges

A data mining system which is quick and correct on some small training sets, could behave completely different when applied to a larger database. A data mining system may work perfect for consistent data and may not perform well when a little noise is added to the training set. The most prominent challenges for data mining systems today are:

- Noisy Data
- Difficult Training Set
- Databases are Dynamic
- Databases may be Huge

Noisy Data: In a large database, many of the attribute values will be inexact or incorrect. This may be due to erroneous instruments measuring some property, or human error when registering it. We will distinguish between two forms of noise in the data, both described below:

Corrupted Values: Sometimes some of the values in the training set are altered from what they should have been. This may result in one or more tuples in the database conflict with the rules already established. The system may then regard these extreme values as noise, and ignore them. Alternatively, one may take the values into account possibly changing correct patterns recognized. The problem is that one never knows if the extreme values are correct or not, and the challenge is how to handle ``weird" values in the best manner.

Missing Attribute Values: One or more of the attribute values may be missing both for examples in the training set and for object which are to be classified. If attributes are missing in the training set, the system may either ignore this object totally, try to take it into account by for instance finding what is the missing attribute's most probable value, or use the value ``unknown" as a separate value for the attribute. When an attribute value is missing for an object during classification, the system may check all matching rules and calculate the most probable classification.

Difficult Training Set: Sometimes the training set is not the ultimate training set due to several reasons. These are the following:

Not Representative Data: If the data in the training set is not representative for the objects in the domain, we have a problem. If rules for diagnosing patients are being created and only elderly people are registered in the training set, the result for diagnosing a kid based on these data probably will not be good. Even though this may have serious consequences, we would say that not representative data is mainly a problem of machine learning when the learning is based on few examples. When using large data sets, the rules created probably are representative, as long as the data being classified belongs to the same domain as those in the training set.

No Boundary Cases: To find the real differences between two classes, some boundary cases should be present. If a data mining system for instance is to classify animals, the property counting for a bird might be that it has wings and not that it can fly. This kind of detailed distinction will only be possible if e.g. penguins are registered. Limited Information: In order to classify an object to a specific class, some condition attributes are investigated. Sometimes, two objects with the same values for condition attributes have a different classification. Then, the objects have some properties which are not among the attributes in the training set, but still make a difference. This is a problem for the system, which does not have any way of distinguish these two types of objects.

Databases are Dynamic: Databases usually change continually. We would like rules which reflect the content of the database at all times, in order to make the best possible classification. Many existing data mining systems require that all the training examples are given at once. If something is changed at a later time, the whole learning process may have to be conducted again. An important challenge for data mining systems is to avoid this, and instead change its current rules according to updates performed.

Databases may be Huge: The size of databases seem to be ever increasing. Most machine learning algorithms have been created for handling only a small training set, for instance a few hundred examples. In order to use similar techniques in databases thousands of times bigger, much care must be taken. Having very much data is advantageous since they probably will show relations really existing, but the number of possible descriptions of such a dataset is enormous. Some possible ways of coping with this problem, are to design algorithms with lower complexity and to use heuristics to find the best classification rules. Simply using a faster computer is seldom a good solution.

References

Breiman, L., Freidman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*. Wadsworth, Belmont CA.

Larose, DT.(2006).*Data Mining Methods and Models*. Wiley-Interscience, New Jersey, USA.

Han, J., Kamber, M., Pei, J. (2012).*Data mining: concepts and techniques.* Morgan Kaufmann, Elsevier, USA.

CLASSIFICATION AND REGRESSION TREE (CART) AND SELF ORGANIZING MAPS (SOM)

Ramasubramanian V.
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
r.subramanian@icar.gov.in

1. Introduction

In certain research studies, development of a reliable decision rule, which can be used to classify new observations into some predefined categories, plays an important role. The existing traditional statistical methods are inappropriate to use in certain specific situations, or of limited utility, in addressing these types of classification problems. There are a number of reasons for these difficulties. First, there are generally many possible “predictor” variables which makes the task of variable selection difficult. Traditional statistical methods are poorly suited for this sort of multiple comparisons. Second, the predictor variables are rarely nicely distributed. Many variables (in agriculture and other real life situations) are not normally distributed and different groups of subjects may have markedly different degrees of variation or variance. Third, complex interactions or patterns may exist in the data. For example, the value of one variable (e.g., age) may substantially affect the importance of another variable (e.g., weight). These types of interactions are generally difficult to model and virtually impossible to model when the number of interactions and variables becomes substantial. Fourth, the results of traditional methods may be difficult to use. For example, a multivariate logistic regression model yields a probability for different classes of the dependent variable, which can be calculated using the regression coefficients and the values of the explanatory variable. But practitioners generally do not think in terms of probability but, rather in terms of categories, such as “presence” versus “absence.” Regardless of the statistical methodology being used, the creation of a decision rule requires a relatively large dataset.

Classification methods include the conventional clustering methods (e.g. K-means), discriminant function method and SOFMs while predictive models include decision trees (e.g., CART - Classification And Regression Trees), neural networks (the most popular type of architectures being MLP – MultiLayer Perceptron) and statistical models (e.g. MLR - Multiple Linear Regression, Logistic regression etc.). Decision trees are nothing but classification systems that predict or classify future observations based on a set of decision rules and are sometimes called rule induction methods because the reasoning process behind them is clearly evident when browsing the trees. Neural network models are used when the underlying relationship between the different variables in the system are unknown (which are complex and typically non-linear). Self-Organizing Feature Maps (SOFMs) also known as Kohonen neural networks which comes under the category of unsupervised learning which are used when the study or main or dependent variable is a categorical variable and hence such networks are used for classification purposes.

The Kohonen architecture of neural networks is a special type of architecture and is totally different from other types and solely meant for classification rather than prediction. Kohonen network offers a considerably different approach to ANNs and are designed primarily for unsupervised learning rather than for supervised problems. The very first thing to be aware of while employing any classification method or prediction model is of ascertaining whether the nature of the problem requires a ‘supervised’ or an ‘unsupervised’ approach. The supervised problem occurs when there is a known membership class or output associated with each input in the ‘training’ data set i.e. the set upon which the method or model will be fitted or employed. The unsupervised problem means that one deals with a set of data which have no specific associated classes or outputs attached.

In this write-up, two chief methods viz., CART and SOM in the context of classification (i.e. when the main or study or dependent variable is categorical) are discussed in detail.

1. Classification And Regression Tree (CART)

CART analysis is a tree-building technique which is different from traditional data analysis methods. In a number of studies, CART has been found to be quite effective for creating decision rules which perform as well or better than rules developed using more traditional methods aiding development of DSS (Decision Support Systems). In addition, CART is often able touncover complex interactions between predictors which may be difficult or impossible using raditional multivariate techniques. It is now possible to perform a CART analysis with a simple understanding of each of the multiple steps involved in its procedure. Classification tree methods such as CART are convenient way to produce a prediction rule from a set of observations described in terms of a vector of features and a response value. The aim is to define a general prediction rule which can be used to assign a response value to the cases solely on the bases of their predictor (explanatory) variables. Tree-structured classifications are not based on assumptions of normality and user-specified model statements, as are some conventional methods such as discriminant analysis and ordinary least square regression.

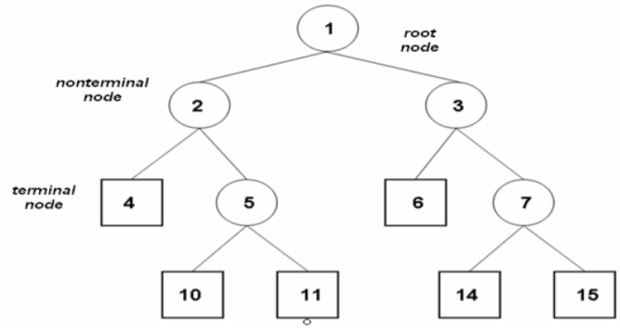
Tree based classification and regression procedure have greatly increased in popularity during the recent years. Tree based decision methods are statistical systems that mine data to predict or classify future observations based on a set of decision rules and are sometimes called rule induction methods because the reasoning process behind them is clearly evident when browsing the trees. The CART methodology have found favour among researchers for application in several areas such as agriculture, medicine, forestry, natural resources management etc. as alternatives to the conventional approaches such as discriminant function method, multiple linear regression, logistic regression etc. In CART, the observations are successively separated into two subsets based on associated variables significantly related to the response variable; this approach has an advantage of providing easily comprehensible decision strategies. CART can be applied either as a classification tree or as a regressive tree depending on whether the response variable is categorical or continuous. Tree based methods are not based on any stringent assumptions. These methods can handle large number of variables, are resistant to outliers, non-parametric, more versatile, can handle categorical variables, though computationally more intensive. They can be applied to data sets having both a large number of cases and a large number of variables, and are extremely robust to outliers. These are not based on assumptions such as normality and user-specified model statements, as are some conventional methods such as discriminant analysis or ordinary least square (OLS) regression. Yet, unlike the case for other nonparametric methods for classification and regression, such as kernel-based methods and nearest neighbor methods, the resulting tree-structured predictors can be relatively simple functions of the predictor variables which are easy to use.

CART can be a good choice for the analysts as they give fairly accurate results quickly, than traditional methods. If more conventional methods are called for, trees can still be helpful if there are a lot of variables, as they can be used to identify important variables and interactions. These are also invariant to the monotonic transformations of the explanatory variables and do not require the selection of the variable in advance as in regression analysis.

Agriculture being a highly uncertain occupation, classification and prediction in the field of agriculture aid planners to take proactive measures. Keeping in view the requirements to develop a sound classificatory system and that the potentials of the tree based methods for this purpose has not fully been explored, it will be of interest to employ these methodologies upon a suitable data set in the field of agriculture. More importantly, since the real world data often does not satisfy the usual assumptions like that of normality, homoscedasticity etc it can be taken up as a motivation to find such a classificatory rule where assumptions of such rules fail. Apart from all these, tree based methods are one among the promising data mining tools that provide easily comprehensible decision strategy.

Tree based applications originated in the 1960s with the development of AID (Automatic Interaction Detector) by Morgan and Sonquistin the 1960s as regression trees. Further modifications in this technique was carried out to result in THAID (THeta AID) by Morgan and Messenger (1973) to produce classification trees and CHAID (CHi AID) by Kass in the late 1970s.Breiman*et al.*(1984) developed CART (Classification and Regression Trees) which is a sophisticated program for fitting trees to data. Breiman, again in 1994, developed the bagging predictors which is a method of generating multiple versions of a predictor and using them to get an aggregated predictor. A good account of the CART methodology can be found in many recent books, say, Izenman (2008).An application of classification trees in the field of agriculture can be found in Sadhu *et al.* (2014).

Theconventional CART methodologyis outlined briefly. Following is a schematic representation of aconventional CART tree structure:



The unique starting point of,say, a classification tree, is called a root node and consists of the entire learning set $\mathcal{L}$ at the top of the tree. A node is a subset of the set of variables, and it can be terminal or nonterminal node. A nonterminal (or parent) node is a node that splits into two left and right child nodes (binary split). Such a binary split is determined by a condition on the value of a single variable, where the condition is either satisfied or not satisfied by the observed value of that variable. All observations in $\mathcal{L}$ that have reached a particular (parent) node and satisfy the condition for that variable drop down to one of the two *child* nodes; the remaining observations at that (parent) node that do not satisfy the condition drop down to the other *child* node. A node that does not split is called a terminal node and is assigned a class label. Each observation in $\mathcal{L}$ falls into one of the terminal nodes. When an observation of unknown class is “dropped down” the tree and ends up at a terminal node, it is assigned the class corresponding to the class label attached to that node. There may be more than one terminal node with the same class label. To produce a tree-structured model using recursive binary partitioning, CART determines the best split of the learning set $\mathcal{L}$ to start with and thereafter the best splits of its subsets on the basis of various issues such as identifying which variable should be used to create the split, and determining the precise rule for the split, determining when a node of the tree is a terminal one, and assigning a predicted class to each terminal node. The assignment of predicted classes to the terminal nodes is relatively simple, as is determining how to make the splits, whereas determining the right-sized tree is not so straightforward. After growing a fully expanded tree, a tree of optimum size is obtained. In a particular type of tree building called 'exhaustive search', at each stage of recursive partitioning, all of the allowable ways of splitting a subset of $\mathcal{L}$ are considered, and the one which leads to the greatest increase in node purity is chosen. This can be accomplished using what is called an "impurity function", which is nothing but a function of the proportion of the learning sample belonging to the possible classes of the response variable. To choose the best split over all variables, first the best split for a given variable has to be determined. To assess the goodness of a potential split, the value of the 'impurity function' such as Gini diversity index and the Entropy function can be calculated using the cases in the learning sample corresponding to the parent node, and subtract from this the weighted average of the impurity for the two *child* nodes, with the weights proportional to the number of cases of the learning sample corresponding to each of the *child* nodes, to get the decrease in the overall impurity that would result from the split. To select the way to split a subset of$\mathcal{L}$in the tree growing procedure, all allowable ways of splitting can be considered, and the one which will result in the greatest decrease in node impurity (or, in other words, greatest increase in the node purity) can be chosen.

In order to grow a tree, the starting point is the root node, which consists of the learning set$\mathcal{L}$. Using the "goodness of split" criterion for a single variable, the tree algorithm finds the best split at the root node for each of the variables. The best split s at the root node is then defined as the one that has the largest value of this goodness of split criterion over all single-variable best splits at that node. Next is to split each of the *child* nodes of the root node in the same way. The above computations are repeated for each of the *child* nodes except that this time only the observations in that specific *child* node are considered for the calculations rather than all the observations. When these splits are completed, the splitting is continued with the subsequent nodes. This sequential splitting procedure of building a tree layer-by-layer is hence called recursive partitioning. If every parent node splits in two *child* nodes, the result is called a binary tree. If the binary tree is grown until none of the nodes can be split any further, then the tree is said to be saturated. Usually, first a very large tree is grown, splitting subsets in the current partition of $\mathcal{L}$ even if a split does not lead to an appreciable decrease in impurity. Then a sequence of smaller trees can be created by "pruning" the initial large tree, where in the pruning process, splits that were made are removed and a tree having a fewer number of nodes is produced. The crucial part of creating a good tree-structured classification model is determining how complex the tree should be. If nodes continue to be created until no two distinct values of the independent variables for the cases in the learning sample belong to the same node, the tree may be over fitting the learning sample and not be a good classifier of future cases. On the other hand, if a tree has only a few terminal nodes, then it may be that it is not making enough use of information in the learning sample, and classification accuracy for future cases will suffer. Initially, in the tree-growing procedure, the predictive accuracy typically increases as more nodes are created and the partition gets finer. But it is usually seen that at some point the misclassification rate for future cases will start to get worse as the tree becomes more complex. In order to compare the prediction accuracy of various tree-structured models, there needs to be a way to estimate a given tree's misclassification rate for the future observations, a measure named 'resubstitution estimate' of the misclassification rate is obtained by using the tree to classify the members of the learning sample (that were used to create the tree), and observing the proportion that are misclassified. More often, a better estimate of a tree's misclassification rate can be obtained using an independent "test set", which is a collection of cases coming from the same population or distribution as the learning set. Like the learning set, for the test set the true class for each case is known in addition to the values for the predictor variables. The test set estimate of the misclassification rate is just the proportion of the test set cases that are misclassified when predicted classes are obtained using the tree created from the learning set. The learning set and the test set are both composed of cases for which the true class is known in addition to the values for the predictor variables. Generally, about one third of the available cases should be set aside to serve as a test set, and the rest of the cases should be used as learning set. But sometimes a smaller fraction, such as one tenth, is also used and then resorting to 10-fold cross validation. A specific way to create a useful sequence of different-sized trees is to use “minimum cost-complexity pruning”. In this process, a nested sequence of subtrees of the initial large tree is created by “weakest-link cutting”. With weakest-link cutting (pruning), all of the nodes that arise from a specific nonterminal node are pruned off (leaving that specific node itself as terminal node), and the specific node selected is the one for which the corresponding pruned nodes provide the smallest per node decrease in the resubstitution misclassification rate. If two or more choices for a cut in the pruning process would produce the same per node decrease in the resubstitution misclassification rate, then pruning off the largest number of nodes is preferred. The sequence of subtrees produced by the pruning procedure serves as the set of candidate subtrees for the model, and to obtain the classification tree, all that remains to be done is to select the one which will hopefully have the smallest misclassification rate for future observations. The selection of final tree is based on estimated misclassification rates, obtained using a test set or by cross validation.

1. Self Organizing Map (SOM)

In SOM, the training data set contains only input variables and no outputs. It is a 'self-organizing' system, which automatically adapts itself in such a way that similar input objects are associated with the topological close neurons in the ANN. The phrase 'topological close neurons' means that neurons that are physically located close to each other will react similar to similar inputs, while the neurons that are far apart in the lay-out of the ANN will react quite different to similar inputs. A practical treatment on SOFM based Kohonen networks can be found in Haykin (1996).

The principal goal is to transform an incoming input pattern of arbitrary dimension into a two dimensional discrete map. Neurons in the network are arranged in a two dimensional grid and there happens a competition among these neurons to represent the input pattern. The ‘winning’ neurons and the similar pattern neurons i.e. the neighboring neurons are placed in contiguous locations in output space. The neurons learn to pin-point the location of the neuron in the ANN that is most 'similar' to the input vector. Here, the phrase 'location of the most similar neuron' has to be taken in a very broad sense. It can mean the location of the closest neuron with the smallest or with the largest Euclidean distance to the input vector, or it can mean the neuron with the largest output in the entire network for this particular input vector etc. In other words, in the Kohonen network, a ‘rule’ deciding which of all neurons will be selected after the input vector enters the ANN is mandatory. During the training in the Kohonen’s ANN, the multidimensional neurons self-organise themselves in the two-dimensional plane in such a way that the objects from the multidimensional measurement space are mapped into the plane of neurons with respect to some internal property correlated to the m-dimensional measurement space of objects.

Bullinaria (2004) has explained the above discussion in the following manner. Neurons are placed at the nodes of a lattice that is usually two-dimensional and undergo the following three steps:

(i) Competition

Neurons become selectively tuned to various input patterns (stimuli). Such “winning” neurons become ordered w.r. to each other in such a way that a meaningful coordinate system for different input features is created over the lattice. The competitive learning is characterized by formation of a topographic map of the inputs in which spatial locations of the neurons in the lattice are indicative of intrinsic features contained in the inputs, hence the name SOFM.

(ii) Cooperation

The winning neurons determines the spatial location of a topographic neighbourhood of excited neurons, thereby providing the basis for cooperation

(iii) Adaptation

The excited neurons adapts their individual values of its functional form in relation to the input pattern through suitable adjustments applied to their synaptic weights. Thus the response of the winning neuron to the subsequent application of a similar input pattern is enhanced

The correction of weights is carried out after the input of each input object in the following four steps:

(i) the neuron with the most ’distinguished’ response of all (in a sense explained above) is selected and named the ’central’ or the ’most excited’ neuron

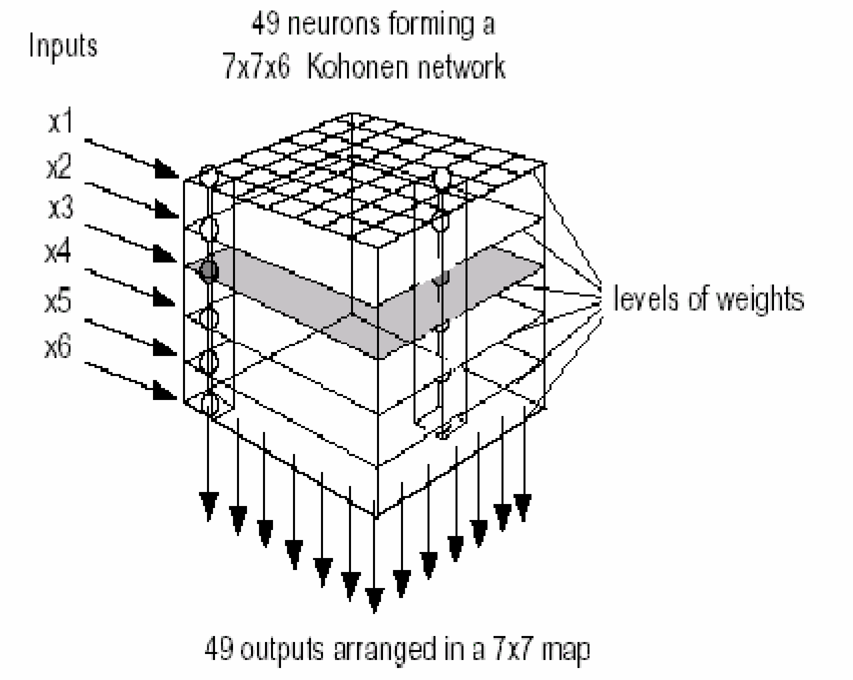
(ii) the maximal neighbourhood around this central neuron is determined.

(iii) the ‘correction factor’ is calculated for each neighbourhood ring separately (the correction changes according to the distance and time of training)

(iv) the ‘weights’ in neurons of each neighbourhood are corrected according to a pre-specified equation

The most important difference is that the neurons in the error back propagation learning (in that of the most famous multi-layer perceptron type of architectured neural network) tries to yield quantitatively an answer as close as possible to the target, while in the Kohonen approach the neurons learn to pin-point the location of the neuron in the ANN that is most ’similar’ to the input vector.

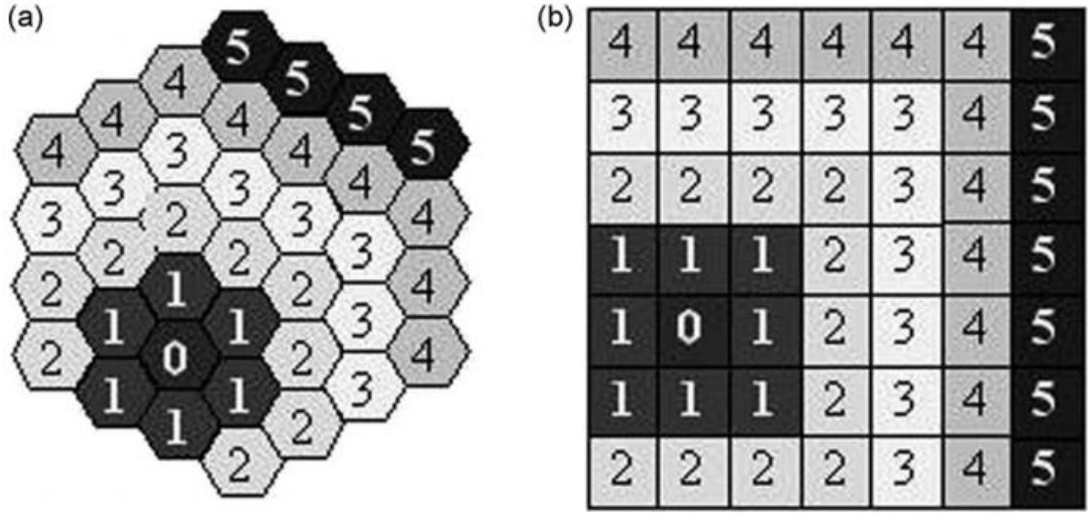
In order to make things clear, let us consider the following figure wherein there are six input variables along with a two-dimensional map of order 7x7. The neurons are in the columns associating the input variables with the (i, j)-th neuron in the output map, with weights at various levels corresponding to the inputs. That is, because the Kohonen ANN has only one layer of neurons, the specific input variable, let us say the i-th variable x_i is always received in all neurons of the ANN by the weight placed in the i-th position. If the neurons are presented as columns of weights then all i-th weights in all neurons can be regarded as the weights of the i-th level (Zupan, 1994).



Because the Kohonen ANN has only one layer of neurons the specific input variable, let us say the i-th variable, x_i, is always received in all neurons of the ANN, by the weight placed at the i-th position. If the neurons are presented as columns of weights then all i-th weights in all neurons can be regarded as the weights of the i-th level. This is especially important because the neurons are usually ordered in a two-dimensional formation.

Thus the main goal of Kohonen is to perform a non-linear mapping from an high-dimensional variable space to a low-dimensional (usually 2D) target space so that the distance and proximity relations between the samples or, in a single word, the topology, are preserved. The target space used in Kohonen mapping is a two-dimensional array of neurons fully connected to the input layer, onto which the samples are mapped. Introducing the preservation of topology, results in specifying for each node in the Kohonen layer, a defined number of neurons as nearest neighbors, second-nearest neighbors and so on.

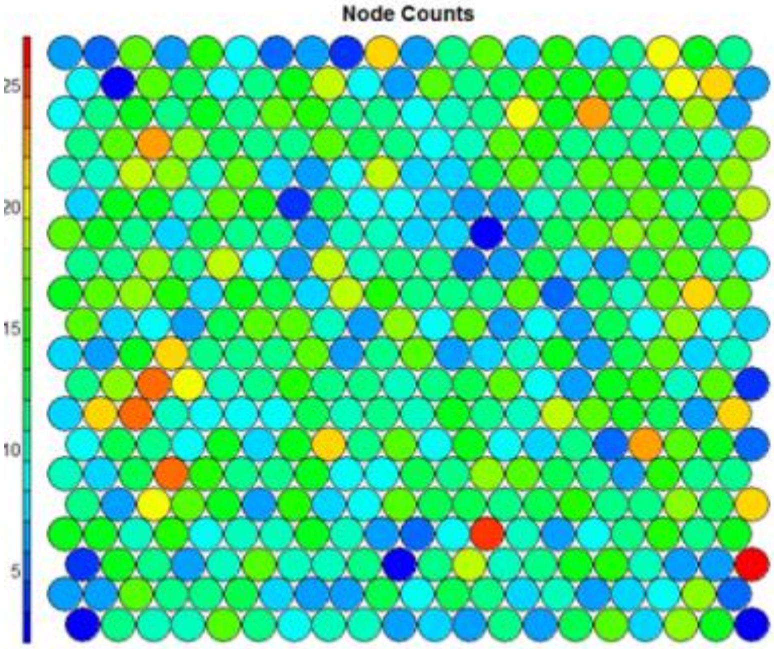
The layout of neurons in the Kohonen ANN is an important feature to be discussed (Marini *et al.*, 2007). The neighborhood of a neuron is usually considered to be hexagonal [see (a) in figure below] or square [see (b) in figure below] which means that each neuron has eight or six nearest neighbors, respectively.



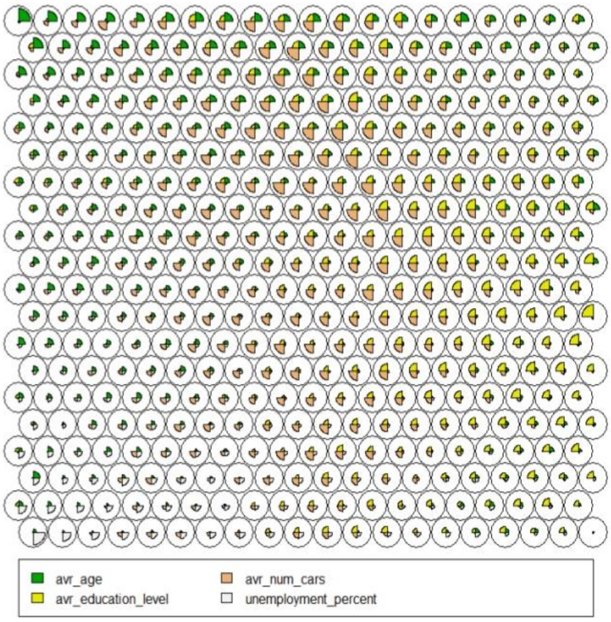
4	4	4	4	4	4	5
3	3	3	3	3	4	5
2	2	2	2	3	4	5
1	1	1	2	3	4	5
1	0	1	2	3	4	5
1	1	1	2	3	4	5
2	2	2	2	3	4	5

The main issue in Kohonen learning is that similar input vectors excite neurons which are very close in the 2D layer. From an algorithmic point of view, Kohonen mapping implements competitive learning, i.e. only one neuron in the 2D layer is selected after each input is presented to the network (winner takes-all). The winning neuron c is selected as the one having the weight vector most similar to the input pattern. After the winning neuron in the Kohonen layer is selected, the weights of each other neuron in the Kohonen layer are updated on the basis of the difference between their old value and the values of the input vector; this correction is scaled according to the topological distance from the winner.

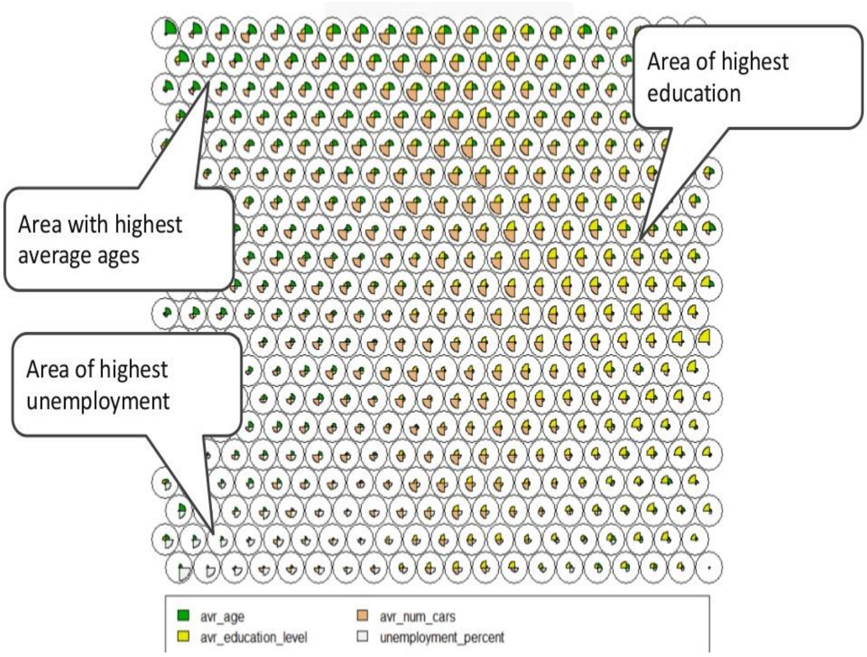
Lynn (2014) have extensively discussed about the theKohonen package available in the open source and freely available R software. This Kohonen R package allows us to visualise the count of how many samples are mapped to each node on the map. This metric can be used as a measure of map quality – ideally the sample distribution is relatively uniform. Large values in some map areas suggests that a larger map would be benificial. Empty nodes indicate that the map size is too big for the number of samples. He suggest that one should aim for at least 5-10 samples per node when choosing map size. One such output where node counts are visualized is given subsequently.



The node weight vectors, or "codes", are made up of normalised values of the original variables used to generate the SOM. Each node's weight vector is representative / similar of the samples mapped to that node. By visualising the weight vectors across the map, we can see patterns in the distribution of samples and variables. Such a visualisation of the weight vectors can be done using a "fan diagram", where individual fan representations of the magnitude of each variable in the weight vector is shown for each node. One such fan diagram is given below.



- Fan diagram shows distribution of variables across map.
- Can see patterns by examining dominant colours etc.
- This type of representation is useful for SOMs when the number of variables is less than ~ 5
- Good to get a grasp of general patterns in SOM



Area of highest education

Area with highest average ages

Area of highest unemployment

References

Breiman, L., Freidman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*. Wadsworth, Belmont CA.

Bullinaria, J.A. (2004). https://www.cs.bham.ac.uk/~jxb/NN/l16.pdf, accessed on 29th February, 2020.

Haykin, S. (1996). *Neural networks: A comprehensive foundation*, Pearson Education, Asia.

Izenman, A.J. (2008). *Modern multivariate statistical techniques: Regression, classification and manifold learning*. Springer, New York.

Lynn, S. (2014). https://www.slideshare.net/shanelynn/2014-0117-dublin-r-selforganising-maps-for-customer-segmentation-shane-lynn, accessed on 23rd December, 2019.

Marini, F., Magri, A. L., Bucci, R. and Magri, A.D. (2007). Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils, *AnalyticaChimicaActa*, **599**, 232–240.

Morgan, J.N. and Messenger, R.C. (1973). THAID: a sequential search program for the analysis of nominal scale dependent variables. Institute for Social Research, University of Michigan, Ann Arbor, MI.

Sadhu, S.K., Ramasubramanian, V., Rai, A. and Kumar, A. (2014). Decision tree based models for classification in agricultural ergonomics, *Statistics and Applications*, **12(1&2)**, 21-33.

Zupan, J. (1994). Introduction to Artificial Neural Network (ANN) methods: What they are and how to use them, *Acta Chimica Slovenica* ,**41 (3)**, 327-352.

CLUSTER ANALYSIS USING R

Alka Arora
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
Alka.Arora@icar.gov.in

1. Introduction

Clustering algorithms maps the data items into clusters, where clusters are natural grouping of data items based on similarity methods. Unlike classification and prediction which analyzes class-label data objects, clustering analyzes data objects without class-labels and tries to generate such labels. Clustering has many applications. In business/ marketing, clustering can help in identifying different customer groups and appropriate marketing campaign can be carried out targeting different groups. In agriculture, it can be used to derive plant and animal taxonomies, characterization of diseases and varieties, in bioinformatics- categorization of genes with similar functionally. Further it can be used to group similar documents on the web for faster discovery of content. It can be used to group geographical locations based on crime, amenities, weather etc. As data mining function, cluster analysis is used to gain insight into distribution of data, to observe the characteristics of each cluster and to focus on a particular set of clusters for further analysis.

1. Similarity Measures

Similarity is fundamental to majority of clustering algorithms. *Similarity is quantity that reflects the strength of relationship between two objects or two features.* This quantity is usually having range of either -1 to +1 or normalized into 0 to 1. If the similarity between feature *i* and feature *j* is denoted by s_{ij} , we can measure this quantity in several ways depending on the scale of measurement (or data type) that we have. Dissimilarity is opposite to similarity. There are many types of distance and similarity measures.

Similarity and dissimilarity can be measured for two objects based on several features/ variables. After the distance or similarity of each variable is determined, we can aggregate all features/ variables together into single Similarity (or dissimilarity) index between the two objects.

2.1 Distance for binary variables

We often face variables that only binary value such as Yes and No, or Agree and Disagree, True and False, Success and Failure, 0 and 1, Absence or Present, Positive and Negative, etc. Similarity of dissimilarity (distance) of two objects that represented by binary variables can be measured in term of number of occurrence (frequency) of positive and negative in each object.

For example:

Feature of Fruit	**Sphere shape**	**Sweet**	**Sour**	**Crunchy**
Object i =Apple	Yes	Yes	Yes	Yes
Object j =Banana	No	Yes	No	No

The coordinate of Apple is (1,1,1,1) and coordinate of Banana is (0,1,0,0). Because each object is represented by 4 variables, we say that these objects have 4 dimensions.

Let *p* = number of variables that positive for both objects .

q = number of variables that positive for the *i* th objects and negative for the *j* th object

r= number of variables that negative for the i th objects and positive for the j th object

s= number of variables that negative for both objects

t= *p*+*q*+*r*+*s* = total number of variables.

		Object j	
		Yes	**No**
object i	**Yes**	p	q
	No	r	s

For our example above, we have measured Apple and Banana have *p=1*, *q=3* and *r=0*, *s=0*. Thus, *t= p+q+r+s=4.*

The most common use of binary dissimilarity (distance) is

Simple Matching distance $d_{ij} = \frac{q+r}{t}$

Jaccard's distance $d_{ij} = \frac{q+r}{p+q+r}$

Hamming distance $d_{ij} = q+r$

Example: Simple matching distance between Apple and Banana is 3/4.

Jaccard's distance between Apple and Banana is 3/4.

Hamming distance between Apple and Banana is 3.

2.2 Distance for quantitative variables

Variable which have quantitative values.

	Features k			
	cost	time	weight	incentive
Object A	0	3	4	5
Object B	7	6	3	-1

We can represent the two objects as points in 4 dimension. Point A has coordinate (0, 3, 4, 5) and point B has coordinate (7, 6, 3, -1). Dissimilarity (or similarity) between the two objects are based on these coordinates.

Euclidean Distance: Euclidean Distance is the most common use of distance. In most cases when people said about distance, they will refer to Euclidean distance. Euclidean distance or simply 'distance' examines the *root of square differences* between coordinates of a pair of objects.

Formula $d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$

$$d_{BA} = \sqrt{(0-7)^2 + (3-6)^2 + (4-3)^2 + (5+1)^2}$$
$$= \sqrt{49+9+1+36} = 9.747$$

Euclidean distance is a special case of Minkowski distance with $\lambda = 2$

City block (Manhattan) distance : It is also known as *Manhattan* distance, *boxcar* distance, *absolute value* distance. It examines the *absolute differences* between coordinates of a pair of objects. City block distance is a special case of Minkowski distance with $\lambda = 1$

Formula: $d_{ij} = \sum_{k=1}^{n} \left|x_{ik} - x_{jk}\right|$

The City Block Distance between point A and B is

$$d_{BA} = |0-7| + |3-6| + |4-3| + |5+1|$$
$$= 7+3+1+6 = 17$$

Chebyshev Distance : Chebyshev distance is also called Maximum value distance. It examines the *absolute magnitude of the differences* between coordinates of a pair of objects. This distance can be used for both ordinal and quantitative variables.

Formula $d_{ij} = \max_{k} |x_{ik} - x_{jk}|$ and B is

$$d_{BA} = \max\{|0-7|,|3-6|,|4-3|,|5+1|\}$$
$$= \max\{7,3,1,6\} = 7$$

Minkowski Distance: This is the generalized metric distance. When $\lambda = 1$ it becomes city block distance and when $\lambda = 2$, it becomes Euclidean distance. Chebyshev distance is a special case of Minkowski distance with $\lambda = \infty$ (taking a limit). This distance can be used for both ordinal and quantitative variables.

Formula $d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^{\lambda}}$

2. Clustering Algorithms

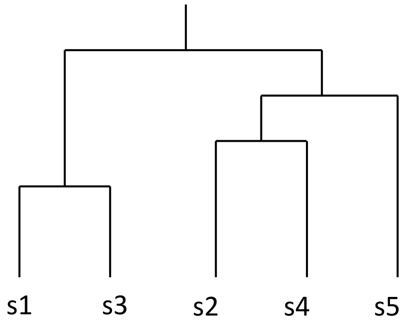
There are many clustering algorithms available in literature, choice of appropriate algorithm depends on the data type and desired results. We will be focusing on commonly used clustering algorithms.

3.1 Hierarchical Algorithms

A hierarchical method creates a hierarchical decomposition of data objects in the form of tree like diagram which is called a dendogram. There are two approaches to building a cluster hierarchy.

Agglomerative approach also called bottom up approach starts with each object forming a separate group and successively merges the objects close to one another, until all the groups are merged into one.

Divisive approach also called top-down approach starts with all the objects in same cluster, until each object is in one cluster.



Process flow of agglomerative hierarchical clustering method is given below:

- Convert object features to distance matrix.

- Set each object as a cluster (thus if we have 6 objects, we will have 6 clusters in the beginning)
- Iterate until number of cluster is 1
    1. Merge two closest clusters
    2. Update distance matrix

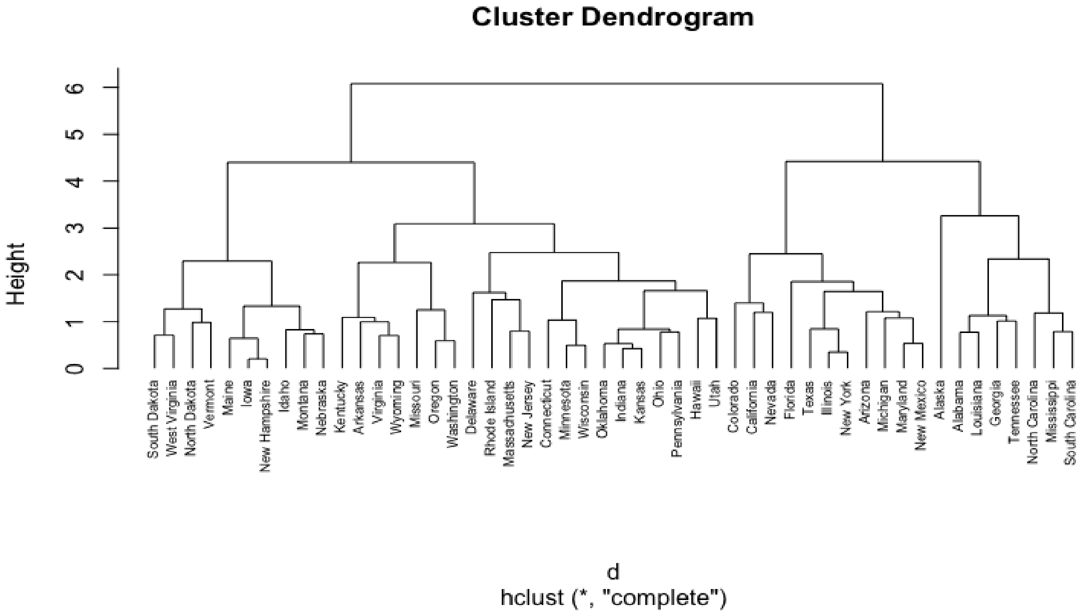
First distance matrix is computed using any valid distance measure between pairs of objects. The choice of which clusters to merge is determined by a linkage criterion, which is a function of the pairwise distances between observations. Commonly used linkage criteria are mentioned below:

- Complete Linkage: The maximum distance between elements of each cluster $\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}.$
- Single Linkage: The minimum distance between elements of each cluster $\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}.$
- Average Linkage /UPGMA: The mean distance between elements of each cluster
$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y).$$

3.1.1 Hierarchical Clustering (HC) using R:

In R, function hclust() performs hierarchical clustering. First the dissimilarity values are computed with dist function. Feed these values into hclust and specify the agglomeration method to be used (i.e. “complete”, “average”, “single”, “ward.D”). Then plot the dendrogram.

```
# Dissimilarity matrix
d <- dist(df, method = "euclidean")
# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )
# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)
```



Alternatively, you can use the agnes function. These functions behave very similarly; however, with the agnes function you can also get the agglomerative coefficient, which measures the amount of clustering structure found (values closer to 1 suggest strong clustering structure).

```
# Compute with agnes
hc2 <- agnes(df, method = "complete")
# Agglomerative coefficient
hc2$ac
## [1] 0.8531583
```

This allows us to find certain hierarchical clustering methods that can identify stronger clustering structures. Here we see that Ward’s method identifies the strongest clustering structure of the four methods assessed.

```
# methods to assess
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")
# function to compute coefficient
ac <- function(x) {
  agnes(df, method = x)$ac
}
map_dbl(m, ac)
##   average    single  complete      ward
## 0.7379371 0.6276128 0.8531583 0.9346210
hc3 <- agnes(df, method = "ward")
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of agnes")
```

Similarly, HC can be performed using function diana. diana works similar to agnes; however, there is no method to provide.

```
# compute divisive hierarchical clustering
hc4 <- diana(df)
# Divise coefficient; amount of clustering structure found
hc4$dc
## [1] 0.8514345
# plot dendrogram
pltree(hc4, cex = 0.6, hang = -1, main = "Dendrogram of diana")
```

Working with Dendrograms

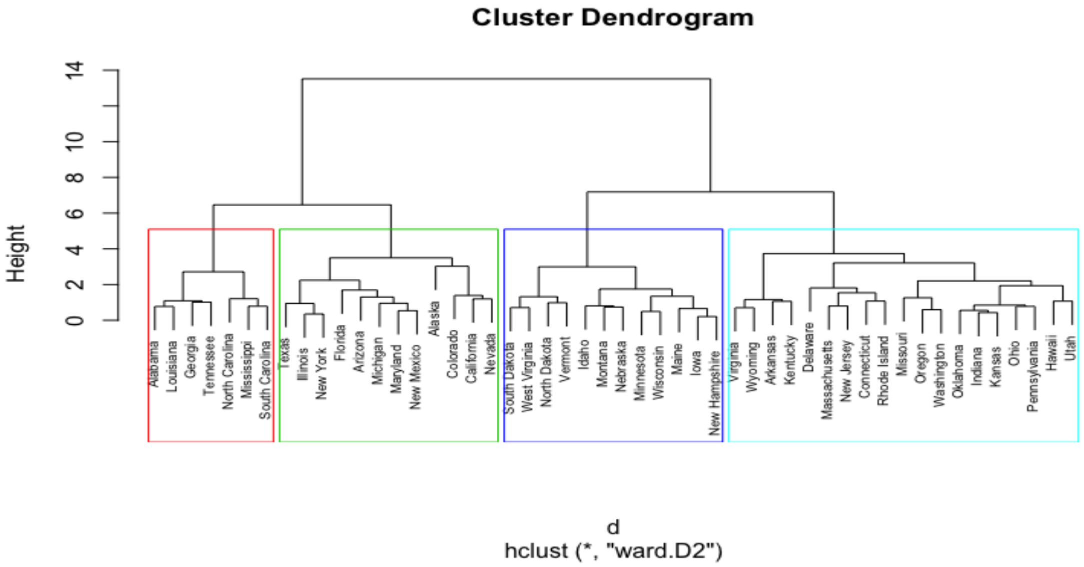
In the dendrogram displayed above, each leaf corresponds to one observation. As we move up the tree, observations that are similar to each other are combined into branches, which are themselves fused at a higher height. The height of the fusion, provided on the vertical axis, indicates the (dis)similarity between two observations.

The height of the cut to the dendrogram controls the number of clusters obtained. we can cut the dendrogram with cutree ():

```
# Ward's method
hc5 <- hclust(d, method = "ward.D2" )
# Cut tree into 4 groups
sub_grp <- cutree(hc5, k = 4)
# Number of members in each cluster
table(sub_grp)
## sub_grp
##  1  2  3  4
##  7 12 19 12
```

It’s also possible to draw the dendrogram with a border around the 4 clusters. The argument border is used to specify the border colors for the rectangles:

```
plot(hc5, cex = 0.6)
rect.hclust(hc5, k = 4, border = 2:5)
```



3.2 Partitional Algorithms

It basically involves segmenting data objects into k partitions, optimizing some criteria, over t iterations. These methods are popularly known as iterative relocation methods.

3.2.1 K-means Algorithm

K-means is the most popularly used algorithm in this category. It randomly selects k objects as cluster mean or center. It works towards optimizing square error criteria function, defined as:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - m_i\|^2 \text{ , where } m_i \text{ is the mean of cluster } C_i .$$

Main steps of k-means algorithm are:

1) Assign initial means m_i

2) Assign each data object x to the cluster C_i for the closest mean

3) Compute new mean for each cluster

4)Iterate until criteria function converges, that is, there are no more new assignments.

The k-means algorithm is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data.

3.2.2 k-means clustering in R :

We can compute k-means in R with the kmeans function. In this example, data is grouped into two clusters (centers = 2). The kmeans function also has an nstart option that attempts multiple initial configurations and reports on the best one. For example, adding nstart = 25 will generate 25 initial configurations.

```
k2 <- kmeans(df, centers = 2, nstart = 25)
str(k2)
## List of 9
##  $ cluster     : Named int [1:50] 1 1 1 2 1 1 2 2 1 1 ...
##   ..- attr(*, "names")= chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ centers     : num [1:2, 1:4] 1.005 -0.67 1.014 -0.676 0.198 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ totss       : num 196
##  $ withinss    : num [1:2] 46.7 56.1
##  $ tot.withinss: num 103
##  $ betweenss   : num 93.1
##  $ size        : int [1:2] 20 30
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

The output of kmeans is a list with several bits of information. The most important being:

cluster: A vector of integers (from 1:k) indicating the cluster to which each point is allocated.

centers: A matrix of cluster centers.

totss: The total sum of squares.

withinss: Vector of within-cluster sum of squares, one component per cluster.

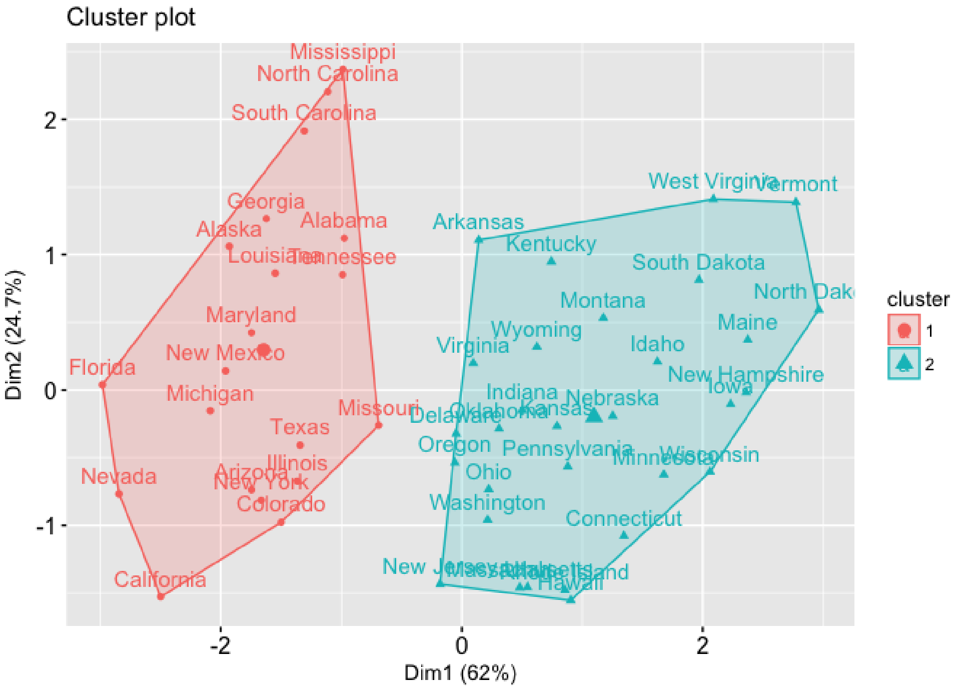
tot.withinss: Total within-cluster sum of squares, i.e. sum(withinss).

betweenss: The between-cluster sum of squares, i.e. $totss-tot.withinss$.

size: The number of points in each cluster.

We can also view the results by using fviz_cluster. This provides a nice illustration of the clusters. If there are more than two dimensions (variables). fviz_cluster will perform principal component analysis (PCA) and plot the data points according to the first two principal components that explain the majority of the variance.

```
fviz_cluster(k2, data = df)
```



References

A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, *ACM Computing Surveys*, 31(3): 264-323, 1999, ISSN: 0360-0300.

B. Mirkin, *Clustering for Data Mining: Data Recovery Approach*, Chapman & Hall/CRC, 2005.

I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*, Morgan Kaufmann publishers, 1999.

J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann Publisher, 2006, ISBN 1-55860-901-6

L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, 1990.

http://en.wikipedia.org/wiki/K-means_clustering

http://people.revoledu.com/kardi/tutorial

R. Xu, D. Wunsch, Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, May 2005

S. Mitra, T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics*, John Wiley & Sons, 2004, ISBN 9812-53-063-0.

https://uc-r.github.io/hc_clustering

https://uc-r.github.io/kmeans_clustering

https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/hclust

ARCH/ GARCH FAMILY OF NON-LINEAR MODELS

Ranjit Kumar Paul
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
ranjitstat@gmail.com, ranjit.paul@icar.gov.in

Introduction

The most widely used technique for analysis of time-series data is; undoubtedly, the Box Jenkins' Autoregressive integrated moving average (ARIMA) methodology (Box *et al.*, 2007). However, it is based on some crucial assumptions, like linearity and homoscedastic prediction error variances. In reality, underlying relationships among variables are highly complex and cannot be described satisfactorily through a linear modelling approach. There are many features, like existence of threshold value, which can be described only through a nonlinear approach. During the last few decades a new area of "Nonlinear time-series modelling" is fast coming up. Here, there are basically two approaches, viz. Parametric or Nonparametric. Evidently, if in a particular situation, we are quite sure about the functional form, we should use the former; otherwise the latter may be employed.

When dealing with nonlinearities, Campbell *et al.* (1997) made the distinction between:

- *Linear Time-Series:* shocks are assumed to be uncorrelated but not necessarily identically and independently distributed (*iid*).
- *Nonlinear Time-Series:* shocks are assumed to be *iid*, but there is a nonlinear function relating the observed time-series $\{X_t\}_{t=0}^{\alpha}$ and the underlying shocks, $\{\varepsilon_t\}_{t=0}^{\alpha}$.

A nonlinear process is described as

$$X_t = g(\varepsilon_{t-1}, \varepsilon_{t-2}, ...) + \varepsilon_t h(\varepsilon_{t-1}, \varepsilon_{t-2}, ...).\ E[X_t / \psi_{t-1}] = g(\varepsilon_{t-1}, \varepsilon_{t-2}, ...)$$

$$Var[X_t / \psi_{t-1}] = E\left[\{(X_t - E(X_t))/\psi_{t-1}\}^2\right] = \{h(\varepsilon_{t-1}, \varepsilon_{t-2}, ...)/\psi_{t-1}\}^2$$

where function $g(\cdot)$ corresponds to conditional mean of X_t, and function $h(\cdot)$ is coefficient of proportionality between innovation in X_t and shock ε_t. The general form above leads to a natural division in Nonlinear time-series literature in two branches:

- Models Nonlinear in Mean: $g(\cdot)$ is nonlinear;
- Models Nonlinear in Variance: $h(\cdot)$ is nonlinear.

The most promising parametric nonlinear time series models like ARCH and GARCH models are described below.

Autoregressive Conditional Heteroscedastic (ARCH) Model

The most promising parametric nonlinear time-series model has been the Autoregressive conditional heteroscedastic (ARCH) model, which was introduced by Engle (1982). It allows the conditional variance to change over time as a function of squared past errors leaving the unconditional variance constant. The presence of ARCH type effects in financial and macro-economic time-series is a well established fact. The combination of ARCH specification for conditional variance and the Autoregressive (AR) specification for conditional mean has many appealing features, including a better specification of the forecast error variance. Ghosh and Prajneshu (2003) employed AR(*p*)-ARCH(*q*)-in-Mean model for carrying out modelling and forecasting of volatile monthly onion price data. The AR-ARCH model has also been used as the basic "building blocks" for Markov switching and mixture models (See e.g. Lanne and Saikkonen 2003 and Wong and Li 2001).

The ARCH (*q*) model for series $\{\varepsilon_t\}$ is defined by specifying the conditional distribution of ε_t given information available up to time $t-1$. Let ψ_{t-1} denote this information. It consists of the knowledge of all available values of the series, and anything which can be computed from these values, *e.g.* innovations, and squared observations. In principle, it may even include knowledge of the values of other related time-series, and anything else which might be useful for forecasting and is available by time $t-1$.

We say that the process $\{\varepsilon_t\}$ is ARCH (*q*), if the conditional distribution of $\{\varepsilon_t\}$ given available information ψ_{t-1} is

$$\varepsilon_t \mid \psi_{t-1} \sim N(0, h_t) \text{ and } h_t = a_0 + \sum_{i=1}^{q} a_i \, \varepsilon_{t-i}^2 \qquad (1)$$

where $a_0 > 0$, $a_i \geq 0$ for all *i* and $\sum_{i=1}^{q} a_i < 1$

Properties of the ARCH model (Tsay,2005)

To study the properties of ARCH model, consider the simple ARCH (*1*) model. The conditional variance equation of the this model is defined as

$$\varepsilon_t = \eta_t h_t^{1/2},$$

η_t is white noise and conditional variance h_t satisfies

$$h_t = a_0 + a_1\varepsilon_{t-1}^2$$

where $a_0 > 0$, $a_1 \geq 0$.The important properties of ARCH models are mentioned below:

(i) The unconditional mean of ε_t remains zero because,

$$E(\varepsilon_t) = E[E(\varepsilon_t|\Psi_{t-1})] = E\left[\sqrt{h_t}E(\varepsilon_t)\right] = 0$$

(ii) The unconditional variance of ε_t can be defined as

$$\text{var}(\varepsilon_t) = E(\varepsilon_t^2) = E[E(\varepsilon_t^2 \mid \psi_{t-1})] = E(a_0 + a_1\varepsilon_{t-1}^2) = a_0 + a_1E(\varepsilon_{t-1}^2).$$

If ε_t is a stationary process with $E(\varepsilon_t) = 0$, $\text{var}(\varepsilon_t) = \text{var}(\varepsilon_{t-1}) = E(\varepsilon_{t-1}^2)$. Therefore, $\text{var}(\varepsilon_t) = a_0 + a_1 \text{var}(\varepsilon_t)$ and so $var(\varepsilon_t) = a_0 / (1 - a_1)$. Since variance of ε_t must be positive, therefore $0 \leq a_1 < 1$.

(iii) In some applications, higher order moments of ε_t are required to exist and, hence, a_1 must satisfy some additional constraints. For instance, to study its tail behavior, we require that the fourth moment of ε_t is finite.

Heavy tails are a common aspect of financial data, and hence the ARCH models are very popular in this field. Besides that, Bera and Higgins (1993) mention the following reasons for the ARCH success:

- ARCH models are simple and easy to handle.
- ARCH models take care of clustered errors.
- ARCH models take care of nonlinearities.
- ARCH models take care of changes in the econometrician’s ability to forecast.

Forecasting

Forecasts of the ARCH model can be obtained recursively as those of an AR model. Consider an ARCH (q) model. At the forecast origin t, the one-step ahead forecast is

$$h_t(1) = a_0 + a_1\varepsilon_t^2 + \ldots + a_q\varepsilon_{t+1-q}^2 \quad (2)$$

The two-step ahead forecast is $h_t(2) = a_0 + a_1h_t(1) + a_2\varepsilon_t^2 + \ldots + a_q\varepsilon_{t+2-q}^2$, and l- step ahead forecast is $h_t(l) = a_0 + \sum_{i=1}^{q} a_ih_t(l-i)$ where $h_t(l-i) = \varepsilon_{t+l-i}^2$ if $l - i \leq 0$.

However, ARCH model has some drawbacks. Firstly, when the order of ARCH model is very large, estimation of a large number of parameters is required. Secondly, conditional variance of ARCH(q) model has the property that unconditional autocorrelation function (Acf) of squared residuals; if it exists, decays very rapidly compared to what is typically observed, unless maximum lag q is large. To overcome these difficulties, Bollerslev (1986) proposed the Generalized ARCH (GARCH) model in which conditional variance is also a linear function of its own lags. This model is also a weighted average of past squared residuals, but it has declining weights that never go completely to zero. It gives parsimonious models that are easy to estimate and, even in its simplest form, has proven surprisingly successful in predicting conditional variances. Angelidis *et al*. (2004) evaluated the performance of GARCH models in modelling the daily Value-at-Risk (VaR) of perfectly distributed portfolios in five stock indices, using a number of distributional assumptions and sample sizes. Paul *et al*. (2009, 2014) applied GARCH model for forecasting of spices export and wheat yield respectively.

Generalized ARCH(GARCH) Model

To overcome the weaknesses of ARCH model, Bollerslev (1986) and Taylor (1986) proposed the Generalized ARCH (GARCH) model independently of each other, in which conditional variance is also a linear function of its own lags and has the following form

$$\varepsilon_t = \xi_t h_t^{1/2} \; h_t = a_0 + \sum_{i=1}^{q} a_i \, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j \, h_{t-j} \qquad (3)$$

where ξ_t ~ IID($0,1$). A sufficient condition for the conditional variance to be positive is

$$a_0 > 0, \; a_i \geq 0, \; i = 1,2,...,q. \;\; b_j \geq 0, \;\; j = 1,2,...,p$$

The GARCH (p, q) process is weakly stationary if and only if $\sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1$.

The conditional variance defined by (3) has the property that the unconditional autocorrelation function of ε_t^2 ; if it exists, can decay slowly. For the ARCH family, the decay rate is too rapid compared to what is typically observed in financial time-series, unless the maximum lag q is long. As (3) is a more parsimonious model of the conditional variance than a high-order ARCH model, most users prefer it to the simpler ARCH alternative.

The most popular GARCH model in applications is the GARCH(*1,1*) model. To express GARCH model in terms of ARMA model, denote $\eta_t = \varepsilon_t^2 - h_t$. Then from equation (3)

$$\varepsilon_t^2 = a_0 + \sum_{i=1}^{Max(p,q)} (a_i + b_i)\varepsilon_{t-i}^2 + \eta_t + \sum_{j=1}^{p} b_j \eta_{t-j} \qquad (4)$$

Thus a GARCH model can be regarded as an extension of the ARMA approach to squared series { ε_t^2 }. Using the unconditional mean of an ARMA model, we have

$$\mathrm{E}(\varepsilon_t^2) = \frac{a_0}{1 - \sum_{i=1}^{Max(p,q)} (a_i + b_i)} \qquad (5)$$

provided that the denominator of the prior fraction is positive.

Properties of GARCH model

The most widely used GARCH specification asserts that the best predictor of the variance in the next period is a weighted average of the long-run average variance, the variance predicted for this period, and the new information in this period that is captured by the most recent squared residual. Such an updating rule is a simple description of adaptive or learning behavior and can be thought of as Bayesian updating.

The properties of GARCH models can easily be studied by focusing on the simplest GARCH(*1,1*) model with

$$\varepsilon_t = \xi_t h_t^{1/2} \quad h_t = a_0 + a_1 \varepsilon_{t-1}^2 + b_1 h_{t-1}, \qquad (6)$$

where ξ_t ~ IID(*0,1*) and $0 \le a_1, b_1 \le 1, (a_1 + b_1) < 1$.

The GARCH model that has been described is typically called the GARCH(*1,1*) model. The (*1,1*) in parentheses is a standard notation in which the first number refers to how many autoregressive lags, or ARCH terms, appear in the equation, while the second number refers to how many moving average lags are specified, which here is often called the number of GARCH terms. Sometimes models with more than one lag are needed to find good variance forecasts.

First a large ε_{t-1}^2 or h_{t-1} gives rise to a large h_t. This means that a large ε_{t-1}^2 tends to followed by another large ε_t^2, generating again the well known behavior of volatility clustering in financial time-series.

Second it can be seen that if $1-2a_1^2-(a_1+b_1)^2>0$, then

$$\frac{E(\varepsilon_t^4)}{[E(\varepsilon_t)]^2}=\frac{3\left[1-(a_1+b_1)^2\right]}{1-(a_1+b_1)^2-2a_1^2}>3$$

Consequently, similar to ARCH models, the tail distribution of a GARCH(*1,1*) process is heavier than that of a normal distribution.

Third, the model provides a simple parametric function that can be used to describe the volatility evolution.

Forecasting volatility by GARCH model

Forecasts of a GARCH model can be obtained using methods similar to those of an ARMA model. Although this model is directly set up to forecast for just one period, it turns out that based on the one-period forecast, a two-period forecast can be made. Ultimately, by repeating this step, long-horizon forecasts can be constructed. For the GARCH(*1,1*), the two-step forecast is a little closer to the long-run average variance than is the one-step forecast, and, ultimately, the distant-horizon forecast is the same for all time periods as long as $(a_1+b_1)<1$. This is just the unconditional variance. Thus, the GARCH models are mean reverting and conditionally heteroscedastic, but have a constant unconditional variance.

Consider the GARCH(*1,1*) model in (6) and assume that the forecast origin is *t*, the one-step ahead forecast is $h_t(1)=a_0+a_1\varepsilon_t^2+b_1h_t$

For multi-step ahead forecasts, use $\varepsilon_t^2=\xi_t^2h_t$ and rewrite the volatility equation in (6) as

$$h_{t+1}=a_0+(a_1+b_1)h_t+a_1h_t(\varepsilon_t^2-1)$$

For two-step ahead forecasts $h_{t+2}=a_0+(a_1+b_1)h_{t+1}+a_1h_{t+1}(\varepsilon_{t+1}^2-1)$ Since

$$E((\varepsilon_{t+1}^2-1)/\psi_t)=0,$$

The two-step ahead volatility forecast at the forecast origin *t* satisfies the equation

$$h_t(2)=a_0+(a_1+b_1)h_t(1)$$

In general we have $h_t(l)=a_0+(a_1+b_1)h_t(l-1)$, *l>1*

This result is exactly the same as that of an ARMA(*1,1*) model. By repeated substitution in the equation (7), the one- step ahead forecast can be written as

$$h_t(l)=\frac{a_0\left[1-(a_1+b_1)^{l-1}\right]}{1-a_1-b_1}+(a_1+b_1)^{l-1}h_t(1)$$

Therefore, $h_t(l) \rightarrow \dfrac{a_0}{1-a_1-b_1}, as\, l \rightarrow \infty$,provided that a_1+b_1<1.

Consequently, the multi-step ahead volatility forecast of a GARCH(*1,1*) model converge to the unconditional variance of ε_t as the forecast horizon increases to infinity provided that Var(ε_t)exists.

In order to estimate the parameters of GARCH model, three types of estimator are available in literature. They are the conditional maximum likelihood estimator, Whitle's estimator and the least absolute deviation estimator.

Conditional maximum likelihood estimator

Similar to the estimation for ARMA models, the most frequently used estimators for ARCH/GARCH models are those derived from a (conditional) Gaussian likelihood function.

The loglikelihood function of a sample of *T* observations, apart from constant, is

$$L_T(\theta) = T^{-1}\sum_{t=1}^{T}\left(log\, h_t + \varepsilon_t^2 h_t^{-1}\right), \text{ where } h_t = a_0 + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j\, h_{t-j}$$

For a general GARCH model the conditional variance (h_t) cannot be expressed in terms of a finite number of the past observations. Some truncation is inevitable. By induction, it is possible to derive

$$h_t = \frac{a_0}{1-\sum_{i=1}^{q} a_i} + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 + \sum_{i=1}^{q} a_i \sum_{k=1}^{\infty}\sum_{j_1=1}^{p} \dots \sum_{j_k=1}^{p} b_{j_1}..b_{j_k}\, \varepsilon_{t-i-j_1-\dots-j_k}^2$$

where the multiple sum vanishes if *q = 0*. It is to be noted that the multiple sum above converges with probability *1* since each a_i and b_i is nonnegative, and since the expected value of the multiple series is finite. In practice the above expression of h_t is replaced by truncation version

$$\tilde{h}_t = \frac{a_0}{1-\sum_{i=1}^{q} a_i} + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 + \sum_{i=1}^{q} a_i \sum_{k=1}^{\infty}\sum_{j_1=1}^{p} \dots \sum_{j_k=1}^{p} b_{j_1}..b_{j_k}\, \varepsilon_{t-i-j_1-\dots-j_k}^2\, I\left(t-i-j_1-\dots-j_k \ge 1\right)$$

where*t*>*q*.

In general, suppose that *f*(.) is the probability density function of ε_t. However, generally, maximum likelihood estimators are derived by minimizing

$$L_T(\theta)=T^{-1}\overset{T}{\underset{t=v}{\Sigma}}\left(log\sqrt{\tilde{h}_t}-log\ f\left(\frac{\varepsilon_t}{\sqrt{\tilde{h}_t}}\right)\right)$$

where $\tilde{h}_t$ is the truncated version of h_t (Fan and Yao, 2003).

Whitle's estimator

For GARCH(p,q) defined by (3), the conditional variance can be written a

$$h_t=\frac{a_0}{1-\sum_{i=1}^{q}a_i}+\sum_{i=1}^{\infty}d_i\,\varepsilon_{t-i}^2 \text{ where} d_i \geq 0 \text{and } \sum_{i=1}^{\infty}d_i=\frac{\sum_{j=1}^{p}b_j}{1-\sum_{i=1}^{q}a_i}$$

Suppose that $\{\varepsilon_t\}$ is fourth-order stationary in the sense that its first four moments are all time-invariant. $x_t=\varepsilon_t^2$ then $\{x_t\}$ is a stationary AR(∞) process satisfying

$$x_t=\frac{a_0}{1-\sum_{i=1}^{q}a_i}+\sum_{i=1}^{\infty}d_i\,x_{t-i}+e_t \text{ where} e_t \text{ is a martingale difference}$$

$$e_t=\left(\varepsilon_t^2-1\right)\left\{\frac{a_0}{\left(1-\sum_{i=1}^{q}a_i\right)}+\sum_{i=1}^{\infty}d_i\,x_{t-i}\right\}$$ with σ_e^2 = Var(e_t) <∞. Therefore, the spectral density of the process $\{x_t\}$ is $g(\omega)=\frac{\sigma_e^2}{2\pi}\left|1-\sum_{i=1}^{\infty}d_i\,e^{ij\omega}\right|^{-2}$

Whitle's estimators for a_i and b_i are obtained by minimizing $\sum_{j=1}^{T-1}I_T(\omega_j)/\,g(\omega_j)$

whereI_T(.) is the periodogram of $\{x_t\}$ and $\omega_j=2\pi j/2$.

Whittle's estimator suffer from the lack of efficiency, as e_t is unlikely to be normal even when η_t is normal.

Least absolute deviations estimator

Both the estimators discussed above are derived from maximizing a Gaussian likelihood or an approximate Gaussian likelihood. In time-series they are known as L_2 - estimators. Empirical evidence suggests that some financial time-series exhibit heavy-tailed than those of a normal distribution would be more appropriate. Based on this consideration, Peng and Yao (2003) proposed Least absolute deviations estimation (LADE) which minimizes

$\sum_{t=v}^{T}\left|log\varepsilon_t^2 - log(h_t)\right|$wherev $= p+1$, if $q = 0$ and $v > p+1$, if $q > 0$.

The idea behind this implies implicitly a reparameterization of model (3) such that E(ξ_t) = *0* and the median (instead of variance) of η_t^2 is equal to *1*. Peng and Yao (2002) showed that under very mild conditions, the least absolute deviations estimators are asymptotically normal with the standard convergence rate $T^{1/2}$ regardless of whether the distribution of η_t has heavy tails or not. This is in marked contrast to the conditional maximum likelihood estimators, which will suffer from slow convergence when ξ_t is heavy-tailed.

Fan and Yao (2003) and Straumann (2005) have given a good description of various estimation procedures for conditionally heteroscedastic time- series models.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for GARCH model with Gaussian distributed errors are computed by:

$$AIC = \sum_{t=1}^{T}\left(log\tilde{h}_t + \varepsilon_t^2\tilde{h}_t^{-1}\right) + 2(p + q + 1) \qquad (7)$$

and

$$BIC = \sum_{t=1}^{T}\left(log\tilde{h}_t + \varepsilon_t^2\tilde{h}_t^{-1}\right) + 2(p + q + 1)\ log(T - v + 1) \qquad (8)$$

where*T* is the total number of observations.

Evidently, the likelihood equations are extremely complicated. Fortunately, the estimates can be obtained by using a software package, like EViews, SAS, SPLUS GARCH, GAUSS, TSP, R, MATLAB, and RATS.

Testing for ARCH Effects

Let $\varepsilon_t = y_t - \phi y_{t-1}$ be the residual series. The squared series $\{\varepsilon_t^2\}$ is then used to check for conditional heteroscedasticity, which is also known as the ARCH effects. To this end, two tests, briefly discussed below, are available. The first one is to apply the usual Ljung-Box statistic *Q*(*m*) to the $\{\varepsilon_t^2\}$series. The null hypothesis is that the first *m* lags of autocorrelation functions of the $\{\varepsilon_t^2\}$ series are zero. The second test for conditional heteroscedasticity is the Lagrange multiplier test of Engle (1982). This test is equivalent to usual *F*-statistic for testing $H_0 : a_i = 0,\ i = 1, 2, \ldots, q$ in the linear regression

$$\varepsilon_t^2 = a_0 + a_1\varepsilon_{t-1}^2 + \ldots + a_q\varepsilon_{t-q}^2 + e_t,\ t = q+1, \ldots, T \qquad (9)$$

where e_t denotes the error term, q is the prespecified positive integer, and T is the sample size.

Let $SSR_0 = \sum_{t=q+1}^{T}\left(\varepsilon_t^2 - \varpi\right)^2$, where $\varpi = \sum_{t=q+1}^{T} \varepsilon_t^2 / T$ is the sample mean of $\left\{\varepsilon_t^2\right\}$, and $SSR_1 = \sum_{t=q+1}^{T} \hat{e}_t^2$, where $\hat{e}_t$ is the least squares residual of (9). Then, under H_0,

$$F = \frac{\left(SSR_0 - SSR_1\right)/q}{SSR_1\left(T - q - 1\right)} \tag{10}$$

is asymptotically distributed as chi-squared distribution with q degrees of freedom. The decision rule is to reject H_0 if $F > \chi_q^2(\alpha)$, where $\chi_q^2(\alpha)$ is the upper $100(1-\alpha)$th percentile of χ_q^2 or, alternatively, the p-value of F is less than α.

Illustration(Paul *et al*., 2009)

Paul *et al*. (2009) found that AR(*1*)-GARCH (*1*,*1*) model was better than ARIMA model for modeling and forecasting of all-India data of monthly export of spices during the period April, 2000 to November, 2006.First of all they fitted ARIMA model. The appropriate model was chosen on the. ARIMA(*1*,*1*,*1*) model is selected for modelling and forecasting of the export of spices based on minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. The estimates of parameters of above model are reported in Table 1. The graph of fitted model along with data points is exhibited in Fig. 1. Evidently, the fitted ARIMA(*1*,*1*,*1)* model is not able to capture successfully the volatility present at various time-epochs, like October, 2001; May, 2002; March, 2004; and March, 2006.

Table 1. Estimates of parameters along with their standard errors for fitted ARIMA(*1*,*1*,*1)* model

Parameter	Estimate	Standard error
AR1	-0.100	0.159
MA1	0.696	0.119
Constant	1.468	0.966

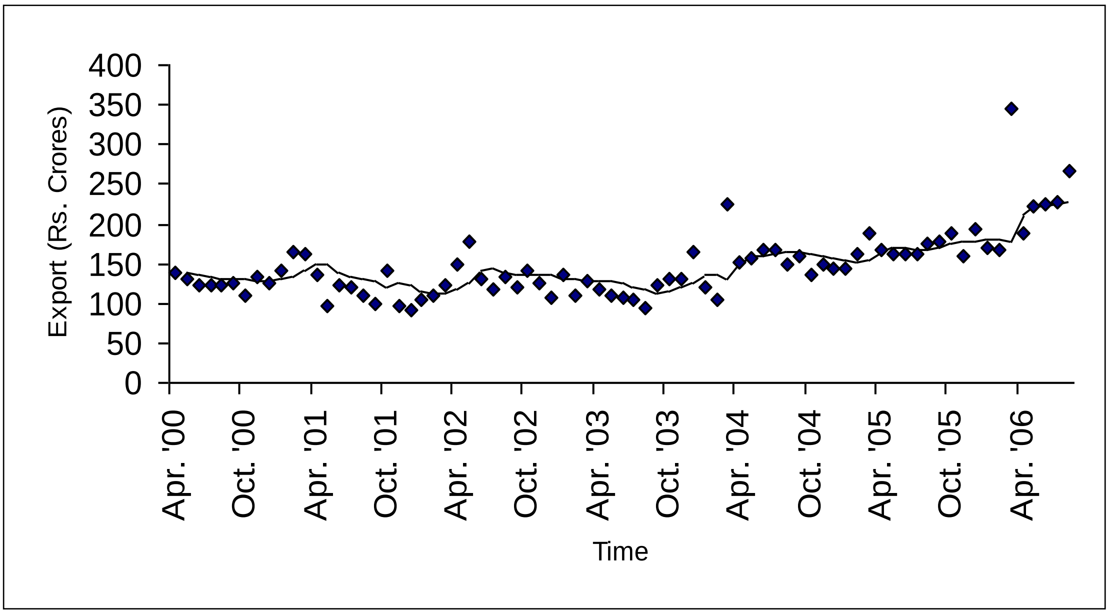


Fig. 1. Fitted ARIMA(*1*,*1*,*1*) model along with data points

Fitting of GARCH Model

On investigating autocorrelation of the squared residuals of the fitted ARIMA(*1*,*1*,*1*) model it was found that the autocorrelation was highest at lag 24, which was 0.265. The ARCH-LM test statistic at lag 24 computed using equation (10) was 37.48, which was significant at 5% level of significance. But it is not reasonable to apply ARCH model of order 24 in view of the enormously large number of parameters. Therefore, the parsimonious GARCH model is applied. The AR(*1*)-GARCH(*1*,*1*) model is selected on the basis of minimum AIC and BIC values. The estimates of parameters of the above model along with their corresponding standard errors in brackets ( ) are

$y_t = 157.99 + 0.829 y_{t-1} + \varepsilon_t$

(33.692) (0.087)

where $\varepsilon_t = h_t^{1/2} \xi_t$, and h_t satisfies the variance equation

$h_t = 1427.855 + 0.354 \varepsilon_{t-1}^2 + 0.509 h_{t-1}$

(237.058) (0.277) (0.206)

Using eqs. (7) and (8), the AIC and BIC values for fitted AR(*1*) – GARCH(*1*,*1*) model, are respectively computed as 479.77 and 521.97. To study the appropriateness of the fitted GARCH model, the autocorrelation function of the standardized residuals and squared standardized residuals are computed and it is found that, in both situations, the autocorrelation function is insignificant at 5% level of significance, thereby confirming that the mean and variance equations are correctly specified.The graph of fitted model along with data points is exhibited in Fig. 2. Obviously, the fitted GARCH model is able to capture the volatility present in the data set.EViews software package was employed for fitting of these models.

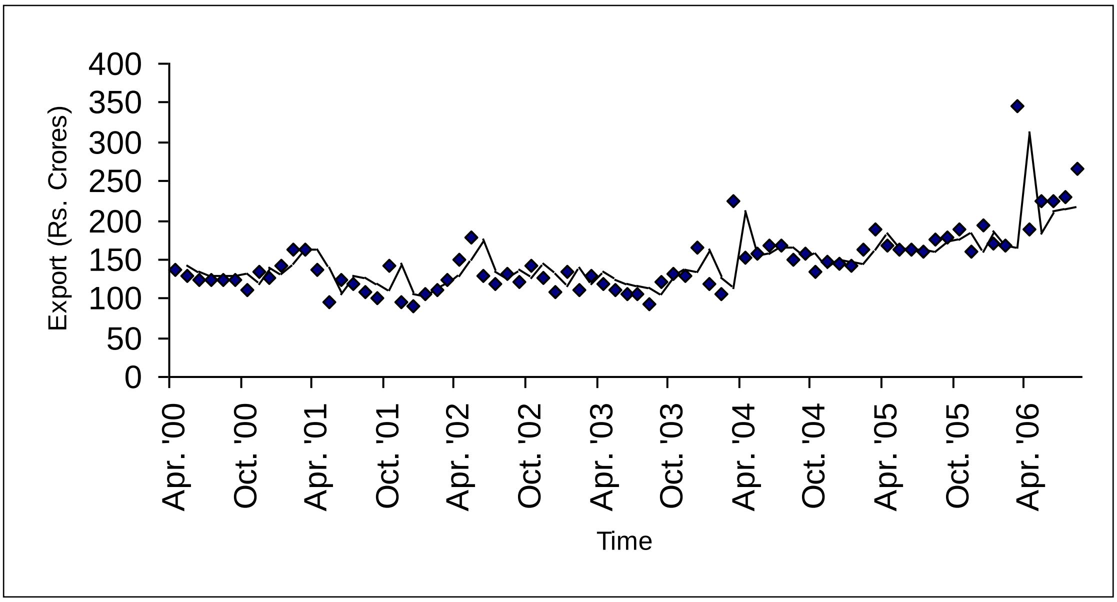


Fig. 4. Fitted AR(*1*) – GARCH(*1*,*1*) model along with data points

Forecasting

One-step ahead forecasts of export of spices along with their corresponding standard errors inside the brackets ( ) for the months of September, 2006 to November, 2006 in respect of above fitted models are reported in Table 2. A perusal indicates that, for fitted GARCH model, all the forecast values lie within one standard error of forecasts. However, this attractive feature does not hold for fitted ARIMA model.

The Mean square prediction error (MSPE) values and Mean absolute prediction error (MAPE) values for fitted GARCH model are respectively computed as 18.14 and 15.00, which are found to be lower than the corresponding ones for fitted ARIMA model, viz. 33.17 and 29.02 respectively.

Table 2. One-step ahead forecasts of export of spices (Rs. Crores) for fitted models

Months	**Actual Price**	**Forecasts by**	
		ARIMA(*1*,*1*,*1*)	**AR(*1*)-GARCH(*1*,*1*)**
Sep. '06	270.91	235.67(29.58)	247.14 (40.93)
Oct. '06	232.59	240.27 (30.12)	231.89 (48.17)
Nov. '06	286.21	241.50 (31.16)	265.68 (53.31)

To sum up, it may be concluded that the AR(*1*)-GARCH(*1,1*) model has performed better than the ARIMA(*1,1,1*) model for present data for both modelling as well as forecasting purposes.

References

Angelidis, T., Benos, A. and Degiannakis, S. (2004). The use of GARCH models in VaR estimation. *Stat. Meth.*, **1**, 105-128.

Bera, A. K., and Higgins, M. L. (1993), "ARCH Models: Properties, Estimation and Testing," *J. Econ.Surv*.,**7**, 307-366.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econ*., **31** 307-327.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2007). *Time-Series Analysis: Forecasting and Control*. 3rd edition. Pearson education, India.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*, Princeton, New Jersey: Princeton University Press.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987-1008.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, U.S.A.

Ghosh, H. and Prajneshu. (2003).Nonlinear time-series modelling of volatile onion price data using AR(p)-ARCH(q)-in-mean. *Cal. Stat. Assn. Bull*., **54**, 231 – 47

Lanne, M. and Saikkonen, P. (2003). Modeling the U.S. short-term interest rate by mixture autoregressive processes. *J. Fin. Econ*., **1**, 96 – 125.

Paul, R. K., Ghosh, H., and Prajneshu (2009). GARCH Nonlinear Time Series Analysis for Modelling and Forecasting of India s Volatile Spices Export Data.J. Ind. Soc. Agri. Stat.**62** (2) 123-132

Paul, R. K., Ghosh, H. and Prajneshu (2014). Development of out-of-sample forecast formulae for ARIMAX-GARCH model and their application. *Journal of the Indian Society of Agricultural Statistics*, **68**(1),85-92.

Peng, L. and Yao, Q. (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika*, **90**, 967-975.

Straumann, D. (2005). *Estimation in conditionally heteroscedastic time series models*. Springer, Germany.

Taylor, S. J. (1986). *Modeling financial time series*. Wiley, New York.

Tsay, R. S. (2005). *Analysis of financial time series*. 2nd Ed. John Wiley, U.S.A.

Wong, C. S. and Li, W. K. (2001). On a mixture autoregressive conditional heteroscedastic model. *J. Amer. Stat. Assoc*., **96**, 992-995.

ENSEMBLE METHODS

Shashi Dahiya
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
shashi.dahiya@icar.gov.in

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. They combine multiple algorithms to produce better classification performance.It is a machine learning approach to combine multiple other models in the prediction process. The combined models increase the accuracy of the results significantly.Those models are referred to as base estimators. It is a solution to overcome the following technical challenges of building a single estimator:High variance: The model is very sensitive to the provided inputs to the learned features.

- Low accuracy: One model or one algorithm to fit the entire training data might not be good enough to meet expectations.
- Features noise and bias: The model relies heavily on one or a few features while making a prediction.

Bagging is used to reduce the variance of weak learners. Boosting is used to reduce the bias of weak learners. Stacking is used to improve the overall accuracy of strong learners.

Ensemble Algorithm

A single algorithm may not make the perfect prediction for a given dataset. Machine learning algorithms have their limitations and producing a model with high accuracy is challenging. If we build and **combine** multiple models, the overall accuracy could get boosted. The combination can be implemented by aggregating the output from each model with two objectives: reducing the model error and maintaining its generalization. The way to implement such aggregation can be achieved using some techniques. Some textbooks refer to such architecture as *meta-algorithms*.

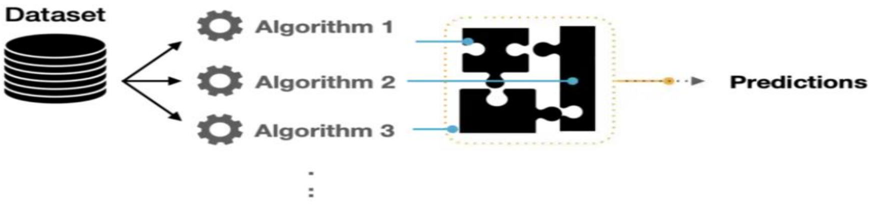


Figure 1: Diversifying the model predictions using multiple algorithms.

Ensemble Learning

Building ensemble models is not only focused on the variance of the algorithm used. For instance, we could build multiple C45 models where each model is learning a specific pattern specialized in predicting one aspect. Those models are called **weak learners** that can be used to obtain a meta-model. In this architecture of ensemble learners, the inputs are passed to each weak learner while collecting their predictions. The combined prediction can be used to build a final ensemble model.

One important aspect to mention is those weak learners can have different ways of mapping the features with variant decision boundaries.

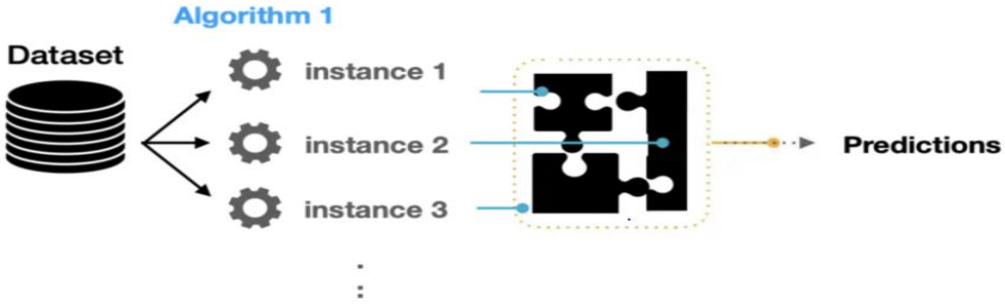


Figure 2: Aggregated predictions using multiple weak learners of the same algorithm.

Ensemble Techniques

Bagging

We use bagging for combining weak learners of high variance. Bagging aims to produce a model with lower variance than the individual weak models. These weak learners are homogenous, meaning they are of the same type.

Bagging is also known as Bootstrap aggregating. It consists of two steps: bootstrapping and aggregation.

Bootstrapping

Involves resampling subsets of data with replacement from an initial dataset. In other words, subsets of data are taken from the initial dataset. These subsets of data are called bootstrapped datasets or, simply, bootstraps. Resampled 'with replacement' means an individual data point can be sampled multiple times. Each bootstrap dataset is used to train a weak learner.

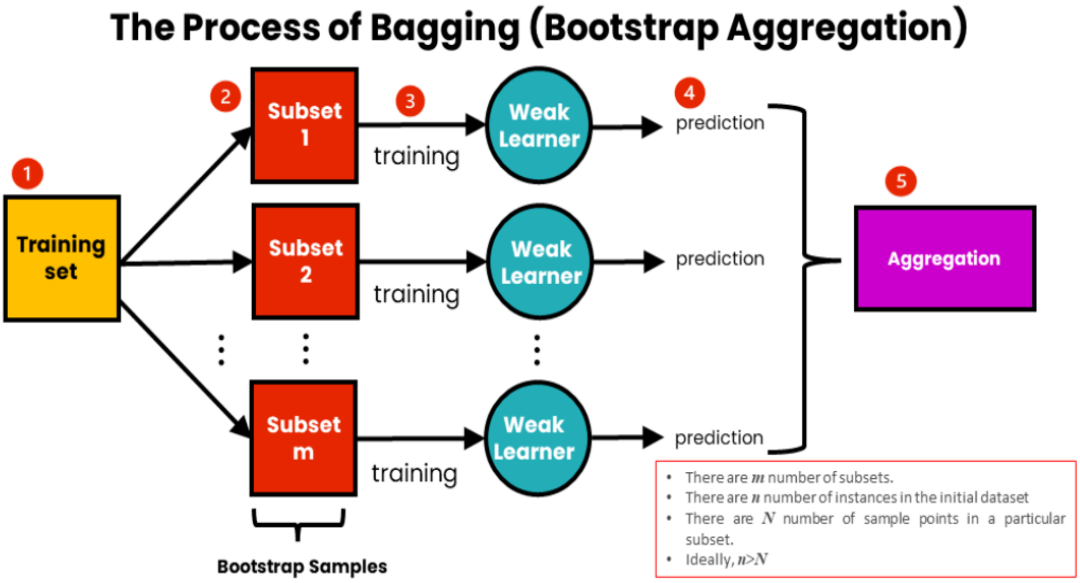
Aggregating

The individual weak learners are trained independently from each other. Each learner makes independent predictions. The results of those predictions are aggregated at the end to get the overall prediction. The predictions are aggregated using either max voting or averaging.

Max Voting is commonly used for classification problems. It consists of taking the mode of the predictions (the most occurring prediction). It is called voting because like in election voting, the premise is that 'the majority rules'. Each model makes a prediction. A prediction from each model counts as a single 'vote'. The most occurring 'vote' is chosen as the representative for the combined model.

Averaging is generally used for regression problems. It involves taking the average of the predictions. The resulting average is used as the overall prediction for the combined model.

It is one of the most straightforward and most intuitive ensemble-based algorithms that create separate samples of the training dataset. Each training dataset is used to train a different classification.



- There are m number of subsets.
- There are n number of instances in the initial dataset
- There are N number of sample points in a particular subset.
- Ideally, $n>N$

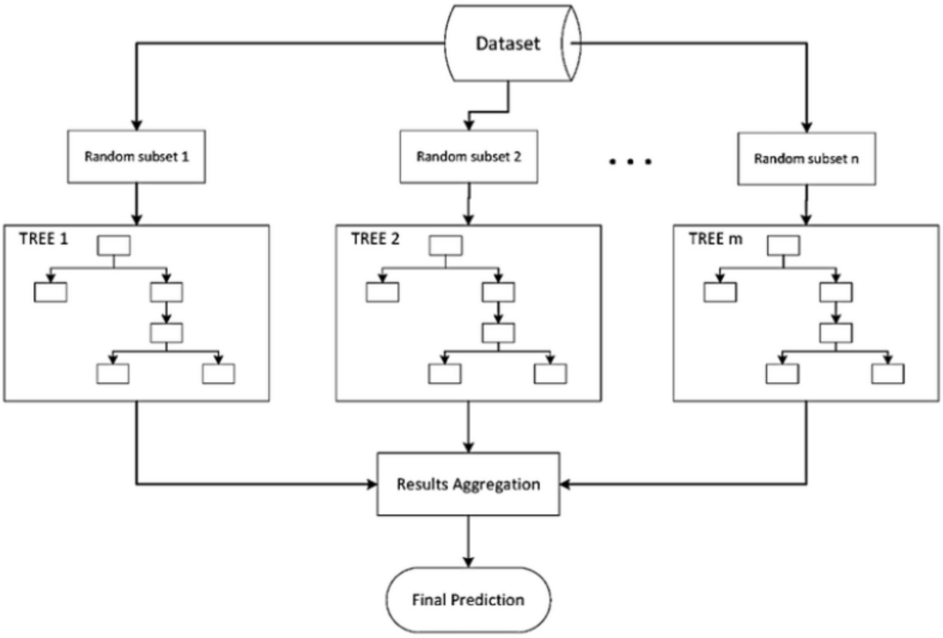
➢ **Bagging**

The idea of bagging is based on making the training data available to an iterative process of learning. Each model learns the error produced by the previous model using a slightly different subset of the training dataset. Bagging reduces variance and minimizes overfitting. One example of such a technique is the Random Forest algorithm.

The steps of Bagging are as follows:

1. We have an initial training dataset containing n-number of instances.

2. We create a m-number of subsets of data from the training set. We take a subset of N sample points from the initial dataset for each subset. Each subset is taken with replacement. This means that a specific data point can be sampled more than once.
3. For each subset of data, we train the corresponding weak learners independently. These models are homogeneous, meaning that they are of the same type.
4. Each model makes a prediction.
5. The predictions are aggregated into a single prediction. For this, either max voting or averaging is used.



Given a Dataset, bootstrapped subsamples are pulled. A Decision Tree is formed on each bootstrapped sample. The results of each tree are aggregated to yield the strongest, most accurate predictor.

Bagging Algorithm:

Input:

Data Set D = {(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)}

Number of iteration T

Process:

Step 1: for i = 1 to T

(a) Through sampling data points with replacement, create a dataset sample S_m.

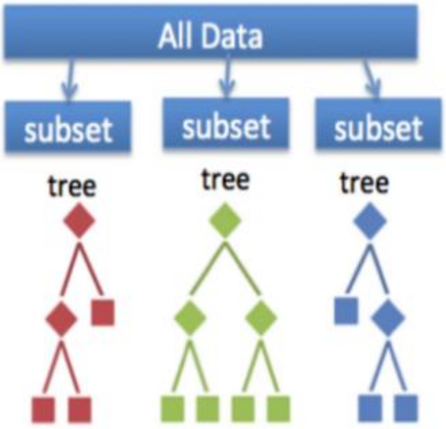
(b) From each dataset sample, S_m learns a classifier C_m.

Step 2: for every test example.

(a) Try all classifiers C_m.

(b) Estimate the class that earns the largest number of votes.

- **Random Forest:** Random Forest is another ensemble machine learning algorithm that follows the bagging technique. It is an extension of the bagging estimator algorithm. The base estimators in random forest are decision trees. Unlike bagging meta estimator, random forest randomly selects a set of features which are used to decide the best split at each node of the decision tree. It uses subset of training samples as well as subset of features to build multiple split trees. Multiple decision trees are built to fit each training set. The distribution of samples/features is typically implemented in a random mode.



A random forest takes a random subset of features from the data, and creates n random trees from each subset. Trees are aggregated together at end.

Looking at it step-by-step, this is what a random forest model does:

1. Random subsets are created from the original dataset (bootstrapping).
2. At each node in the decision tree, only a random set of features are considered to decide the best split.
3. A decision tree model is fitted on each of the subsets.
4. The final prediction is calculated by averaging the predictions from all decision trees.

Note: The decision trees in random forest can be built on a subset of data and features. Particularly, the sklearn model of random forest uses all features for *decision tree and a subset of features are randomly selected for splitting at each node.*

To sum up, Random forest **r**andomly selects data points and features, and builds multiple trees (Forest)**.**

- **Extra-Trees Ensemble:** is another ensemble technique where the predictions are combined from many decision trees. Similar to Random Forest, it combines a large number of decision trees. However, the Extra-trees use the whole sample while choosing the splits randomly.
- **Boosting:**

We use boosting for combining weak learners with high bias. Boosting aims to produce a model with a lower bias than that of the individual models. Like in bagging, the weak learners are homogeneous.

Boosting involves sequentially training weak learners. Here, each subsequent learner improves the errors of previous learners in the sequence. A sample of data is first taken from the initial dataset. This sample is used to train the first model, and the model makes its prediction. The samples can either be correctly or incorrectly predicted. The samples that are wrongly predicted are reused for training the next model. In this way, subsequent models can improve on the errors of previous models.

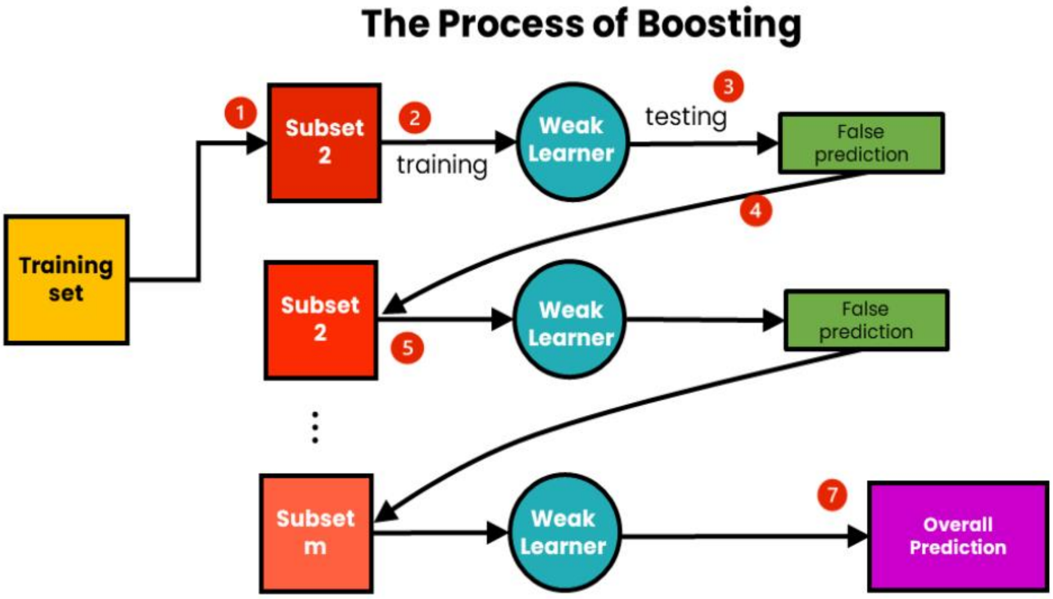
Unlike bagging, which aggregates prediction results at the end, boosting aggregates the results at each step. They are aggregated using weighted averaging.

Weighted averaging involves giving all models different weights depending on their predictive power. In other words, it gives more weight to the model with the highest predictive power. This is because the learner with the highest predictive power is considered the most important.

Boosting works with the following steps:

1. We sample m-number of subsets from an initial training dataset.
2. Using the first subset, we train the first weak learner.
3. We test the trained weak learner using the training data. As a result of the testing, some data points will be incorrectly predicted.
4. Each data point with the wrong prediction is sent into the second subset of data, and this subset is updated.
5. Using this updated subset, we train and test the second weak learner.

6. We continue with the following subset until the total number of subsets is reached.
7. We now have the total prediction. The overall prediction has already been aggregated at each step, so there is no need to calculate it.



Algorithm:

Input:

Data set D = {(X1, Y1), (X2, Y2), ..., (Xn, Yn)}

Number of iteration T

Process:

Step 1: Initialize Weight: Each case receives the same weight.

Wi = 1/N, where i = 1, 2, 3 … N.

Step 2: Construct a classifier using current weight, Compute its error:

$$Em = \frac{\sum wi \times I\{Yi \neq gm(xi)\}}{\sum wi}$$

Step 3: Get a classifier influence and update example weight.

$$am = \log\left(\frac{1 - Em}{Em}\right)$$

Step 4: Go to step 2.

- **Adaptive Boosting (AdaBoost):** is an ensemble of algorithms, where we build models on the top of several weak learners. As we mentioned earlier, those learners are called weak because they are typically simple with limited prediction capabilities. It is one of the simplest boosting algorithms. Usually, decision trees are used for modelling. Multiple sequential models are created, each correcting the errors from the last model. AdaBoost assigns weights to the observations which are incorrectly predicted and the subsequent model works to predict these values correctly.

The adaptation capability of AdaBoost made this technique one of the earliest successful binary classifiers. **Sequential** decision trees were the core of such adaptability where each tree is adjusting its weights based on prior knowledge of accuracies. Hence, we perform the training in such a technique in sequential rather than parallel process. In this technique, the process of training and measuring the error in estimates can be repeated for a given number of iteration or when the error rate is not changing significantly.

AdaBoost was the first boosting technique and is still now widely used in several domains. AdaBoost, in theory, is not prone to overfitting. Stage-wise estimation may slow down the learning process since parameters aren't jointly optimized. AdaBoost may be used to increase the accuracy of the weak classifiers, allowing it to be more flexible. It requires no normalization and has a low generalization error rate. However, training the algorithm takes enormous time. The method is also susceptible to noisy data and outliers. Therefore, removing them before employing them is strongly advised.

Looking at it step-by-step, this is what a AdaBoost model does:

1. Initially, all observations in the dataset are given equal weights.
2. A model is built on a subset of data.
3. Using this model, predictions are made on the whole dataset.
4. Errors are calculated by comparing the predictions and actual values.
5. While creating the next model, higher weights are given to the data points which were predicted incorrectly.
6. Weights can be determined using the error value. For instance, higher the error more is the weight assigned to the observation.
7. This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

- **Gradient Boosting:** Gradient Boosting or GBM is another ensemble machine learning algorithm that works for both regression and classification problems. GBM uses the boosting technique, combining a number of weak learners to form a strong learner. Regression trees used as a base learner, each subsequent tree in series is built on the errors calculated by the previous tree.Gradient boosting algorithms are great techniques that have high predictive performance. Xgboost, LightGBM, and CatBoost are popular boosting algorithms that can be used for regression and classification problems. Their popularity has significantly increased after their proven ability to win some Kaggle competitions.

- **Stacking**

Stacking, also known as Stacked Generalization,is use to improve the prediction accuracy of strong learners. Stacking aims to create a single robust model from multiple heterogeneous strong learners.

Stacking differs from bagging and boosting in that:

- It combines strong learners
- It combines heterogeneous models
- It consists of creating a Metamodel. A metamodel is a model created using a new dataset.

Individual heterogeneous models are trained using an initial dataset. These models make predictions and form a single new dataset using those predictions. This new data set is used to train the metamodel, which makes the final prediction. The prediction is combined using weighted averaging.

Because stacking combines strong learners, it can combine bagged or boosted models.

Stackingis a method similar to boosting. It is an interesting way of combining different models where multiple different algorithms are applied to the training dataset to create a model. The Meta classifier is used to predict unseen data accurately. They produce more robust predictors. It is a process of learning how to create such a stronger model from all weak learners' predictions.

It is an ensemble technique that combines multiple classifications or regression models via a meta-classifier or a meta-regressor. The base-level models are trained on a complete training set, then the meta-model is trained on the features that are outputs of the base-level model. The base-level often consists of different learning algorithms and therefore stacking ensembles are often heterogeneous.

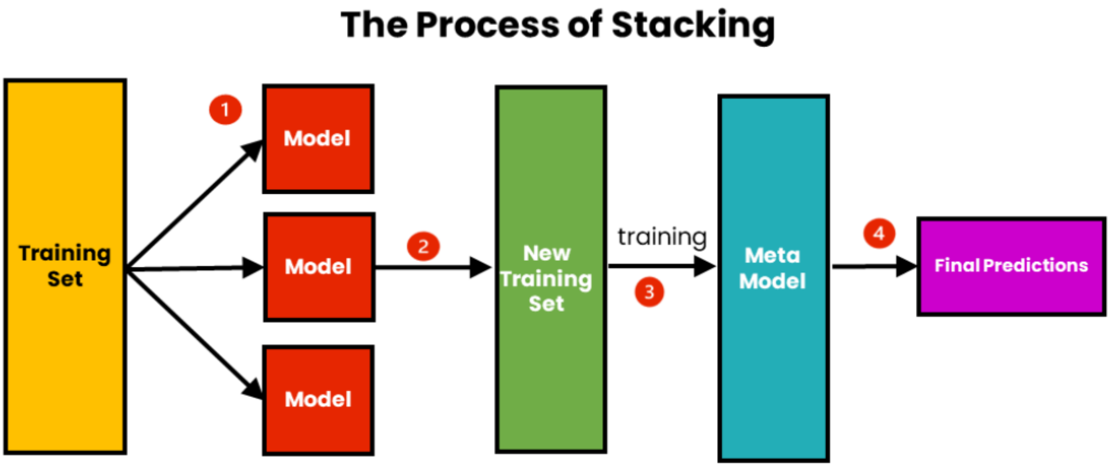
The models(Base-Model) in stacking are typically different (e.g. not all decision trees) and fit on the same dataset. Also, a single model(Meta-model) is used to learn how to best combine the predictions from the contributing models.

The architecture of a stacking model involves two or more base models, often referred to as level-0 models and a meta-model. Meta-model, also referred to as a level-1 model combines the predictions of the base models.

The steps of Stacking are as follows:

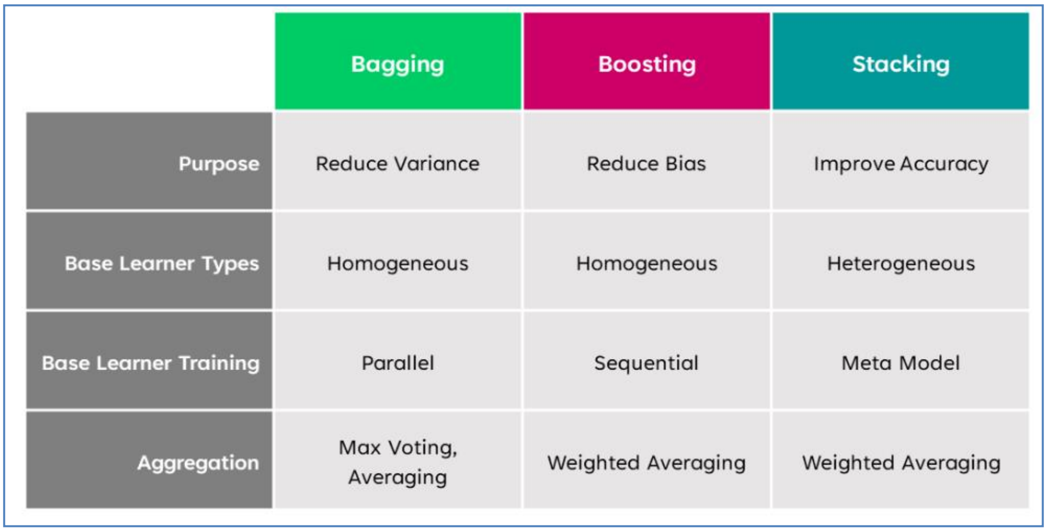
1. We use initial training data to train m-number of algorithms.
2. Using the output of each algorithm, we create a new training set.
3. Using the new training set, we create a meta-model algorithm.
4. Using the results of the meta-model, we make the final prediction. The results are combined using weighted averaging.

The outputs from the base models used as input to the meta-model may be real values in the case of regression, and probability values, probability like values, or class labels in the case of classification.



Please note that what is being learned here (as features) is the prediction from each model.

When to use Bagging, Boosting and Stacking?



	Bagging	Boosting	Stacking
Purpose	Reduce Variance	Reduce Bias	Improve Accuracy
Base Learner Types	Homogeneous	Homogeneous	Heterogeneous
Base Learner Training	Parallel	Sequential	Meta Model
Aggregation	Max Voting, Averaging	Weighted Averaging	Weighted Averaging

- If you want to reduce the overfitting or variance of your model, you use bagging. If you are looking to reduce underfitting or bias, you use boosting. If you want to increase predictive accuracy, use stacking.
- Bagging and boosting both works with homogeneous weak learners. Stacking works using heterogeneous solid learners.
- All three of these methods can work with either classification or regression problems.
- One disadvantage of boosting is that it is prone to variance or overfitting. It is thus not advisable to use boosting for reducing variance. Boosting will do a worse job in reducing variance as compared to bagging.
- On the other hand, the converse is true. It is not advisable to use bagging to reduce bias or underfitting. This is because bagging is more prone to bias and does not help reduce bias.
- Stacked models have the advantage of better prediction accuracy than bagging or boosting. But because they combine bagged or boosted models, they have the disadvantage of needing much more time and computational power. If you are looking for faster results, it’s advisable not to use stacking. However, stacking is the way to go if you’re looking for high accuracy.

References

- Larose, DT.(2006).*Data Mining Methods and Models*. Wiley-Interscience, New Jersey, USA.
- Han, J., Kamber, M., Pei, J. (2012).*Data mining: concepts and techniques.* Morgan Kaufmann, Elsevier, USA.

ASSOCIATION RULES MINING USING R

Dr. Anshu Bharadwaj
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
Anshu.Bharadwaj@icar.gov.in

1. Introduction

Mining association rules is one of the most useful data mining applications. Association rules, were first introduced in 1993 [Agrawal1993], and are used to identify relationships among a set of items in a database. These relationships are not based on inherent properties of the data themselves (as in the case of functional dependencies), but are rather based on co-occurrence of the data items. Association rules are mainly used to analyze transactional data. The association rules are useful in management, to increase the effectiveness and/or reduce the cost associated with advertising, marketing, inventory, stock location on the floor etc. Association rules also provide assistance in other applications such as prediction by identifying what events occur before a set of particular events. An association rule may be one of the following types: Boolean, Spatial, Temporal, Generalized, Quantitative, Interval, and Multiple Min-Support Association etc or a mix of them.

Formally the association rule as stated in [Agrawal1993] and [Cheung1996] is,

Let D be a transaction database and $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct items (attributes) of D, where each transaction (record) T is a set of items such that $T \subseteq I$ and has unique identifier. A transaction T is said to contain a set of item A if and only if $A \subseteq T$. An *association rule* is of the form of an implication expression $A \Rightarrow B$, where A, $B \subset$ I, are sets of items called *itemsets*, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction data D with *support (s)* where s is the ratio (in percent) of the records that contain $A \cup B$ (i.e. both A and B) to the total number of records in the database, i.e. the probability $P(A \cup B)$. The rule $A \Rightarrow B$ has *confidence* (c) in the D, the ratio (in percent) of the number of records that contain $\mathrm{A} \cup \mathrm{B}$ to the number of records that contain A. This is taken to be the conditional probability $P(B \mid A)$. Mining of association rules from a database consists of finding all rules that meet the user-specified thresholds of support and confidence termed as minimum support and minimum confidence. The problem of mining association rules has been decomposed into the following two subproblems [Agrawal1994]:

1) To find all sets of items which occur with a frequency that is greater than or equal to the user-specified threshold support, say *s*.
2) To generate the rules using the frequent itemsets, which have confidence greater than or equal to the user-specified threshold confidence, say c.

The Association relationships are not based on inherent properties of the data themselves but rather based on co-occurrence of the data items. Application of association rules spans across a wide range of domains such as, business, finance, health, geographical information system, weather forecast and many such areas of real life application. The association rules in management may be handy to increase the effectiveness and/or reduce the cost associated with advertising, marketing, inventory, stock location on the floor etc. Association rules could assist in prediction of an event co-occurrence of a set of events. Association rules are generally categorized in following types: Boolean, Spatial, Temporal, Generalized, Quantitative, and Interval or may be mixed of them. The above definition of association rule is also known as Boolean Association Rule.

Association rule mining is:

- Unsupervised learning
- Used for pattern discovery
- Each rule has form: A -> B, or Left -> Right

For example: "70% of customers who purchase 2% milk will also purchase whole wheat bread."

Data mining using association rules is the process of looking for strong rules:

1. Find the large itemsets (i.e. most frequent combinations of items)
2. Generate association rules for the above itemsets.

2. Performance Evaluation Measure of Association Rules

How to measure the strength of an association rule? Using support/confidence

Support: Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both A and B, i.e.

Support = Probability(A and B)

Support = (# of transactions involving A and B) / (total number of transactions).

Confidence: Confidence is the strength of implication of a rule; it is the percentage of transactions that contain B if they contain A, ie.

Confidence = Probability (B if A) = P(B/A)

Confidence =

(# of transactions involving A and B) / (total number of transactions that have A).

3. The Apriori Algorithm

The Apriori Algorithm is an influential algorithm for miningfrequent itemsets for boolean association rules. Some keyconcepts for Apriori algorithm are:

- Frequent Itemsets: The sets of item which hasminimum support (denoted by Li for ith-Itemset).
- Apriori Property: Any subset of frequent itemset mustbe frequent.
- Join Operation: To find Lk , a set of candidate kitemsets is generated by joining Lk-1 with itself.

Very first algorithm proposed for association rules miningwas the Apriori for frequent itemset mining. The mostpopular algorithm for pattern mining is without a doubt Apriori.It is designed to be applied on a transaction database to discover patterns in transactions made by customers in stores. But it can also be applied in several other applications. A transaction is defined a set of distinct items (symbols).

Aprioritakes as input

(1) a minsup threshold set by the user and

(2) atransaction database containing a set of transactions.

Apriorioutputs all frequent itemsets, i.e. groups of items shared by noless than minsup transactions in the input database. Forexample, consider the following transaction data base containing four transactions. Given a minsup of twotransactions, frequent itemsets are“bread, butter”, “breadmilk”, “bread”, “milk” and “butter”.

T1: bread, butter, spinach

T2: butter, salmon

T3: bread, milk, butter

T4: cereal, bread, milk

The Apriori algorithm employs the downward closureproperty if an item set is not frequent, any superset of it cannotbe frequent either. The Apriori algorithm performs a breadthfirstsearch in the search space by generating candidate k+1-itemsets from frequent k itemsets.

The frequency of an item set is computed by counting its occurrence in each transaction. Apriori is an significantalgorithm for mining frequent itemsets for Boolean associationrules. Since the Algorithm uses prior knowledge of frequentitem set it has been given the name Apriori. Apriori is aniterative level wise search Algorithm, where k- itemsets areused to explore (k+1)-itemsets. First, the set of frequents 1- itemsets is found.

This set is denoted by L1. L1 is used to find L2, the set offrequent 2-itemsets , which is used to find L3 and so on , untilno more frequent k-itemsets can be found. The finding of eachLk requires one full scan of database.

There are twosteps for understanding that how Lk-1 is usedto find Lk:-

1) The join step: To find Lk , a set of candidate k-itemsets isgenerated by joining Lk-1 with itself. This set ofcandidates is denoted Ck.

2) The prune step: Ck is a superset of Lk , that is , itsmembers may or may not be frequent , but all of thefrequent k-itemsets are included in Ck .

A scan of the database to determine the count of eachcandidate in Ck would result in the determination of Lk. Ck,however, can be huge, and so this could involve heavycomputation.

To reduce the size of Ck , the Apriori property is used as follows:

i. Any (k-1)-item set that is not frequent cannot be asubset of frequent k-item set.
ii. Hence, if (k-1) subset of a candidate k item set is notin Lk-1 then the candidate cannot be frequent eitherand so can be removed from C.

Based on the Apriori property that all subsets of a frequentitemset must also be frequent, we can determine that four lattercandidates cannot possibly be frequent. How?

For example, let’s take {I1, I2, I3}. The 2-item subsets of itare {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1,I2, I3} are members of L2, We will keep {I1, I2, I3} in C3.

Let’s take another example of {I2, I3, I5} which shows howthe pruning is performed. The 2-item subsets are {I2, I3}, {I2,I5} & {I3,I5}.

BUT, {I3, I5} is not a member of L2 and hence it is notfrequent violating Apriori Property. Thus, we will have toremove {I2, I3, I5} from C3.

Therefore, C3 = {{I1, I2, I3}, {I1, I2, I5}} after checking forall members of result of Join operation for Pruning.

Example : The Titanic Dataset

The Titanic dataset in the datasets package is a 4-dimensional table with summarized information on the fate of passengers on the Titanic according to social class, sex, age and survival. I To make it suitable for association rule mining, we reconstruct the raw data as titanic.raw, where each row represents a person. The reconstructed raw data can also be downloaded at http://www.rdatamining.com/data/titanic.raw.rdata.

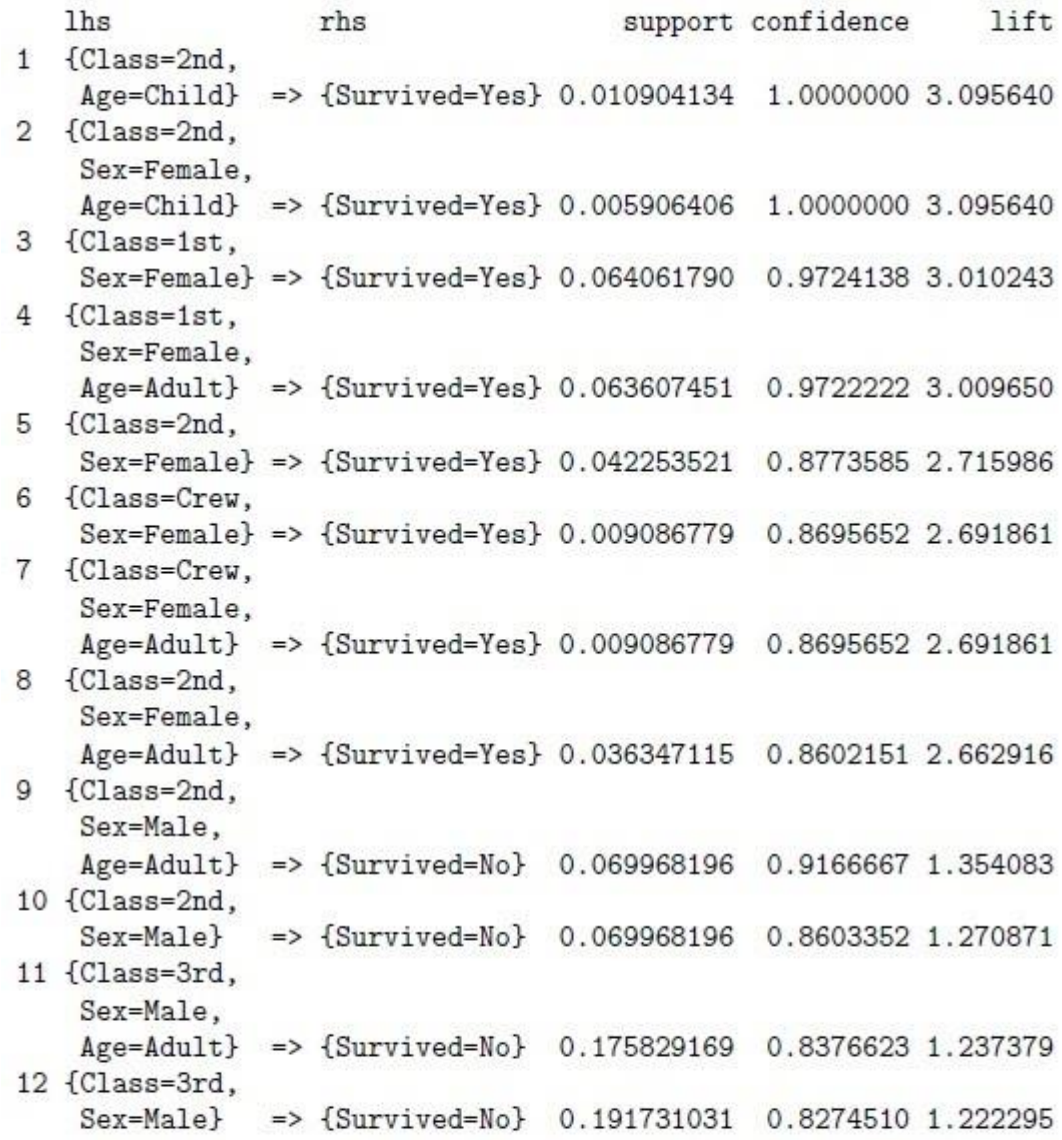
```
> str(titanic.raw)
'data.frame': 2201 obs. of 4 variables:
$ Class : Factor w/ 4 levels "1st","2nd","3rd",..: 3 3 3 3 3 3 3 3 3 3 ...
$ Sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
$ Age : Factor w/ 2 levels "Adult","Child": 2 2 2 2 2 2 2 2 2 2 ...
$ Survived: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Association Rule Mining

```
> library(arules)
> # find association rules with default settings
> rules <- apriori(titanic.raw)
> inspect(rules)
  lhs              rhs          support   confidence lift
1 {}            => {Age=Adult} 0.9504771 0.9504771  1.0000000
2 {Class=2nd}    => {Age=Adult} 0.1185825 0.9157895  0.9635051
3 {Class=1st}    => {Age=Adult} 0.1449341 0.9815385  1.0326798
4 {Sex=Female}   => {Age=Adult} 0.1930940 0.9042553  0.9513700
5 {Class=3rd}    => {Age=Adult} 0.2848705 0.8881020  0.9343750
6 {Survived=Yes} => {Age=Adult} 0.2971377 0.9198312  0.9677574
7 {Class=Crew}   => {Sex=Male}  0.3916402 0.9740113  1.2384742
```

We then set rhs=c("Survived=No", "Survived=Yes") in appearance to make sure that only "Survived=No" and "Survived=Yes" will appear in the rhs of rules.

```
> # rules with rhs containing "Survived" only
> rules <- apriori(titanic.raw,
  + parameter = list(minlen=2, supp=0.005, conf=0.8),
  + appearance = list(rhs=c("Survived=No", "Survived=Yes"),
  + default="lhs"),
  + control = list(verbose=F))
> rules.sorted <- sort(rules, by="lift")
> inspect(rules.sorted)
```

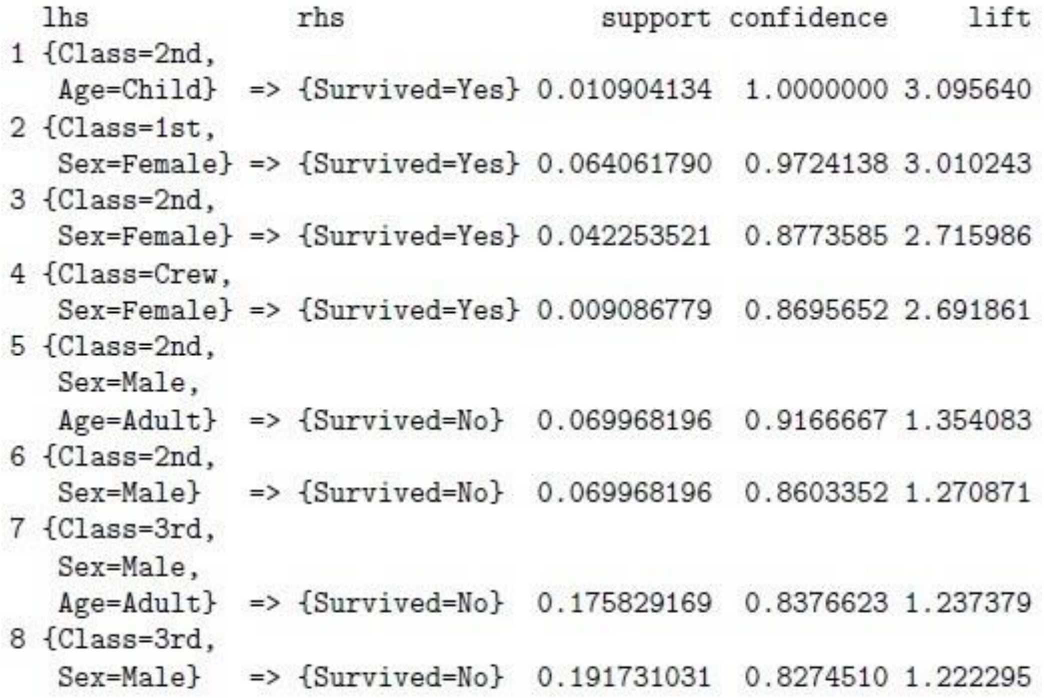


```
   lhs               rhs                support confidence     lift
1  {Class=2nd,
    Age=Child}   => {Survived=Yes} 0.010904134  1.0000000 3.095640
2  {Class=2nd,
    Sex=Female,
    Age=Child}   => {Survived=Yes} 0.005906406  1.0000000 3.095640
3  {Class=1st,
    Sex=Female} => {Survived=Yes} 0.064061790  0.9724138 3.010243
4  {Class=1st,
    Sex=Female,
    Age=Adult}   => {Survived=Yes} 0.063607451  0.9722222 3.009650
5  {Class=2nd,
    Sex=Female} => {Survived=Yes} 0.042253521  0.8773585 2.715986
6  {Class=Crew,
    Sex=Female} => {Survived=Yes} 0.009086779  0.8695652 2.691861
7  {Class=Crew,
    Sex=Female,
    Age=Adult}   => {Survived=Yes} 0.009086779  0.8695652 2.691861
8  {Class=2nd,
    Sex=Female,
    Age=Adult}   => {Survived=Yes} 0.036347115  0.8602151 2.662916
9  {Class=2nd,
    Sex=Male,
    Age=Adult}   => {Survived=No}  0.069968196  0.9166667 1.354083
10 {Class=2nd,
    Sex=Male}    => {Survived=No}  0.069968196  0.8603352 1.270871
11 {Class=3rd,
    Sex=Male,
    Age=Adult}   => {Survived=No}  0.175829169  0.8376623 1.237379
12 {Class=3rd,
    Sex=Male}    => {Survived=No}  0.191731031  0.8274510 1.222295
```

Pruning Redundant Rules

In the above result, rule 2 provides no extra knowledge in addition to rule 1, since rules 1 tells us that all 2nd-class children survived. Generally speaking, when a rule (such as rule 2) is a super rule of another rule (such as rule 1) and the former has the same or a lower lift, the former rule (rule 2) is considered to be redundant. Below we prune redundant rules.

```
> # find redundant rules
> subset.matrix <- is.subset(rules.sorted, rules.sorted)
> subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
> redundant <- colSums(subset.matrix, na.rm=T) >= 1
> which(redundant)
[1] 2 4 7 8
> # remove redundant rules
> rules.pruned <- rules.sorted[!redundant]
> inspect(rules.pruned)
```

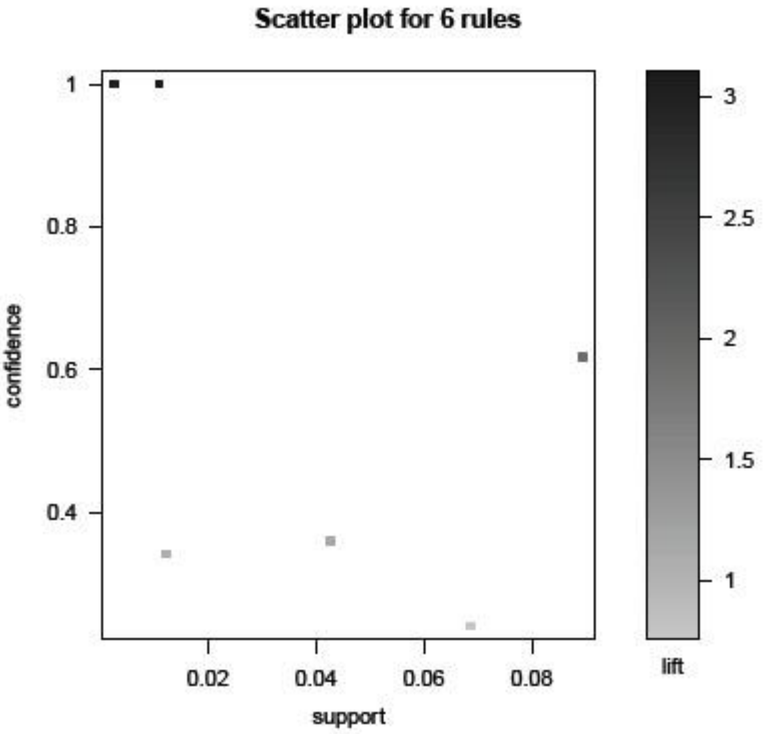


```
  lhs              rhs                 support confidence     lift
1 {Class=2nd,
   Age=Child}  => {Survived=Yes} 0.010904134  1.0000000 3.095640
2 {Class=1st,
   Sex=Female} => {Survived=Yes} 0.064061790  0.9724138 3.010243
3 {Class=2nd,
   Sex=Female} => {Survived=Yes} 0.042253521  0.8773585 2.715986
4 {Class=Crew,
   Sex=Female} => {Survived=Yes} 0.009086779  0.8695652 2.691861
5 {Class=2nd,
   Sex=Male,
   Age=Adult}  => {Survived=No}  0.069968196  0.9166667 1.354083
6 {Class=2nd,
   Sex=Male}   => {Survived=No}  0.069968196  0.8603352 1.270871
7 {Class=3rd,
   Sex=Male,
   Age=Adult}  => {Survived=No}  0.175829169  0.8376623 1.237379
8 {Class=3rd,
   Sex=Male}   => {Survived=No}  0.191731031  0.8274510 1.222295
```

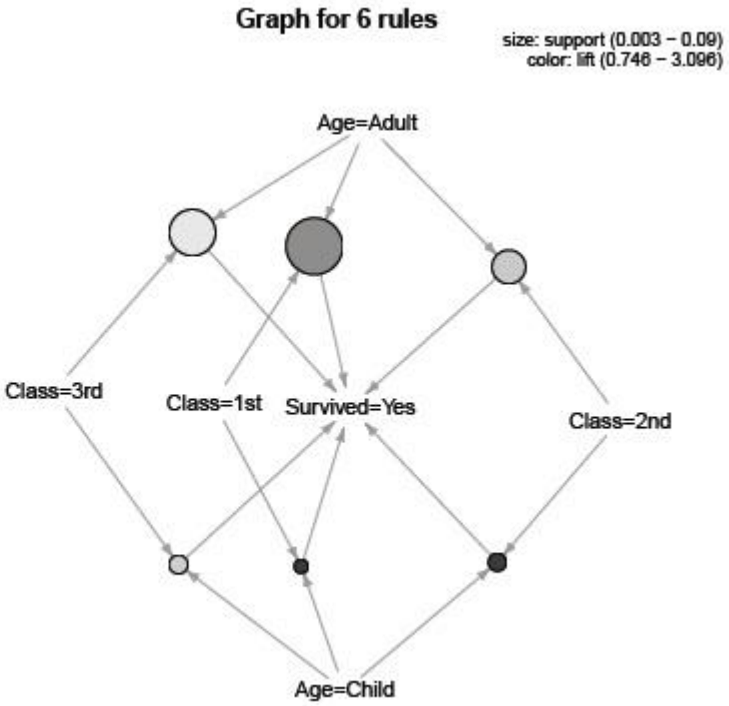
Visualizing Association Rules

Package arules Viz supports visualization of association rules with scatter plot, balloon plot, graph, parallel coordinates plot, etc.

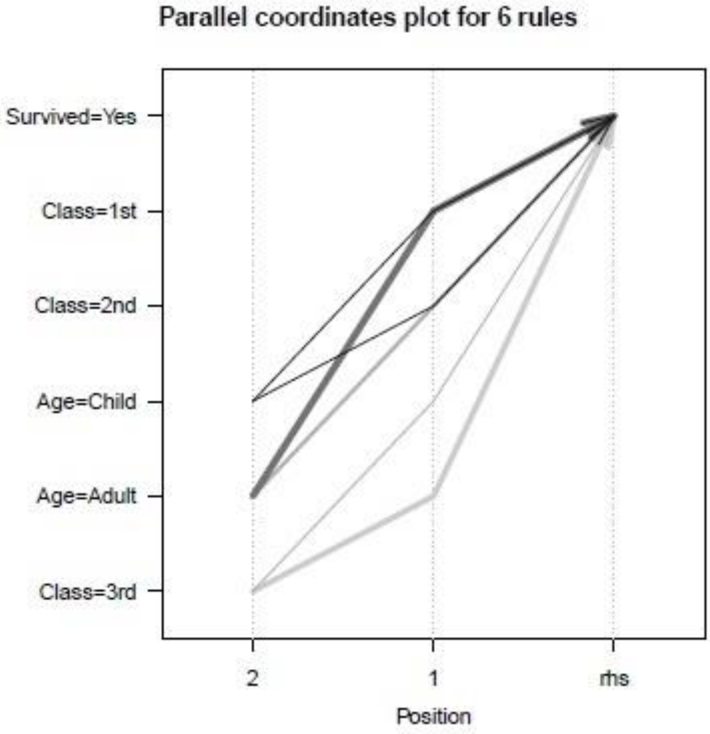
```
> library(arulesViz)
> plot(rules)
```



```
> plot(rules, method="graph", control=list(type="items"))
```



```
> plot(rules, method="paracoord", control=list(reorder=TRUE))
```



4. Frequent Pattern (FP) Growth Method

The FP-Growth Algorithm is an alternative way to find frequent itemsets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the itemset association information.

In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step it divides the compressed database into a set of conditional databases, each one associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity.

In large databases, it’s not possible to hold the FP-tree in the main memory. A strategy to cope with this problem is to firstly partition the database into a set of smaller databases (called projected databases), and then construct an FP-tree from each of these smaller databases.

4.1 FP-Tree structure

The frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database [4].

Han defines the FP-tree as the tree structure io below [1]:

1. One root labeled as “null” with a set of item-prefix subtrees as children, and a frequent-item-header table (presented in the left side of Figure 1);
2. Each node in the item-prefix subtree consists of three fields:
    1. Item-name: registers which item is represented by the node;
    2. Count: the number of transactions represented by the portion of the path reaching the node;
    3. Node-link: links to the next node in the FP-tree carrying the same item-name, or null if there is none.
   1. Each entry in the frequent-item-header table consists of two fields:
       1. Item-name: as the same to the node;
       2. Head of node-link: a pointer to the first node in the FP-tree carrying the item-name.

   Additionally the frequent-item-header table can have the count support for an item. The Figure below show an example of a FP-tree.

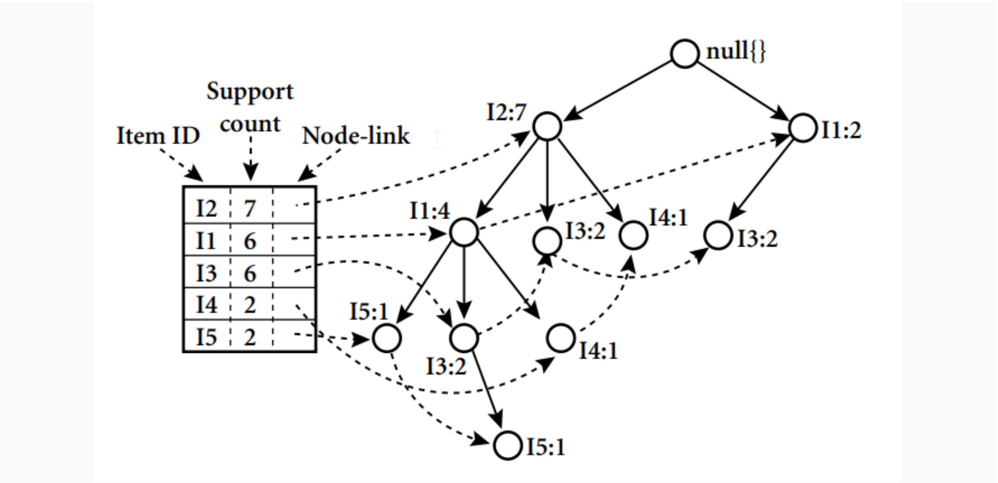
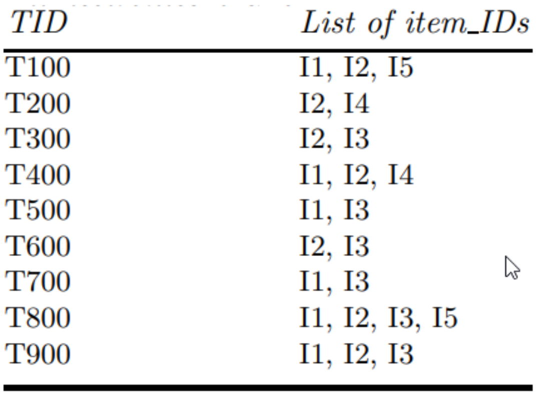


Figure 1: An FP-tree registers compressed, frequent pattern information

Table 1: Transactional data for an AllElectronics branch.

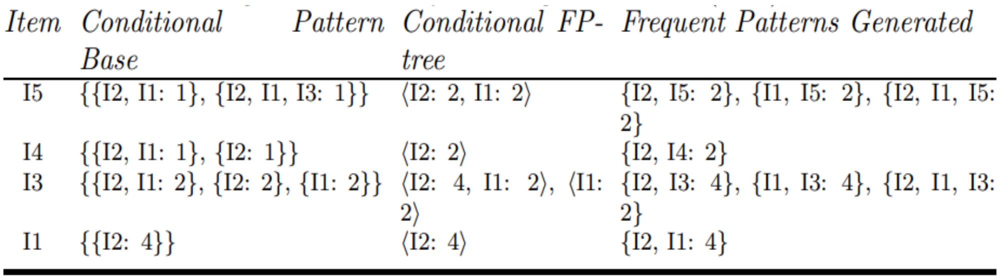


TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

The first scan of the database is the same as Apriori, which derives the set of frequent items (1-itemsets) and their support counts (frequencies). Let the minimum support count be 2. The set of frequent items is sorted in the order of descending support count. This resulting set or list is denoted by L. Thus, we have L ={{I2: 7}, {I1: 6}, {I3: 6}, {I4: 2}, {I5: 2}}. An FP-tree is then constructed as follows. First, create the root of the tree, labeled with "null." Scan database D a second time. The items in each transaction are processed inL order (i.e., sorted according to descending support count), and a branchis created for each transaction. For example, the scan of thefirst transaction, "T100: I1, I2, I5," which contains three items (I2, I1, I5 in L order), leads to the construction of the first branch of the tree with three nodes,hI2: 1i,hI1:1i, and hI5: 1i, where I2islinked as a child to the root, I1islinked to I2, and I5islinked to I1. The second transaction, T200, contains theitems I2 and I4inLorder, whichwould result in a branch where I2 is linked to the root and I4 is linked to I2. However, this branch would share a common prefix, I2, with the existing path for T100. Therefore, we insteadincrement the count of the I2 node by 1, and create a new node,hI4: 1i, which is linked as a child to hI2: 2i. In general, when considering the branch to be addedfor a transaction, the count of each node along a common prefix is incremented by 1, and nodes for the items following the prefix are created and linked accordingly. To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links. The tree obtained after scanning all of the transactions is shown in Figure 6.7 with the associated node-links. In this way, the problem of mining frequent patterns in databases is transformed to that of mining the FP-tree. The FP-tree is mined as follows. Start from each frequent length-1 pattern (as an initial suffix pattern), construct its conditional pattern base (a "sub-database," which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern), then construct its (conditional) FP-tree, and perform mining recursively on such a tree. The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree.

Mining of the FP-tree is summarized in Table 2 and detailed as follows. We first consider I5, which is the last item in L, rather than the first. The reason for starting at the end of the list will become apparent as we explain the FP-tree mining process. I5 occurs in two branches of the FP-tree of Figure 2. (The occurrences of I5 can easily be found by following its chain of node-links.) The paths formed by these branches are hI2, I1, I5: 1i and hI2, I1, I3, I5: 1i. Therefore, considering I5 as a suffix, its corresponding two prefix paths are hI2, I1: 1i and hI2, I1, I3: 1i, which form its conditional pattern base. Using this conditional pattern base as a transaction database, we build an I5-conditional FP-tree, which contains only a single path, hI2: 2, I1: 2i; I3 is not included because its support count of 1 is less than the minimum support count. The single path generates all the combinations of frequent patterns: {I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}. For I4, its two prefix paths form the conditional pattern base, {{I2 I1: 1}, {I2: 1}}, which generates a single-node conditional FP-tree, hI2: 2i, and derives one frequent pattern, {I2, I4: 2}

Table 2: Mining the FP-tree by creating conditional (sub-)pattern bases



Item	*Conditional Pattern Base*	*Conditional FP-tree*	*Frequent Patterns Generated*
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	⟨I2: 2, I1: 2⟩	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	⟨I2: 2⟩	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	⟨I2: 4, I1: 2⟩, ⟨I1: 2⟩	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	⟨I2: 4⟩	{I2, I1: 4}

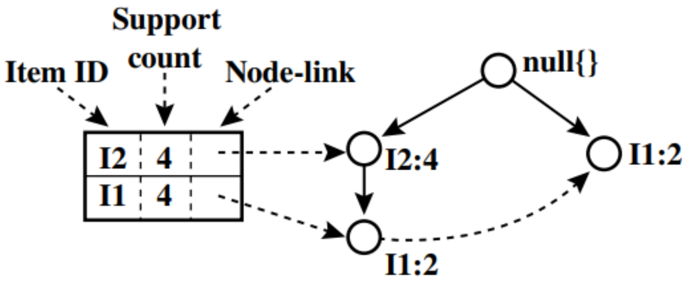


Figure 2: The conditional FP-tree associated with the conditional node I3

Similar to the above analysis, I3's conditional pattern base is {{I2, I1: 2}, {I2: 2}, {I1: 2}}. Its conditional FP-tree has two branches, hI2: 4, I1: 2i and hI1: 2i, as shown in Figure 6.8, which generates the set of patterns {{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}}. Finally, I1's conditional pattern base is {{I2: 4}}, whose FP-tree contains only one node, hI2: 4i, which generates one frequent pattern, {I2, I1: 4}. This mining process is summarized in Figure 6.9. The FP-growth method transforms the problem of finding long frequent patterns to searching for shorter ones in much smaller conditional databases recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs.

5. Basic Association Rules: Problems, Solutions and New Applications

Most of the research efforts in the scope of association rules have been oriented to simplify the rule set and to improve performance of algorithm. But these are not the only problems that can be found and when rules are generated and applied in different domains. Troubleshooting for them should also take into consideration the purpose of association model and data they come from. Some of the major drawbacks of association rule algorithms are as follows:

- Obtaining huge number of rules
- Obtaining non interesting rules

- Low algorithm performance
- Cannot incorporate domain/ user defined knowledge
- Not suitable for supervised learning

Some of the recent studies have focused on overcoming these limitations. Many algorithms for obtaining a reduced number of rules with high **support** and **confidence** have been produced. However these measures are insufficient to determine if discovered associations are really useful. An important property of discovered association rules is that they should be **interesting** and **useful**. Though interestingness of rule is a subjective aspect, many researchers have tried to come up with some ways of **measuring of interest**. It has been suggested that the rules are interesting if they are **unexpected** (unknown to user) and **actionable** (users can do something with them to their advantage). Further some other measures namely: **any-confidence, all confidence and bond** has been suggested as alternative measures of interestingness. Some authors have considered alternative measures of interest as : **gini index, entropy gain or chisquared for database or a measure of implication called conviction.** Most of the approaches for finding interesting rules require user participation to articulate his knowledge or to express what rules are interesting for him. Systems have been developed to analyze the discovered rules against user's knowledge. Discovered rules can be pruned to remove redundant and insignificant rules and further user's evaluation can be used to rank the rules. Unexpected patterns discovered may represent "holes" in domain knowledge which needs to be resolved. These patterns can thus be used to refine already existing beliefs.

Traditionally, association analysis has been considered as an unsupervised technique, so it has been applied for knowledge discovery tasks. Recent studies have shown that knowledge discovery algorithms such as association rule mining can be successfully applied for prediction in classification problems. In such cases the algorithms used for generating association rules must be tailored to peculiarities of predictions in order to build effective classifiers. Some work has been done, where association mining algorithms have been extended so that they can be used for classification/ prediction. A proposal of this category is Classification Based on Association (CBA) algorithm. The algorithm consists of two parts, a rule generator for finding association rules and a classifier builder based on these rules. Main contribution of this algorithm is possibility of making prediction on any attribute in database. Moreover, new incomplete observations can be classified.

In conclusion we can say that association rule mining is an important area of data mining research and a comparatively a younger member of data mining community. In addition to finding co-occurrence relation between items, which is basic objective, the algorithm has been applied for diverse applications. Many extensions of standard methods have been proposed. A major research area on association rules is interestingness of discovered rules. In fact its potential has still to be tapped, so that it can be tailored to solve different types of data mining problems.

STATISTICAL ANALYSIS USING SPSS SOFTWARE

Raju Kumar
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
raju.kumar@icar.gov.in

1. Introduction

SPSS (Statistical Package for the Social Sciences) is a widely used software program for statistical analysis and data management. It provides a comprehensive set of tools and features that enable researchers, data analysts, and students to perform various data-related tasks efficiently. SPSS is known for its user-friendly interface and powerful capabilities, making it a popular choice in both academia and industry.

Originally developed in 1968 by Norman H. Nie, C. Hadlai "Tex" Hull, and Dale H. Bent. The original SPSS manual (Nie*et al.*, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis. Originally it is an acronym of *Statistical Package for the Social Science* but now it stands for *Statistical Product and Service Solutions*. The current versions are officially named IBM SPSS Statistics. Long produced by SPSS Inc., it was acquired by IBM in 2009. During 2009 and 2010 it was called *PASW* (*Predictive Analytics Software*) *Statistics*.SPSS has evolved over the years and is now owned by IBM Corporation. The software has undergone several versions, with each release bringing new functionalities and enhancements to meet the ever-growing demands of statistical analysis.

SPSS allows users to import, manipulate, and analyze data from a wide range of sources, including spreadsheets, databases, and other statistical formats. The software supports both structured and unstructured data, making it versatile for different types of research and analysis. Whether you are working with survey data, experimental data, or observational data, SPSS provides the necessary tools to handle and explore your datasets effectively.

One of the key strengths of SPSS is its extensive range of statistical procedures. The software offers a vast array of statistical techniques, ranging from basic descriptive statistics to advanced multivariate analysis. Users can easily generate frequencies, descriptive statistics, cross-tabulations, and explore relationships between variables. Moreover, SPSS provides options for regression analysis, analysis of variance (ANOVA), factor analysis, cluster analysis, and many other techniques that allow for in-depth data exploration and hypothesis testing.

SPSS also provides a variety of graphical tools for visualizing data. Users can create charts, histograms, scatterplots, and other visual representations to better understand their data and communicate findings effectively. The software supports customization options, enabling users to format and design visuals to suit their specific needs.

In addition to its analytical capabilities, SPSS offers data management features to assist users in preparing and cleaning datasets. With SPSS, users can merge, subset, transform, and recode variables, ensuring data quality and consistency. This helps researchers save time and effort in data preparation, allowing them to focus more on analysis and interpretation.

SPSS is known for its user-friendly interface, making it accessible to users with varying levels of statistical knowledge and programming skills. The software offers a menu-driven interface, where users can perform tasks by selecting options from dropdown menus. However, for more advanced users, SPSS also supports a syntax-based approach, allowing for greater flexibility and automation in data analysis.

Furthermore, SPSS provides options for integration with other statistical software and programming languages. Users can import and export data in various formats, such as Excel, CSV, and SQL, facilitating seamless data exchange between different software tools. SPSS also supports integration with R and Python, allowing users to leverage the power of these programming languages for custom analyses and extensions.

In conclusion, SPSS is a powerful and versatile software program for statistical analysis and data management. With its user-friendly interface, extensive statistical procedures, and data visualization capabilities, SPSS enables researchers and data analysts to explore, analyze, and interpret data efficiently. Its wide range of features and compatibility with other software tools make SPSS a valuable asset in various fields, including social sciences, market research, healthcare, and more.

Some versions of SPSS released in recent years are

- SPSS Statistics 17.0.1 - December 2008
- PASW Statistics 17.0.3 - September 2009
- PASW Statistics 18.0, 18.0.1, 18.0.2, 18.0.3
- IBM SPSS Statistics 19.0 - August 2010
- IBM SPSS Statistics 19.0.1, 20.0, 20.0.1, 21.0, 22.0, 23.0, 24.0,25.0,26.0,27,28,29

Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services).

2.Opening SPSS

Depending on how the computer you are working on is structured, you can open SPSS in one of two ways.

1. If there is an SPSS shortcut like this on the desktop, simply put the cursor on it and double click the left mouse button.

2. Click the left mouse button on the button on your screen, then put your cursor on **Programs** or **All Programs** and left click the mouse. Select **SPSS 17.0 for Windows or IBM SPSS STATISTICS20 by** clicking the left mouse button. Either approach will launch the program.

3. Key Featuresof SPSS

Some of the key features of SPSS are

- It is easy to learn and use with its pull-down menu features
- It includes a full range of data management system and editing tools
- It offers comprehensive range of plotting, reporting and presentation features.
- It provides in-depth statistical analysis capabilities

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary stored in the datafile) are features of the base software. There are varieties of statistics included in the base software. Some of the important statistics are:

Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, DescriptiveRatio Statistics etc.

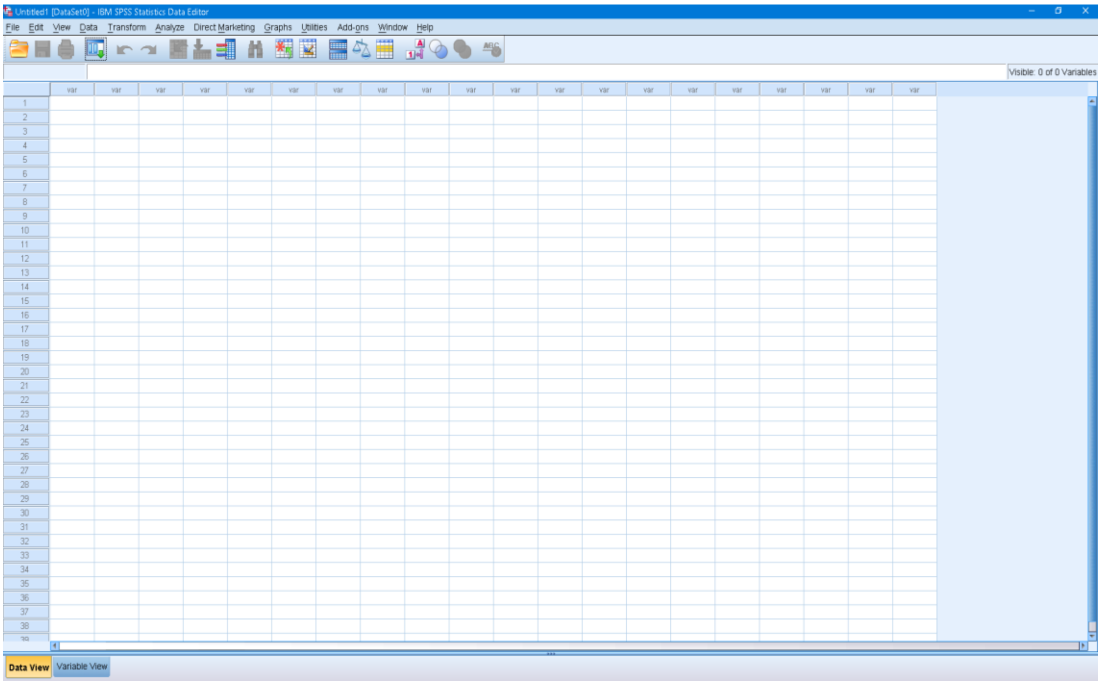
Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), nonparametric tests etc.

Prediction for numerical outcomes: Linear regression, Multiple Regression

Prediction for identifying groups: Factor analysis, Cluster analysis (two-step, K-means,hierarchical),Discriminant analysis etc.

4. Layout of SPSS

Data Editor: This graphical user interface displays the contents of the data file. One can create new data files or modify existing ones. The Data Editor window opens automatically when an SPSS session is started. The Data Editorwindow has two views that can be selected from the lower left hand side of the screen. Data Viewis where you see the data you are using. Variable Viewis where you can specify the format of your data when you are creating a file or where you can check the format of a pre-existing file. The data in the Data Editoris saved in a file with the extension .sav.The data editor offers a simple and efficient spreadsheet-like facility for entering data and browsing the working data file. To invoke SPSS in the windows environment, select the appropriate **SPSS** icon.


Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 0 of 0 Variables
var
Data View
Variable View

One can have only one data file open at a time. This editor has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS window.

- ✓ **Data view**: Displays the actual data values or defined value labels. The 'Data View' shows a spreadsheet view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells. One can modify data values in the Data view in many ways like change data values; cut, copy and paste data values; add and delete cases;
- ✓ **Variable view**: Displays variable definition information contained or metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics. One can modify variable properties in the Variable view for example, add and delete variables, change the order of variables etc.

Cells in both views can be manually edited, defining the file structure and allowing data entry without using command syntax. This may be sufficient for small datasets. Larger datasets such as statistical surveys aremore often created in data entry software, or entered during computer-assisted personal interviewing, by scanning and using optical character recognition and optical mark recognition software, or by direct capture from online questionnaires. These datasets are then read into SPSS. Extension of the saved data file will be ".sav".

Viewer: All results, tables, and charts performed by different statistical analysis are displayed in the Viewer. Extension of the saved output file will be ".spv". One can use the Viewer to browse results, show or hide selected tables and charts, change the display order of results by moving selected items or move items between the Viewer and other applications. The output presented in Viewer can be edited and saved for later use. A Viewer window opens automatically the first time a procedure is run that generates output. The Viewer is divided into two panes:

- ✓ The left pane contains an outline view of the contents. One can click an item in the outline to go directly to the corresponding table or chart.
- ✓ The right pane contains statistical tables, charts, and text output.

Syntax Editor: The pull-down menu interface generates command syntax: this can be displayed in the output. These command syntax can also be pasted into a syntax file in a syntax window using the "paste" button present in each menu. One can then edit the command syntax toutilize special features of SPSS not available through dialog boxes. These commands can be saved in a file for use in subsequent SPSS sessions. Extension of the saved syntax file will be ".sps". Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses. Additionally, some complex applications can only be programmed in syntax that are not accessible through the menu structure.

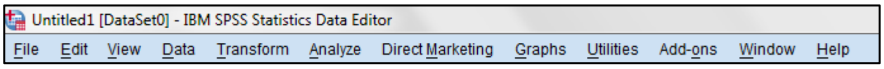
Pivot Table Editor: The results from most statistical procedures are displayed in pivot tables. These pivot tables outputs can be modified in many ways with pivot table editor. One can edit text, swap data in rows and columns, create multidimensional tables, and selectively hide and show results. Changing the layout of the table does not affect the results. Instead, it's a way to display information in a different or more desirable manner.

Text Output Editor: Text output not displayed in pivot tables can be modified with the Text Output Editor. One can edit the output and change font characteristics (type, style, colour, size).

Chart Editor: High-resolution charts and plots can be modified in chart windows. One can change the colours, select different type of fonts and sizes, switch the horizontal and vertical axes, rotate 3-D scatterplots, and even change the chart type.

Script Window: It provides the opportunity to write full-blown programs, in a BASIC-like language. It is a text editor for syntax composition. Extension of the saved script file will be ".sbs"

Many features of SPSS Statistics are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. Many of the tasks that are to be performed with SPSS start with **menu** selections. Each window has its own menu bar with menu selections appropriate for that window type. The various menu options available in SPSS are



Most menu selections open dialog boxes. One can use dialog boxes to select variables and options for analysis. Since most procedures provide a great deal of flexibility, not all of the possible choices can be contained in a single dialog box. The main dialog box usually contains the minimum information required to run a procedure. Additional specifications are made in sub-dialog boxes. All these above mentioned options have further sub-options. To see what applications there are, we simply move the cursor to a particular option and press, when a drop-down menu will appear. To cancel a drop-down menu, place the cursor anywhere outside the option and press the left button.

The three dots after an option term (...) on a drop-down menu, such as **Define Variable**...option in Data option, signifies that a dialog box will appear when this option is chosen. To cancel a dialog box, select the **Cancel** button in the dialog box. A right-facingarrowhead after an option term indicates that a further submenu will appear to the right of thedrop-down menu. An option with neither of these signs means that there are no further dropdownmenus to select. There are five standard command pushbuttons in most dialog boxes.

OK:It runs the procedure. After the variables and additional specifications are selected, clickOK to run the procedure.

Paste:It generates command syntax from the dialog box selections and pastes the syntax into a syntax window.

Reset:It deselects any variables in the selected variable list and resets all specifications in the dialog box.

Cancel:It cancels any changes in the dialog box settings since the last time it was opened and closes the dialog box.

Help:It contains information about the current dialog box.

5. Entering and Editing Data

The easiest way of entering data in SPSS is to type it directly into the matrix of columns and numbered rows in the **Data Editor** window. The columns represent variables and the rows represent cases. The variables can be defined in the variable view. Variable name must be no longer than eight characters and the name must begin with a letter.

Saving data

To be able to retrieve a file, the file must be saved with a proper name. The default extension name for saving files is **sav**. To save this file on a floppy disk, we carry out the following sequence:

> →**File** →**Save As...** [opens**Save Data As** dialog box]→box under **File Name:**delete the asterisk and type file name →**OK**

The output file can also be printed and saved. The extension name for output file is .**spo**.

Retrieving a saved file

To retrieve this file at a later stage when it is no longer the current file, use the following procedure:

> **File**→**Open**→**Data...**[opens the **Open Data File** dialog box] →choose drive from options listed →type name under **File Name:** →file name → **OK**

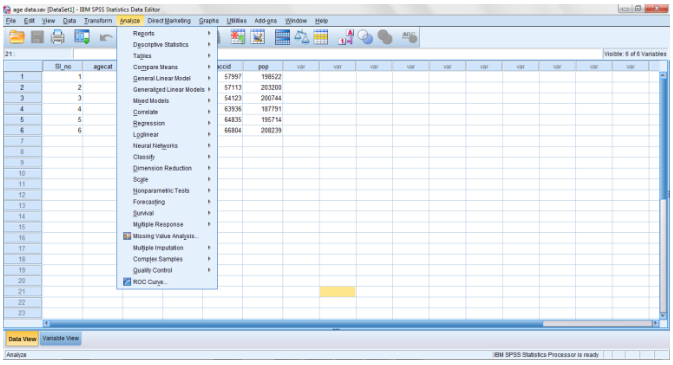
Basic Steps in Data Analysis

- **Get your data into SPSS**. You can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter your data directly in the Data Editor.
- **Select a procedure**. Select a procedure from the menus to calculate statistics or to create a chart.
- **Select the variables for the analysis**. The variables in the data file are displayed in a dialog box for the procedure.
- **Run the procedure**. Results are displayed in the Viewer.

6. Statistical Procedures

After entering the data set in **Data Editor** or reading an ASCII data file, we are now ready to analyse it. The **Analyse** option has the following sub options:

Reports, Descriptive Statistics, Tables, Compare means, General Linear model, Mixed Models, Correlate, Regression, Loglinear, Neural Networks, Classify, Dimension Reduction, Scale, Non parametric tests, Forecasting, Time Series, Survival, Multiple response, Missing value analysis, Multiple imputation, Complex samples, Quality control, ROC curve.



	Sl_no	agecat	ccid	pop
1	1		57997	198522
2	2		57113	203200
3	3		54123	200744
4	4		63936	187791
5	5		64835	195714
6	6		66804	208239

6.1Reports:

This submenu provides techniques for reporting the results. The various sub-sub menus under this are as follows:

Codebook reports the dictionary information such as variable names, variable labels, value labels, missing values and summary statistics for all or specified variables and multiple response sets in the active dataset. For nominal and ordinal variables and multiple response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation, and quartiles.

OLAP (Online Analytical Processing) **Cubes** procedure calculates totals, means, and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.

Case Summaries calculates subgroup statistics for variables within categories of one or more grouping variables. All levels of the grouping variable are cross tabulated. One can choose the order in which the statistics are displayed. Summary statistics for each variable across all categories are also displayed. With large datasets, one can choose to list only the first n cases.

Report Summaries in Rows produces reports in which different summary statistics are laid out in rows. Case listings are also available from this command, with or without summary statistics.

Report Summaries in Columns produces reports in which different summary statistics are laid out in separate columns.

6.2 Descriptive Statistics:

This submenu provides techniques for summarizing data with statistics, charts, and reports. The various sub-sub menus under this are as follows:

Frequencies provides information about the relative frequency of the occurrence of each category of a variable. This can be used it to obtain summary statistics that describe the typical value and the spread of the observations. To compute summary statistics for each of several groups of cases, Means procedure or the Explore procedure can be used.

Descriptivesis used to calculate statistics that summarize the values of a variable like the measures of central tendency, measures of dispersion, skewness, kurtosis etc.

Explore produces and displays summary statistics for all cases or separately for groups of cases. Boxplots, stem-and leaf plots, histograms, tests of normality, robust estimates of location, frequency tables and other descriptive statistics and plots can also be obtained.

Crosstabs is used to count the number of cases that have different combinations of values of two or more variables, and to calculate summary statistics and tests. The variables you use to form the categories within which the counts are obtained should have a limited number of distinct values.

P-P plots provides the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.

Q-Q plots provide the quantiles of a variable's distribution against the quantiles of the normal distribution.

6.3 Tables:

Custom Tables submenu provides attractive, flexible displays of frequency counts, percentages and other statistics.

6.4 Compare Means:

This submenu provides techniques for testing differences among two or more means for both independent and related samples.

Means computes summary statistics for a variable when the cases are subdivided into groups based on their values for other variables.

One-Sample tTest procedure tests whether the mean of a single variable differs from a specified constant. For each test variable: mean, standard deviation, and standard error of the mean.

Independent Sample t test is used if two unrelated samples come from populations with the same mean. The observations should be from two unrelated groups, and for testing, the mean must be an appropriate summary measure for the variable to be compared in the two groups. For more than two independent groups, the *One-way ANOVA* option could be used.

Paired Sample t test is used to compare the means of the same subjects in two conditions or at two points in time i.e. to compare subjects who had been matched to be similar in certain respects and then to test if two related samples come from populations with the same mean. The related, or paired, samples often result from an experiment in which the same person is observed before and after an intervention. If the distribution of the differences of the values between the members of a pair is markedly non-normal you should consider one of the nonparametric tests.

One-Way ANOVA is used to test that several independent groups come from populations with the same mean. To see which groups are significantly different from each other, multiple comparison procedures can be used through *Post Hoc Multiple Comparison option* which consist of the options like *Least-significant difference, Duncan's multiple range test, Scheffe*etc. The contrast analysis can also be performed in order to compare the different groups or treatments by using the *Contrast* option. The data obtained using completely randomised design can be analysed through this option.

6.5 General Linear Model

This submenu provides techniques for testing univariateand multivariate Analysis-of-Variance models, including repeated measures.

Univariatesub-option could be used to analyse the experimental designs like Completely randomised design, Randomised block design, Latin square design, Designs for factorial experiments etc. The covariance analysis can also be performed and alternate methods for partitioning sums of squares can be selected. If only some of the interactions of a particular order are to be included, the *Custom* procedure should be used. If there is only one factor then One-Way ANOVA procedure should be used.

Multivariate analyses analysis-of-variance and analysis-of-covariance designs when you have two or more correlated dependent variables. Multivariate analysis of variance is used to test hypotheses about the relationship between a set of interrelated dependent variables and one or more factor or grouping variables. For example, you can test whether verbal and mathematical test scores are related to instructional method used, sex of the subject, and the interaction of method and sex. This procedure should be used only if there are several dependent variables which are related to each other. For a single dependent variable or unrelated dependent variables, the Univariate ANOVA procedures can be adopted. If the same dependent variable is measured on several occasions for each subject, the Repeated Measures procedure is to be used.

Repeated Measures is used to test hypotheses about the means of a dependent variable when the same dependent variable is measured on more than one occasion for each subject. Subjects can also be classified into mutually exclusive groups, such as males or females, or type of job held. Then you can test hypotheses about the effects of the between-subject variables and the within-subject variables, as well as their interactions.

6.6 Correlate

This submenu provides measures of association for two or more variables measured at the interval level.

Bivariate calculates matrices of Pearson product-moment correlations, and of Kendall and Spearman nonparametric correlations, with significance levels and optional univariate statistics. The correlation coefficient is used to quantify the strength of the linear relationship between two variables. The *Pearson correlation coefficient* should be used only for data measured at the interval or ratio level. Spearman and Kendall correlation coefficients are nonparametric measures which are particularly useful when the data contain outliers or when the distribution of the variables is markedly non-normal. Both the Spearman and Kendall coefficients are based on assigning ranks to the variables.

Partial calculates *partial correlation coefficients* that describe the relationship between two variables, while adjusting for the effects of one or more additional variables. If the value of a dependent variable from a set of independent variables is to be predicted then the Linear Regression procedure may be used. If there are no control variables then the Bivariate Correlations procedure can be adopted. Nominal variables should not be used in the partial correlation procedure.

Distances calculates statistics measuring either similarities or dissimilarities (distances), either between pairs of variables or between pairs of cases. These similarity or distance measures can then be used with other procedures, such as factor analysis, cluster analysis, or multidimensional scaling, to help analyze complex datasets. Dissimilarity (distance) measures for interval data are Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, or customized; for count data, chi-square or phi-square; for binary data, Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, shape, or Lance and Williams. Similarity measures for interval data are Pearson correlation or cosine; for binary data, Russel and Rao, simple matching, Jaccard, etc.

6.7 Regression

This submenu provides a variety of regression techniques, including linear, logistic, nonlinear, weighted, and two-stage least-squares regression.

Linear is used to examine the relationship between a dependent variable and a set of independent variables. If the dependent variable is dichotomous, then the logistic regression procedure should be used. If the dependent variable is censored, such as survival time after surgery, use the Life Tables, Kaplan-Meier, or proportional hazards procedure.

Curve Estimation produces curve estimation regression statistics and related plots for 11 different curve estimation regression models. A separate model is produced for each dependent variable. One can also save predicted values, residuals, and prediction intervals as new variables.

Logistic estimates regression models in which the dependent variable is dichotomous. If the dependent variable has more than two categories, use the Discriminant procedure to identify variables which are useful for assigning the cases to the various groups. If the dependent variable is continuous, use the Linear Regression procedure to predict the values of the dependent variable from a set of independent variables. In recent versions there are two options **Binary Logistic** as well as **Multinomial Logistic.**

Probit performs probit analysis which is used to measure the relationship between a response proportion and the strength of a stimulus. For example, the probit procedure can be used to examine the relationship between the proportion of plants dying and the strength of the pesticide applied or to examine the relationship between the proportion of people buying a product and the magnitude of the incentive offered. The Probit procedure should be used only if the response is dichotomousbuy/not buy, alive/dead--and several groups of subjects are exposed to different levels of some stimulus. For each stimulus level, the data must contain counts of the totals exposed and the totals responding. If the response variable is dichotomous but you do not have groups of subjects with the same values for the independent variables you should use the Logistic Regression procedure.

Nonlinear estimates nonlinear regression models, including models in which parameters are constrained. The nonlinear regression procedure can be used if one knows the equation whose parameters are to be estimated, and the equation cannot be written as the sum of parameters times some function of the independent variables. In nonlinear regression the parameter estimates are obtained iteratively. If the function is linear, or can be transformed to a linear function, then the Linear Regression procedure should be used.

Weight Estimation estimates a linear regression model with differential weights representing the precision of observations. This command is in the Professional Statistics option. If the variance of the dependent variable is not constant for all of the values of the independent variable, weights which are inversely proportional to the variance of the dependent variable can be incorporated into the analysis. This results in a better solution. The Weight Estimation procedure can also be used to estimate the weights when the variance of the dependent variable is related to the values of an independent variable. If you know the weights for each case you can use the linear regression procedure to obtain a weighted least squares solution. The linear regression procedure provides a large number of diagnostic statistics which help you evaluate how well the model fits your data.

2-Stage Least Squares performs two-stage least squares regression for models in which the error term is related to the predictors. This command is in the Professional Statistics option. For example, if you want to model the demand for a product as a function of price, advertising expenses, cost of the materials, and some economic indicators, you may find that the error term of the model is correlated with one or more of the independent variables. Two-stage least squares allows you to estimate such a model.

The **Loglinear** submenu provides general and hierarchical log-linear analysis and logit analysis.

6.8 Classify

This submenu provides cluster and discriminant analysis.

Two Step Cluster performs Two Step Cluster Analysis procedure which is an exploratory data analysis tool designed to reveal natural clustering within a dataset that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques. The Log-likelihood and Euclidean Distance Measures are used as the similarity measure between two clusters.

K-means Cluster performs cluster analysis using an algorithm that can handle large numbers of cases, but that requires you to specify the number of clusters. The goal of cluster analysis is to identify relatively homogeneous groups of cases based on selected characteristics. If the number of clusters to be formed is not known, then Hierarchical Cluster procedure can be used. If the observations are in known groups and one wants to predict group membership based on a set of independent variables, then the Discriminant procedure can be used.

Hierarchical Cluster combines cases into clusters hierarchically, using a memory-intensive algorithm that allows you to examine many different solutions easily.

Discriminant is used to classify cases into one of several known groups on the basis of various characteristics. To use the Discriminant procedure the dependent variable must have a limited number of distinct categories. Independent variables that are nominal must be recoded to dummy or contrast variables. If the dependent variable has two categories, Logistic Regression can be used. If the dependent variable is continuous one may use Linear Regression.

Nearest Neighbor performs Nearest Neighbor Analysis for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

6.9 Dimension Reduction

This submenu provides factor analysis, correspondence analysis, and optimal scaling.

Factor is used to identify factors that explain the correlations among a set of variables. Factor analysis is often used to summarize a large number of variables with a smaller number of derived variables, called factors.

Correspondence Analysis analyzes correspondence tables (such as cross-tabulations) to best measure the distances between categories or between variables. This command is in the Categories option.

Distances computes many different measures of similarity, dissimilarity or distance. Many different measures can be used to quantify how much alike or how different two cases or variables are. Similarity measures are constructed so that large values indicate much similarity and small values indicate little similarity. Dissimilarity measures estimate the distance or unlikeness of two cases. A large dissimilarity value tells that two cases or variables are far apart. In order to decide which similarity or dissimilarity measure to use, one must consider characteristics of the data. Special measures are available for interval data, frequency counts, and binary data. If the cases are to be classified into groups based on similarity or dissimilarity measures, one of the Cluster procedures should be used.

6.10 Scale

This submenu provides reliability analysis and multidimensional scaling.

Reliability analysis allows to study the properties of measurement scales and the items that compose the scales. The Reliability Analysis procedure calculates a number of commonly used measures of scale reliability and also provides information about the relationships between individual items in the scale. This provides several statistics like descriptives for each variable and for the scale, summary statistics across items, inter-item correlations and covariances, reliability estimates, ANOVA table, intraclass correlation coefficients, Hotelling's T2, and Tukey's test of additivity.

6.11 Nonparametric Tests:

This submenu provides nonparametric tests for one sample, or for two and more paired or independent samples. Legacy dialogs sub-submenu consists following tests

Chi-Square is used to test hypotheses about the relative proportion of cases falling into several mutually exclusive groups. For example, if one wants to test the hypotheses that people are equally likely to buy six different brands of cereals, one can count the number buying each of the six brands. Based on the six observed counts Chi-Square procedure could be used to test the hypothesis that all six cereals are equally likely to be bought. The expected proportions in each of the categories don't have to be equal. The hypothetical proportions to be tested should be specified.

Binomial is used to test the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values. For example, to test that the probability of an item on the assembly line is defective is one out of ten (p=0.1), take a sample of 300 items and record whether each is defective or not. Then use the binomial procedure to test the hypothesis of interest.

Runs is used to test whether the two values of a dichotomous variable occur in a random sequence. The runs test is appropriate only when the order of cases in the data file is meaningful.

1-Sample K-S is used to compare the observed frequencies of the values of an ordinal variable, such as rated quality of work, against some specified theoretical distribution. It determines the statistical significance of the largest difference between them. In SPSS, the theoretical distribution can be **Normal, Uniform or Poisson**. Alternative tests for normality are available in the Explore procedure, in the Summarize submenu. The P-P and Q-Q plots in the Graphs menu can also be used to examine the assumption of normality.

2-Independent Samples is used to compare the distribution of a variable between two non-related groups. Only limited assumptions are needed about the distributions from which the sample are selected. The Mann-Whitney U test is an alternative to the two sample t-test. The actual values of the data are replaced by ranks. The Kolmogorov-Smirnov test is based on the differences between the observed cumulative distributions of the two groups. The Wald-Woflowitz runs tests sorts the data values from smallest to largest and then performs a runs test on the group’s numbers. The Moses Test of Extreme Reaction is used to test for differences in range between two groups.

K-Independent Samples is used to compare the distribution of a variable between two or more groups. Only limited assumptions are needed about the distributions from which the samples are selected. The Kruskal-Wallis test is an alternative to one-way analysis of variance, with the actual values of the data replaced by ranks. The Median tests counts the number of cases in each group that are above and below the combined median, and then performs a chi-square test.

2 Related Samples is used to compare the distribution of two related variables. Only limited assumptions are needed about the distributions from which the samples are selected. The Wilcoxon and Sign tests are nonparametric alternative to the paired samples t-test. The Wilcoxon test is more powerful than the Sign test. *McNemar's test*is used to determine changes in proportions for related samples. It is often used for "before and after" experimental designs when the dependent variable is dichotomous. For example, the effect of a campaign speech can be tested by analysing the number of people whose preference for a candidate changed based on the speech. Using McNemar's test you analyse the changes to see if change in both directions is equally likely.

K Related Samples is used to compare the distribution of two or more related variables. Only limited assumptions are needed about the distributions from which the samples are selected. *The Friedman test* is a nonparametric alternative to a single-factor repeated measures analysis of variance. You can use it when the same measurement is obtained on several occasions for a subject. For example, the Friedman test can be used to compare consumer satisfaction of 5 products when each person is asked to rate each of the products on a scale. *Cochran's Q test*can be used to test whether several dichotomous variables have the same mean. For example, if instead of asking each subject to rate their satisfaction with five products, you asked them for a yes/no response about each, you could use Cochran's test to test the hypothesis that all five products have the same proportion of satisfied users. *Kendall's W measures*the agreement among raters. Each of your cases corresponds to a rater, each of the selected variables is an item being rated. For example, if you ask a sample of customers to rank 7 ice-cream flavours from least to most liked, you can use Kendall's W to see how closely the customers agree in their ratings.

6.12 Forecasting

This submenu provides create models, seasonal decomposition, spectral analysis, autocorrelations, cross-correlations etc.

Autocorrelations calculates and plots the autocorrelation function (ACF) and partial autocorrelation function of one or more series to any specified number of lags, displaying the Box-Ljung statistic at each lag to test the overall hypothesis that the ACF is zero at all lags.

Cross-correlations calculates and plots the cross-correlation function of two or more series for positive, negative, and zero lags.

Spectral analysis calculates and plots univariate or bivariate periodograms and spectral density functions, which express variation in a time series (or covariation in two time series) as the sum of a series of sinusoidal components. It can optionally save various components of the frequency analysis as new series.

6.13 Survival:

The submenu provides techniques for analyzing the time for some terminal event to occur, including Kaplan-Meier analysis and Cox regression.

6.14Multiple Response:

This submenu provides facilities to define and analyze multiple-response or multiple-dichotomy sets.

Quality Control submenu provides facilities to for obtaining control charts and Pareto charts.

Complex Samples submenu provides procedures for Sampling from Complex Designs. The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

Other than this **Analyze** menu there are several other important menus available in SPSS.

6.15 Transform

Compute calculates the values for either a new or an existing variable, for all cases or for cases satisfying a logical criterion.

Random Number Seed sets the seed used by the pseudo-random number generator to a specific value, so that you can reproduce a sequence of pseudo-random numbers.

Count creates a variable that counts the occurrences of the same value(s) in a list of variables for each case.

Recode into Same Variables reassigns the values of existing variables or collapses ranges of existing values into new values.

Recode into Different Variables reassigns the values of existing variables to new variables or collapses ranges of existing values into new variables.

Rank Cases creates new variables containing ranks, normal scores, or similar ranking scores for numeric variables.

Automatic Recode reassigns the values of existing variables to consecutive integers in new variables.

Create Time Series creates a time-series variable as a function of an existing series, for example, lagged or leading values, differences, cumulative sums. This command is in the Trends option.

Replace Missing Values substitutes non-missing values for missing values, using the series mean or one of several time-series functions. This command is in the Trends option.

Run Pending Transforms executes transformation commands that are pending due to the Transformation Options setting in the Preferences dialog.

6.16 Utilities

Command Index take you to the dialog box for a command if you know its name in the SPSS command language.

Fonts lets you choose a font, style, and size for SPSS Data Editor, output, and syntax windows.

Variable Information displays the Variables window, which shows information about the variables in your working data file, and allows you to scroll the data editor to a specific variable, or copy variable names to the designated syntax window.

File Information displays information about the working data file in the output window.

Output Page Titles lets you specify a title and subtitle for output from SPSS. They appear in the page header, if it is displayed. (Preferences in the Edit menu controls the page header.)

Define Sets defines sets of variables for use in other dialog boxes.

Use Sets lets you select which defined sets of variables should appear in the source-variable lists of other dialog boxes.

Grid Lines turns grid lines on and off in the Data Editor window. This command is available when the Data Editor is active.

Value Labels turns on and off the display of Value Labels (instead of actual values) in the Data Editor window. When Value Labels are displayed you can edit data with a pop-up menu of labels. This command is available when the Data Editor is active.

Auto New Case turns on and off the automatic creation of new cases by cursor movement below the last case in the Data Editor window. This command is available when the Data Editor is active.

Designate Window designates the active window to receive output from SPSS commands (if it is an output window); or to receive commands pasted from dialog boxes (if it is a syntax window). You can also designate a window by clicking the !button on its icon bar. This command is available when an output or syntax window is active.

6.17 Graphs

The Chart Builder available in Graph menu allows to build charts from predefined gallery charts or from the individual parts (for example, axes and bars). You build a chart by dragging and dropping the gallery charts or basic elements onto the canvas, which is the large area to the right of the Variables list in the Chart Builder dialog box.

Legacy Dialogs submenu provides following graph submenus

Bar generates a simple, clustered, or stacked bar chart of the data.

3-D Bar Charts allows to generate bar graph in 3-dimensional axis.

Line generates a simple or multiple line chart of the data.

Area generates a simple or stacked area chart of the data.

Pie generates a simple pie chart or a composite bar chart from the data.

High-Low plots pairs or triples of values, for example high, low, and closing prices.

Boxplot generates boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.

Error Bar Charts plot the confidence intervals, standard errors, or standard deviations of individual variables.

Scatter/dot generates a simple or overlay scatter plot, a scatter plot matrix, or a 3-D scatter plot from the data.

Histogram generates a histogram showing the distribution of an individual variable.

Practical exercise using SPSS.

Exercise 1: The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

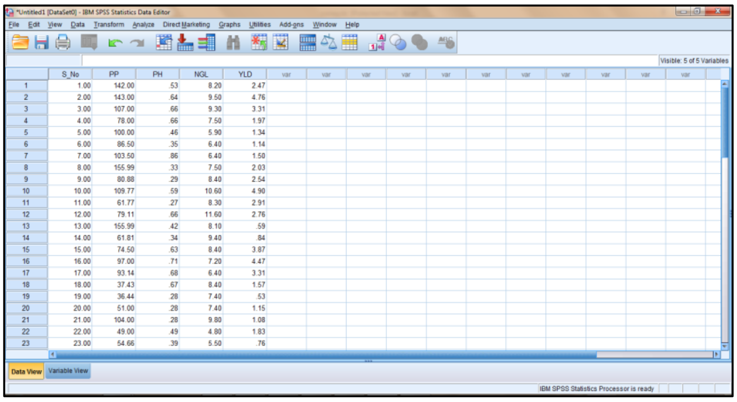
S.No.	PP	PH	NGL	Yield	S.No.	PP	PH	NGL	Yield
1	142.00	0.525	8.2	2.470	24	55.55	0.265	5.0	0.430
2	143.00	0.640	9.5	4.760	25	88.44	0.980	5.0	4.080
3	107.00	0.660	9.3	3.310	26	99.55	0.645	9.6	2.830
4	78.00	0.660	7.5	1.970	27	63.99	0.635	5.6	2.570

5	100.00	0.460	5.9	1.340	28	101.77	0.290	8.2	7.420
6	86.50	0.345	6.4	1.140	29	138.66	0.720	9.9	2.620
7	103.50	0.860	6.4	1.500	30	90.22	0.630	8.4	2.000
8	155.99	0.330	7.5	2.030	31	76.92	1.250	7.3	1.990
9	80.88	0.285	8.4	2.540	32	126.22	0.580	6.9	1.360
10	109.77	0.590	10.6	4.900	33	80.36	0.605	6.8	0.680
11	61.77	0.265	8.3	2.910	34	150.23	1.190	8.8	5.360
12	79.11	0.660	11.6	2.760	35	56.50	0.355	9.7	2.120
13	155.99	0.420	8.1	0.590	36	136.00	0.590	10.2	4.160
14	61.81	0.340	9.4	0.840	37	144.50	0.610	9.8	3.120
15	74.50	0.630	8.4	3.870	38	157.33	0.605	8.8	2.070
16	97.00	0.705	7.2	4.470	39	91.99	0.380	7.7	1.170
17	93.14	0.680	6.4	3.310	40	121.50	0.550	7.7	3.620
18	37.43	0.665	8.4	1.570	41	64.50	0.320	5.7	0.670
19	36.44	0.275	7.4	0.530	42	116.00	0.455	6.8	3.050
20	51.00	0.280	7.4	1.150	43	77.50	0.720	11.8	1.700
21	104.00	0.280	9.8	1.080	44	70.43	0.625	10.0	1.550
22	49.00	0.490	4.8	1.830	45	133.77	0.535	9.3	3.280
23	54.66	0.385	5.5	0.760	46	89.99	0.490	9.8	2.690

Source: Design Resources Server. Indian Agricultural Statistics Research Institute(ICAR), New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on <05-05-2015>).

1. Find mean, standard deviation, minimum and maximum values of all the characters.
2. Find correlation coefficient between each pair of the variables.
3. Give a scatter plot of the variable PP with dependent variable yield.
4. Fit a multiple linear regression equation where yield is dependent variable whereas all other characters as independent variables.

At first enter the entire data in the data editor as given below,

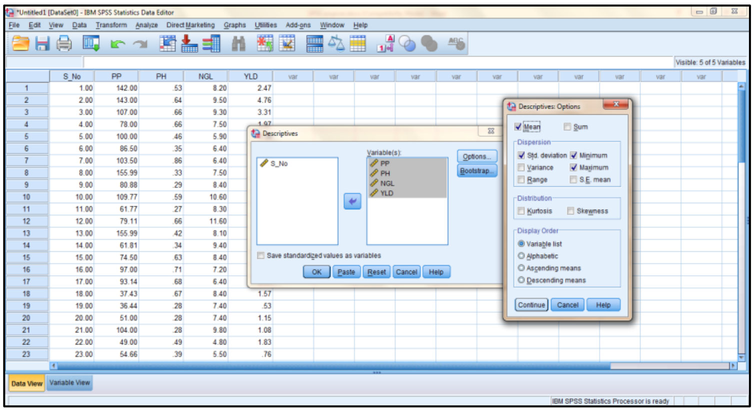


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	4.76
3	3.00	107.00	.66	9.30	3.31
4	4.00	78.00	.66	7.50	1.97
5	5.00	100.00	.46	5.90	1.34
6	6.00	86.50	.35	6.40	1.14
7	7.00	103.50	.86	6.40	1.50
8	8.00	155.99	.33	7.50	2.03
9	9.00	80.88	.29	8.40	2.54
10	10.00	109.77	.59	10.60	4.90
11	11.00	61.77	.27	8.30	2.91
12	12.00	79.11	.66	11.60	2.76
13	13.00	155.99	.42	8.10	.59
14	14.00	61.81	.34	9.40	.84
15	15.00	74.50	.63	8.40	3.87
16	16.00	97.00	.71	7.20	4.47
17	17.00	93.14	.68	6.40	3.31
18	18.00	37.43	.67	8.40	1.57
19	19.00	36.44	.28	7.40	.53
20	20.00	51.00	.28	7.40	1.15
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

There are several ways to answer Q no. 1 in SPSS. Commands following first way is as follows,

Analyze → Descriptive Statistics → Descriptives…→ Put PP, PH, NGL, YLD in the variables list→ Choose appropriate options from Options tab→PressContinue→Ok



Descriptives

S_No

Variable(s):
PP
PH
NGL
YLD

□ Save standardized values as variables

Descriptives: Options

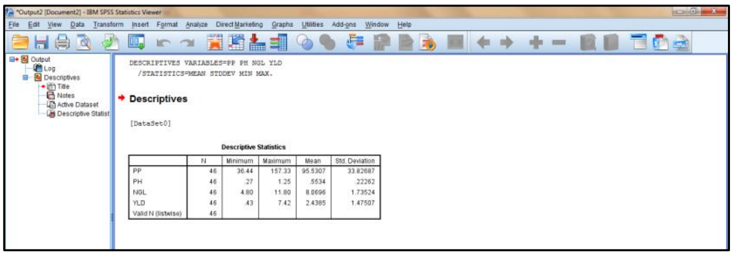
☑ Mean □ Sum

Dispersion
☑ Std. deviation ☑ Minimum
□ Variance ☑ Maximum
□ Range □ S.E. mean

Distribution
□ Kurtosis □ Skewness

Display Order
● Variable list
○ Alphabetic
○ Ascending means
○ Descending means

Output:



```
DESCRIPTIVES VARIABLES=PP PH NGL YLD
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

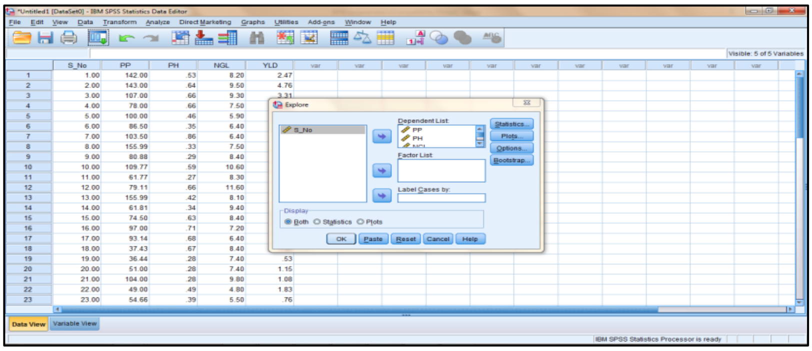
[DataSet0]

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PP	46	36.44	157.33	95.5307	33.82687
PH	46	.27	1.25	.5534	.22262
NGL	46	4.80	11.80	8.0696	1.73524
YLD	46	.43	7.42	2.4385	1.47507
Valid N (listwise)	46				

Another way:

Analyze → Descriptive Statistics → Explore…→ Put PP, PH, NGL, YLD in the Dependent list→ Choose both Statistics and plot→Press Ok



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 5 of 5 Variables
var
Data View
Variable View
IBM SPSS Statistics Processor is ready

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	4.76
3	3.00	107.00	.66	9.30	3.31
4	4.00	78.00	.66	7.50	
5	5.00	100.00	.46	5.90	
6	6.00	86.50	.35	6.40	
7	7.00	103.50	.86	6.40	
8	8.00	155.99	.33	7.50	
9	9.00	80.88	.29	8.40	
10	10.00	109.77	.59	10.60	
11	11.00	61.77	.27	8.30	
12	12.00	79.11	.66	11.60	
13	13.00	155.99	.42	8.10	
14	14.00	61.81	.34	9.40	
15	15.00	74.50	.63	8.40	
16	16.00	97.00	.71	7.20	
17	17.00	93.14	.68	6.40	
18	18.00	37.43	.67	8.40	
19	19.00	36.44	.28	7.40	.53
20	20.00	51.00	.28	7.40	1.15
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

Explore

S_No

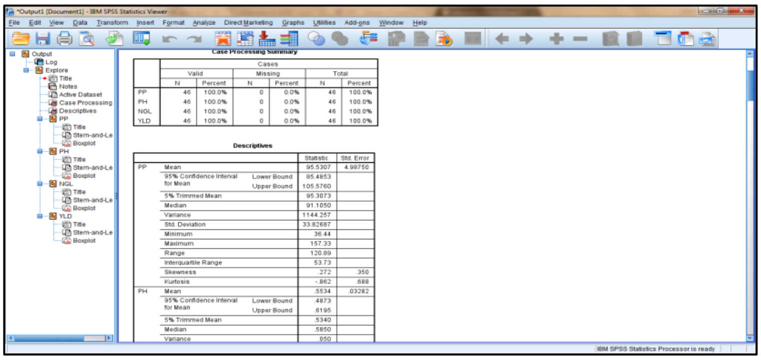
Dependent List:
PP
PH
NGL

Factor List:

Label Cases by:

Display
● Both ○ Statistics ○ Plots

Output:



*Output1 [Document1] - IBM SPSS Statistics Viewer

Case Processing Summary

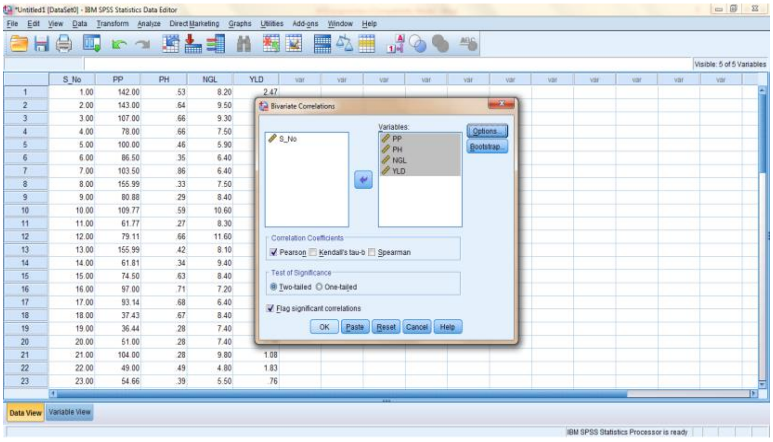
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
PP	46	100.0%	0	0.0%	46	100.0%
PH	46	100.0%	0	0.0%	46	100.0%
NGL	46	100.0%	0	0.0%	46	100.0%
YLD	46	100.0%	0	0.0%	46	100.0%

Descriptives

			Statistic	Std. Error
PP	Mean		95.5307	4.98750
	95% Confidence Interval for Mean	Lower Bound	85.4853	
		Upper Bound	105.5760	
	5% Trimmed Mean		95.3073	
	Median		91.1050	
	Variance		1144.257	
	Std. Deviation		33.82687	
	Minimum		36.44	
	Maximum		157.33	
	Range		120.89	
	Interquartile Range		53.73	
	Skewness		.272	.350
	Kurtosis		-.862	.688
PH	Mean		.5534	.03282
	95% Confidence Interval for Mean	Lower Bound	.4873	
		Upper Bound	.6195	
	5% Trimmed Mean		.5340	
	Median		.5850	
	Variance		.050	

To answer Q no 2 follow the following steps

Analyze → Correlate → Bivariate→ Put PP, PH, NGL, YLD in the Valiables list→ Choose Pearson’s correlation coefficient→Press Ok



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 5 of 5 Variables
var
Data View
Variable View
IBM SPSS Statistics Processor is ready

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	
3	3.00	107.00	.66	9.30	
4	4.00	78.00	.66	7.50	
5	5.00	100.00	.46	5.90	
6	6.00	86.50	.35	6.40	
7	7.00	103.50	.86	6.40	
8	8.00	155.99	.33	7.50	
9	9.00	80.88	.29	8.40	
10	10.00	109.77	.59	10.60	
11	11.00	61.77	.27	8.30	
12	12.00	79.11	.66	11.60	
13	13.00	155.99	.42	8.10	
14	14.00	61.81	.34	9.40	
15	15.00	74.50	.63	8.40	
16	16.00	97.00	.71	7.20	
17	17.00	93.14	.68	6.40	
18	18.00	37.43	.67	8.40	
19	19.00	36.44	.28	7.40	
20	20.00	51.00	.28	7.40	
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

Bivariate Correlations

S_No

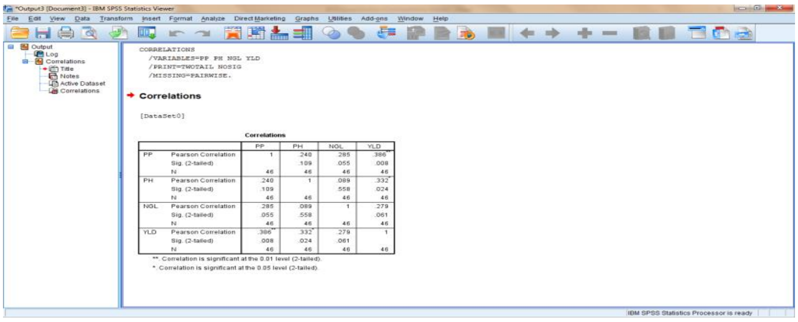
Variables:
PP
PH
NGL
YLD

Correlation Coefficients
☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Output:



*Output3 [Document3] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Correlations
Title
Notes
Active Dataset
Correlations
IBM SPSS Statistics Processor is ready

```
CORRELATIONS
  /VARIABLES=PP PH NGL YLD
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

```
[DataSet0]
```

Correlations

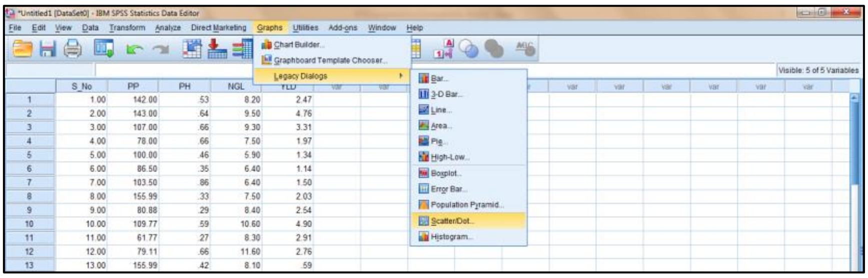
		PP	PH	NGL	YLD
PP	Pearson Correlation	1	.240	.285	.386**
	Sig. (2-tailed)		.109	.055	.008
	N	46	46	46	46
PH	Pearson Correlation	.240	1	.089	.332*
	Sig. (2-tailed)	.109		.558	.024
	N	46	46	46	46
NGL	Pearson Correlation	.285	.089	1	.279
	Sig. (2-tailed)	.055	.558		.061
	N	46	46	46	46
YLD	Pearson Correlation	.386**	.332*	.279	1
	Sig. (2-tailed)	.008	.024	.061	
	N	46	46	46	46

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

To give the scatter plot of the variable PP with dependent variable yield use following steps:

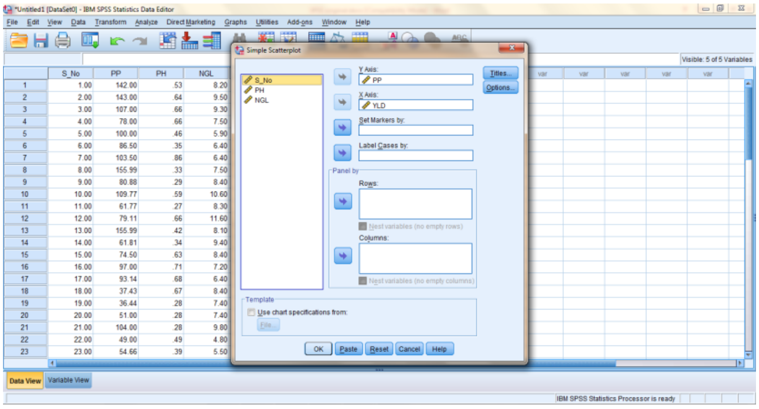
Graphs → Legacy dialogs→ Scatterplot→ Put PP at Y axis and YLD at X axis→ Press Ok



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Chart Builder...
Graphboard Template Chooser...
Legacy Dialogs
Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...
Visible: 5 of 5 Variables

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	4.76
3	3.00	107.00	.66	9.30	3.31
4	4.00	78.00	.66	7.50	1.97
5	5.00	100.00	.46	5.90	1.34
6	6.00	86.50	.35	6.40	1.14
7	7.00	103.50	.86	6.40	1.50
8	8.00	155.99	.33	7.50	2.03
9	9.00	80.88	.29	8.40	2.54
10	10.00	109.77	.59	10.60	4.90
11	11.00	61.77	.27	8.30	2.91
12	12.00	79.11	.66	11.60	2.76
13	13.00	155.99	.42	8.10	.59



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

Simple Scatterplot

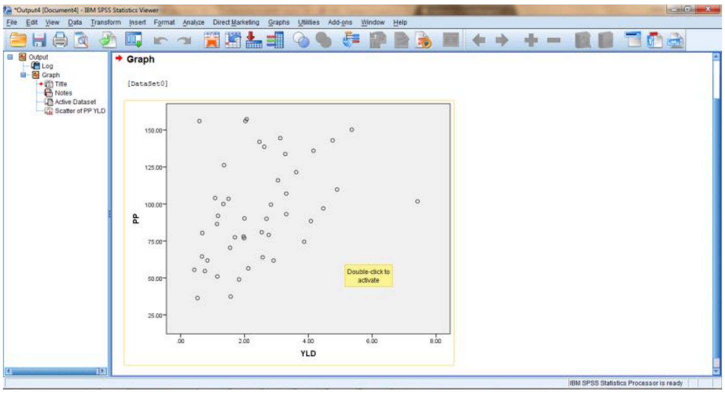
Y Axis: PP

X Axis: YLD

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
S_No
PH
NGL
Set Markers by:
Label Cases by:
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
Titles...
Options...
OK
Paste
Reset
Cancel
Help
Visible: 5 of 5 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

	S_No	PP	PH	NGL
1	1.00	142.00	.53	8.20
2	2.00	143.00	.64	9.50
3	3.00	107.00	.66	9.30
4	4.00	78.00	.66	7.50
5	5.00	100.00	.46	5.90
6	6.00	86.50	.35	6.40
7	7.00	103.50	.86	6.40
8	8.00	155.99	.33	7.50
9	9.00	80.88	.29	8.40
10	10.00	109.77	.59	10.60
11	11.00	61.77	.27	8.30
12	12.00	79.11	.66	11.60
13	13.00	155.99	.42	8.10
14	14.00	61.81	.34	9.40
15	15.00	74.50	.63	8.40
16	16.00	97.00	.71	7.20
17	17.00	93.14	.68	6.40
18	18.00	37.43	.67	8.40
19	19.00	36.44	.28	7.40
20	20.00	51.00	.28	7.40
21	21.00	104.00	.28	9.80
22	22.00	49.00	.49	4.80
23	23.00	54.66	.39	5.50

Output:



*Output4 [Document4] - IBM SPSS Statistics Viewer

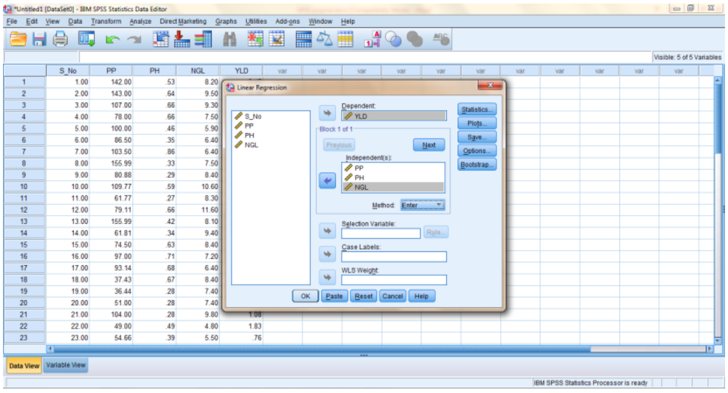
Graph

[DataSet0]

File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Graph
Title
Notes
Active Dataset
Scatter of PP YLD
150.00
125.00
100.00
75.00
50.00
25.00
PP
.00
2.00
4.00
6.00
8.00
YLD
Double-click to activate
IBM SPSS Statistics Processor is ready

To fit a multiple linear regression equation taking yield as dependent variable and all other characters as independent variables perform following steps

Analyze → Regression → Linear → Put Yld in Dependent variable and PP, PH, NGL in independent variable list → Press Ok



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	
2	2.00	143.00	.64	9.50	
3	3.00	107.00	.66	9.30	
4	4.00	78.00	.66	7.50	
5	5.00	100.00	.46	5.90	
6	6.00	86.50	.35	6.40	
7	7.00	103.50	.86	6.40	
8	8.00	155.99	.33	7.50	
9	9.00	80.88	.29	8.40	
10	10.00	109.77	.59	10.60	
11	11.00	61.77	.27	8.30	
12	12.00	79.11	.66	11.60	
13	13.00	155.99	.42	8.10	
14	14.00	61.81	.34	9.40	
15	15.00	74.50	.63	8.40	
16	16.00	97.00	.71	7.20	
17	17.00	93.14	.68	6.40	
18	18.00	37.43	.67	8.40	
19	19.00	36.44	.28	7.40	
20	20.00	51.00	.28	7.40	
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

Linear Regression

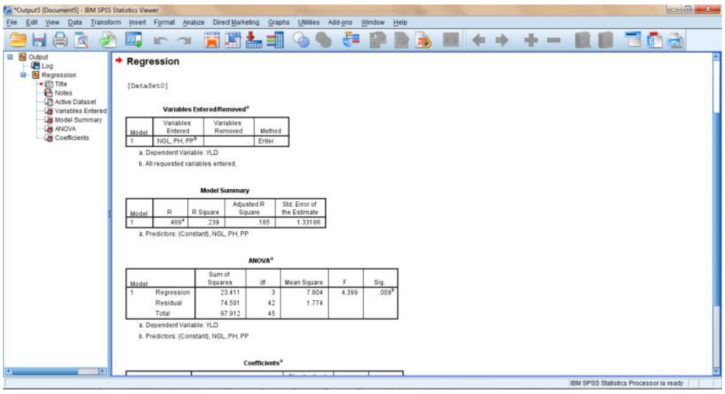
Dependent: YLD

Block 1 of 1

Independent(s): PP, PH, NGL

Method: Enter

Output:



*Output5 [Document5] - IBM SPSS Statistics Viewer

Regression

[DataSet0]

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	NGL, PH, PP[b]	.	Enter

a. Dependent Variable: YLD

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.489[a]	.239	.185	1.33186

a. Predictors: (Constant), NGL, PH, PP

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.411	3	7.804	4.399	.009[b]
	Residual	74.501	42	1.774		
	Total	97.912	45			

a. Dependent Variable: YLD

b. Predictors: (Constant), NGL, PH, PP

Coefficients[a]

Exercise 2. An experiment was conducted to study the hybrid seed production of ottle gourd under open field conditions. The main aim of the investigation was to compare natural pollination. The pollination is performed at noon (1-3pm)} under field conditions. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination. The data were collected on number of fruit set for the period of 45 days, fruit weight (kg), seed yield per plant (g) and seedling length (cm). The data obtained is as given below:

Group	**No. of fruit Set(45days)**	**Fruit weight (kg)**	**Seed yield/plant (g)**	**Seedling length (cm)**
1	**8**	**2.0**	**148.6**	**17.0**
1	**7**	**1.9**	**137.7**	**16.9**
1	**6**	**1.8**	**150.9**	**16.4**
1	**8**	**1.9**	**173.4**	**18.4**
1	**7**	**1.8**	**145.3**	**18.0**
1	**8**	**1.9**	**139.1**	**17.1**
1	**7**	**1.9**	**151.5**	**18.3**
1	**7**	**1.8**	**141.8**	**19.0**
1	**6**	**1.9**	**141.4**	**18.5**
1	**7**	**1.9**	**139.2**	**18.7**
2	**6.3**	**2.6**	**225.6**	**18.3**
2	**6.7**	**2.8**	**198.7**	**18.2**

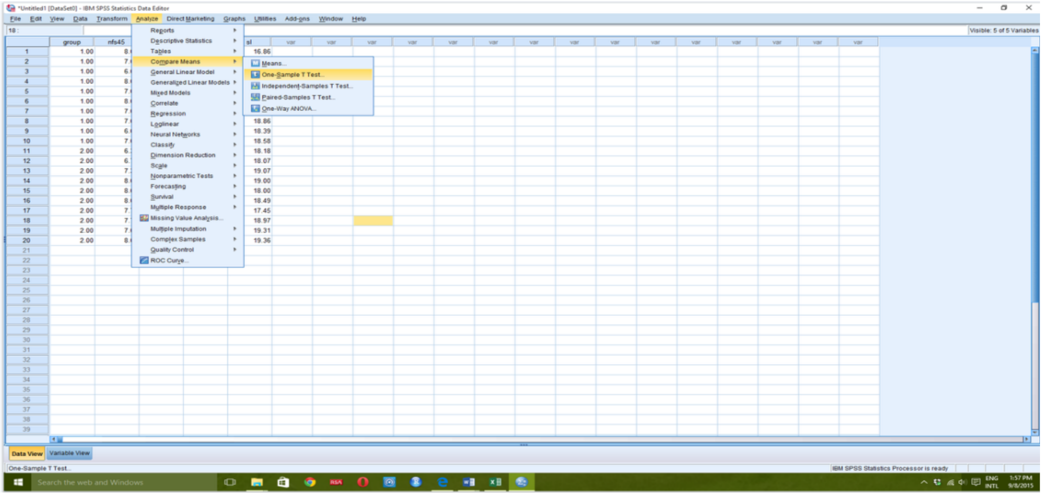
2	7.3	2.6	231.7	19.2
2	8	2.6	218.4	19.1
2	8	2.7	235.2	18.1
2	8	2.6	217.8	18.6
2	7.7	2.4	213.2	17.6
2	7.7	2.7	211.6	19.1
2	7	2.5	201.1	19.4
2	8	2.5	215.6	19.5

1. Test whether the mean of the population of Seed yield/plant (g) is 200 or not.

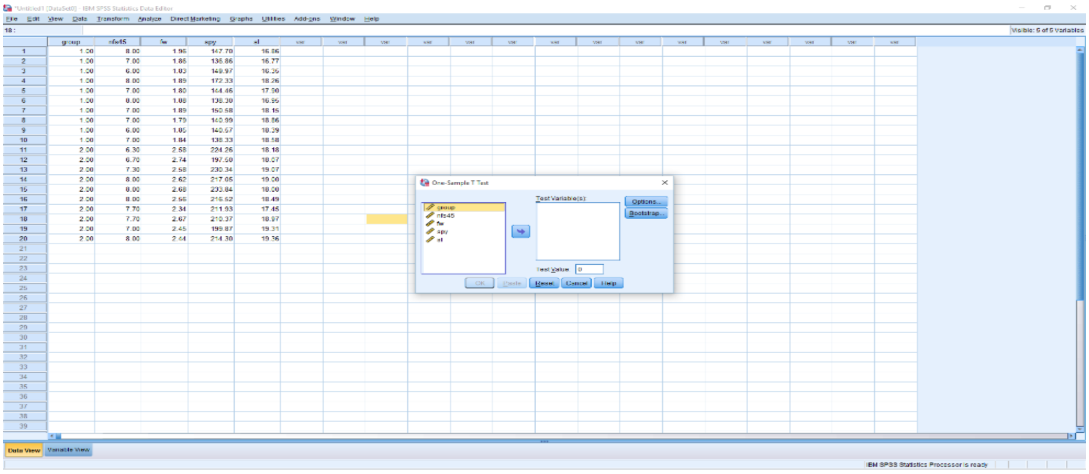
2. Test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different.

Test Procedure in SPSS

1. To test whether the mean of the population of Seed yield/plant (g) is 200 or not use the following steps. Select **Analyze → Compare Means → One-Sample T Test**



This selection displays the following screen


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 5 of 5 Variables
One-Sample T Test
Test Variable(s):
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

	group	nfs45	fw	spy	el
1	1.00	8.00	1.96	147.70	16.86
2	1.00	7.00	1.86	136.86	16.77
3	1.00	6.00	1.83	149.97	16.35
4	1.00	8.00	1.89	172.33	18.26
5	1.00	7.00	1.80	144.46	17.90
6	1.00	8.00	1.88	138.30	16.95
7	1.00	7.00	1.89	150.58	18.15
8	1.00	7.00	1.79	140.99	18.86
9	1.00	6.00	1.85	140.57	18.39
10	1.00	7.00	1.84	138.33	18.58
11	2.00	6.30	2.58	224.26	18.18
12	2.00	6.70	2.74	197.50	18.07
13	2.00	7.30	2.58	230.34	19.07
14	2.00	8.00	2.62	217.05	19.00
15	2.00	8.00	2.68	233.84	18.00
16	2.00	8.00	2.56	216.52	18.49
17	2.00	7.70	2.34	211.93	17.45
18	2.00	7.70	2.67	210.37	18.97
19	2.00	7.00	2.45	199.87	19.31
20	2.00	8.00	2.44	214.30	19.36

group
nfs45
fw
spy
el

Test Value: 0

Select syp and send it to the test variable(s): box and define the Test Value as 200. Click ok.

2. To Test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different.

Steps:

1. select**Analyze → Compare Means → Independent-Samples T Test.**

2. Select group and send it to the Grouping Variables box.

3. nfs45, fw, syp, sl under Test Variables(s) box.

4. Select Define Groups in the Independent-Samples T Test dialog box.

5. **Use Specified values**→ Define Groups as 1 and 2.

6. Click **OK.**

REFERENCES:

1. Design Resources Server. Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India. https://drs.icar.gov.in/

2. Morgan, G.A., Barrett, K. C. Leech, N.L.andGloecknerG.W. (2019).IBM SPSS for Introductory Statistics: Use and Interpretation. Sixth Edition, Routledge.

3. Nie, N. H., Bent, D. H. and Hull, C. H.(1970). SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill.

Online Training Programme on

Analysis of Agricultural Data Using Statistical and Data Mining Techniques

11 - 20 July,2023

by

National Agricultural Higher Education Project - Institution Development Plan (NAHEP-IDP), Rajmata Vijayraje Scindia Krishi Vishwa Vidyalaya, Gwalior, Madhya Pradesh

in collaboration with

National Agricultural Higher Education Project -Component 2 (NAHEP - Comp 2) ICAR-Indian Agricultural Statistics Research Institute,

New Delhi

Editors

RVSKVV, Gwalior

Dr. S.S. Tomar
Dr. V. B. Singh
Dr. Shashi Yadav
Dr. Nisha Singh
Dr. Ankita Sahu
Dr. Purnima Singh

ICAR-IASRI, New Delhi

Dr. Sudeep Marwaha
Dr. Shashi Dahiya
Dr. Mrinmoy Ray

Disclaimer: The information contained in this reference manual has been taken from various web resources. Respective URL's are mentioned in the content.

Citation: Singh, V.B., Dahiya, S., Marwaha, S., S., Ray, M., Sahu, A., Singh, P., Singh, N., Yadav, S., Tomar, S.S. (2023). Analysis of Agricultural Data using Statistical And Data Mining Techniques. Reference Manual, RVSKVV, Gwalior, ICAR-IASRI, NewDelhi

CONTENTS

ARTIFICIAL INTELLIGENCE AND KNOWLEDGE MANAGEMENT IN AGRICULTURE

Sudeep Marwaha
ICAR- Indian Agricultural Statistics Research Institute, New Delhi- 110012
sudeep@icar.gov.in

1. Introduction

The Artificial Intelligence is a very old field of study and has a rich history. Modern AI was formalized by John McCarthy, considered as father of AI. It is a branch of computer science, founded around early 1950's. Primarily, the term Artificial Intelligence (or AI) refers to a group of technique that enables a computer or a machine to mimic the behavior of humans in problem solving tasks. Formally, AI is described as "the study of how to make the computers do things at which, at the moment, people are better" (Rich and Knight, 1991; Rich *et al.*, 2009). The main aim of AI is to program the computer for performing certain tasks in humanly manner such as knowledgebase, reasoning, learning, planning, problem solving etc. The Machine Learning (ML) techniques are the subset of AI which makes the computers/machines/programs the capable of learning and performing tasks without being explicitly programmed. The ML techniques are not just the way of mimicking human behaviour but the way of mimicking how humans learn things. The main characteristics of machine learning is 'learning from experience' for solving any kind of problem. The methods of learning can be categorized into three types: (a) supervised learning algorithm is given with labelled data and the desired output whereas (b) unsupervised learning algorithm is given with unlabelled data and identifies the patterns from the input data and (c) reinforcement learning algorithm allows the ML techniques to capture the learnable things on the basis of rewards or reinforcement. Now, the Deep Learning (DL) technique are the advanced version of machine learning algorithms gained much popularity in the area of image recognition and computer vision. The artificial neural networks (ANNs) clubbed with representation learning are the backbone of the deep learning concepts. These techniques allow a machine to learn patterns in the dataset with multiple levels of abstractions. The DL models are composed of a series of non-linear layers where each of the layer has the capability of transforming the low-level representations into higher-level representations i.e. into a more abstract representations (Le Cun *et al.*, 2015). There are several DL algorithms available now-a-days such as Deep

Convolutional Neural Networks, Deep Recurrent Neural networks, Long Short-term Memory (LSTM) networks that are being applied to different areas of engineering, bioinformatics, agriculture, medical science and many more (Fusco *et al.,* 2021).

2. Applications of Artificial Intelligence in Agriculture:

In present scenario, AI techniques are being exponentially applied in the various areas of the agricultural domain. These areas can be categorized into the following groups: Soil and water management, Crop Health Management, Crop Phenotyping, Recommender-based systems for crops, Semantic web and Ontology driven expert systems for crops and Geo-AI. The application of AI, ML and DL based techniques on these areas are discussed in the following sections.

2.1 Soil and Irrigation Management:

Soil and irrigation are the most viable components of agriculture. The soil and irrigation are the determinant factors for the optimum crop yield. In order to obtain enhanced crop yield and to maintain the soil properties, there is a requirement of appropriate knowledge about the soil resources. The management of irrigation becomes crucial when there are scares of water availability. Therefore, the soil and irrigation related issues should be managed properly and cautiously to ensure a potential yield in crops. In this regards, AI and ML based techniques have shown potential ability to resolve soil and irrigation related issues in crops. A range of machine learning models such as linear regression, support vector machines (or regressors), Artificial neural networks, random forest algorithm and so on are being used. Many researchers have used remote-sensed data with the machine learning techniques for determining soil health parameters. In this section, few significant works in this field are highlighted below:

A. Soil Management:

Besalatpour *et al.,* (2011), Aitkenhead *et al.,* (2012) and Sirsat *et al.,* (2017) used different machine learning techniques such as linear regression, support vector machine, random forests for the prediction of the physical and chemical properties of soil. Rivera *et al.* (2020) and Azizi *et al.,* (2020) worked on estimation and classification of aggregate stability of the soils using conventional machine learning techniques as well as deep learning models. Jha *et al.,* (2018) worked on prediction of microbial dynamics in soils using regression-based techniques. Patil and Dekha (2016) and Mehdizadeh *et Al.* (2017) worked on predicting the evapotranspiration rate in crops using several machine learning techniques. Researchers worked on mapping the soil properties digitally using the remote sensing data with the help of machine learning and deep learning models (Taghizadeh-Mehrjardi *et al.* 2016; Kalambukattu *et al.,* (2018; Padarian *et al.* 2019; Taghizadeh-Mehrjardi *et al.,* 2020).

B. Irrigation management:

Zema *et al.* 2018 applied Data Envelopment Analysis (DEA) with Multiple Regression analysis to improve the irrigation performance Water Users Associations. Ramya *et al.* 2020 and Glória *et al.,* 2021worked on IoT based smart irrigation systems with machine learning models. Agastya *et al*, 2021 and Zhang *et al.* 2018 used deep learning-based CNN models for detection of irrigations using remote sensing data. Jimenez *et al.* 2021 worked on estimating the irrigation based on soil matric potential.

2.2 Crop Health Management:

Every year a significant amount of yield is damaged due to attack of disease causing pathogens and insect-pest infestation. In order to manage the spread of the diseases and insect-pests, proper management practices should be applied at the earliest. Therefore, there is requirement of automatic diseases, pest identification system. In this regard, image-based diagnosis of diseases and pests have become de facto standard of automatic stress identification. This kind of automated detection methodology use sophisticated deep learning-based AI techniques that reduces the intervention of the human experts. There are several attempts have been done to diagnose the diseases as well as insects-pests in crops using deep learning techniques. In this section, some of the significant works in this field have been discussed briefly.

A. Disease identification:

Mohanty *et al.* 2016 worked on disease diagnosis problem using deep CNN models. They used an open-source dataset named PlantVillage (Hughes and Salathe, 2016) containing 54,306 digital images of 26 diseases from 14 crops. Ferentinos, 2018 worked on developing deep CNN-based models for recognising 56 diseases from different crops. Barbedo, 2019 applied transfer learning approach for diagnosis of diseases of 12 different crops. Too et al. 2019, applied pre-trained deep CNN models for identification of diseases of 18 crops using the PlantVillage data. Chen *et al.* 2020 applied a pretrained VGGNet network for classifying the diseases of Rice and Maize crop. Chen *et. al.* 2020 and Rahman *et al.* 2020 worked on identifying the major diseases of Rice crop. Lu *et al.,* 2017; Johannes *et al.* 2017; Picon *et al.* 2019 and Nigam *et al.* 2021 applied deep CNN models for recognising the diseases of wheat crop. Priyadharshini *et al.* 2019; Sibiya & Sumbwanyambe, 2019; Haque *et al.* 2021 used deep learning models for identifying diseases of maize crop.

B. Pest Identification:

Pest Identification problem is inherently different from disease detection. As compared to disease detection there are less number of work has found in the literature. Some of the research of pest identification has been discussed in the following section.

Cheeti *et al.* (2021) developed a model for pest detection and classification of peat using YOLO (You look only once) and CNN. YOLO algorithm is used for detection of pest in an image and Alex net CNN is used for pest classification. Chen et al. (2021) propose an AI-based pest detection system for solving the specific issue of detection of scale pests based on pictures. Deep-learning-based object detection models, such as faster region-based convolutional networks (Faster R-CNNs), single-shot multibox detectors (SSDs), and You Only Look Once v4 (YOLO v4), are employed to detect and localize scale pests in the picture. Taiwan Agricultural Research Institute, Council of Agriculture, has collected images of the three types of pests from the actual fields for decades. Fuentes *et al.* (2017) address disease and pest identification by introducing the application of deep meta-architectures and feature extractors. They proposed a robust deep-learning-based detector for real-time tomato diseases and pests recognition. The system introduces a practical and applicable solution for detecting the class and location of diseases in tomato plants, which in fact represents a main comparable difference with traditional methods for plant diseases classification. Karnik *et al.* (2021) image pre-processing and data augmentation techniques has been performed to get better image.yolov3 classification for classifying plant leaf disease of pepper bell, potato and tomato. This proposed in divided into two stage part first classifier and second stage classifier where in first classifier it will preprocess of median filter and data augmentation is used and trained in yolov3 algorithm and in second stage classifier it will perform the extract plant leaf image output using Resnet50 based. So, it two step classification approach. Based on this research work we achieved 94% accuracy of detection lead diseases. Experiments showed [Li et al. (2020)] that our system with the custom backbone was more suitable for detection of the untrained rice videos than VGG16, ResNet-50, ResNet-101 backbone system and YOLOv3 with our experimental environment. Liu et al.2020 used Yolo V3 model is a little inadequate in the scale when recognizing tomato disease spots and pests.

2.3 Plant Phenomics:

Non-destructive phenotypic measurement with high throughput imaging technique becoming extremely popular. High throughput imaging system produces a large number of images. Deduction of the phenotypic characteristics through image analysis is quick and accurate. A wide range of phenotypic study can be done using phenomics analysis. High throughput imaging system coupled with sophisticated AI technology like deep learning make this field more efficient and accurate. Phenomics is has been used for study of several phenotypic characters like spike detection and counting, yield forecasting, quantification of the senescence in the plant, leaf weight and count, plant volume, convex hull, water stress and many more.

2.4 Recommender Systems:

Recommender systems (RSs) help online users in decision making regarding products among a pile of alternatives. In general, these systems are software solutions which predict liking of a user for unseen items. RSs have been mainly designed to help users in decision making for areas where one is lacking enough personal experience to evaluate the overwhelming number of alternative items that a website has to offer [Resnick & Varian, 1997]. Recommender systems have proved its worth in many different applications like e-commerce, e-library, e-tourism, e-learning, e-business, e-resource services etc. by suggesting suitable products to users [Lu *et al.*, 2015]. RSs are used to introduce new/unseen items to users, to increase user satisfaction etc. Recommendations are generated by processing large amount of historical data on the users and the products to be suggested. Most popular way of gathering users liking on a particular product is in terms of rating either in numerical scale (1 to 5) or ordinal scale (strongly agree, agree, neutral, disagree, strongly disagree). Other techniques of more knowledge – based recommendation are the use of Ontologies [Middleton *et al.*, 2002] of user profiles or item descriptions etc. The core task of a recommendation system is to predict the usefulness of an item to an individual user based on the earlier history of that item or by evaluating the earlier choices of the user. Collaborative way of user modelling [Konstan *et al.*, 1997] is where ratings are predicted for <user,

item> pair, $\overline{R}$<u, i> based on a large number of ratings previously gathered by the system on individual <user, item> pairs. Another way of recommendation is to suggest items that are similar to the ones previously liked by the user, called Content based filtering [Wang *et al.,* 2018; Smyth, 2007]. In a hybrid method of prediction, limitations by the earlier mentioned processes are tackled in various ways.

Agriculture has used recommender systems since 2015 and continues to do so. RSs have been explored to develop crop recommendation strategies based on soil and weather parameters, crop rotation practices, water management, suggestion on suitable varieties, recommendations for management practices etc. It is absolutely essential for the farmers to receive recommendations on the best crop for cultivation. Kamatchi and parvati, 2019 proposed a hybrid RS in combination with Collaborative Filtering, Case-based Reasoning and Artificial Neural Networks (ANN) to predict future climatic conditions and recommendation of crops based on the predicted climate. Crop recommendations have been developed based on season and productivity [Vaishnavi *et al.,* 2021], area and soil type [Pande *et al.,* 2021] by using several machine learning algorithms like Support vector Machine (SVM), Random forest (RF), Multivariate Linear regression (MLR), K- Nearest neighbour (KNN), ANN etc. Ensemble techniques have been used to develop a collaborative system of crop rotation, crop yield prediction, forecasting and fertilizer recommendation [Archana *et al.,* 2020]; to classify soil types into recommended crop types Kharif or Rabi based on specific physical and chemical characteristics, average rainfall and surface temperature [Kulkarni *et al.,* 2018]. Naha and Marwaha, 2020 presented an Ontology driven context aware RS that can recommend land preparation methods, sowing time, seed rate, fertilizer management, irrigation scheduling and harvesting methods to Maize cultivators. Application of RSs has also penetrated in the e-agriculture domain by suggesting parts of agricultural machineries in online ordering [Ballesteros *et al.,* 2021].

2.5 Semantic web, Knowledgebase and Natural Language processing:

Agriculture is vast source of resources and so it is also a vast source of information. The problem with this information is most of the information are unstructured. That unstructured knowledge is merely understandable for machine. It is also has low accessibility for human too. The main objectives of the semantic web and knowledge base system are to make unstructured data into structured one. Semantic web and the knowledgebase mainly facilitated by the ontology in the back end. Ontology is a formal, explicit specification of a shared conceptualization (Gruber, 1993). Making of Ontology that facilitated the semantic web and knowledge base can be made across the agricultural domain to make the unstructured data into structured one. Many ontology has already been developed in accordance with the Bedi and Marwaha, 2004 in the agricultural domain. Saha *et. al.,* (2011) developed an ontology on dynamic maize variety selection in different climatic condition, Sahiram *et. al.,* (2012) developed a ontology on rapeseed and mustard for identification of the variety in multiple languages, Das *et. al.,* (2011) developed a ontology for USDA soil taxonomy and ontology was extended by Deb *et. al*., (2012), Biswas *et. al*., (2012) developed a ontology on microbial taxonomy and was extended by Karn *et. al.* (2014).

2.6 GIS and Remote sensing coupled with AI:

GIS and Remote sensing is helping agricultural community since long. The land use planning, land cover analysis, forest distribution, water distribution, water use pattern, crop rotation and crop calendar analysis can be done by GIS and remote sensing. But when the AI and machine learning coupled with these technology it become more powerful. Machine learning and AI efficiently used for correct land classification and phonological change detection. From Digital soil mapping to yield forecasting, from phenology detection to leaf area index a vast range of the area in agriculture can be handled by GIS and Remote sensing.

References:

Agastya, C., Ghebremusse, S., Anderson, I., Vahabi, H., &Todeschini, A. (2021). Self-supervised Contrastive Learning for Irrigation Detection in Satellite Imagery. arXiv preprint arXiv:2108.05484.

AhilaPriyadharshini, R., Arivazhagan, S., Arun, M., &Mirnalini, A. (2019). Maize leaf disease classification using deep convolutional neural networks. Neural Computing and Applications, 31(12), 8887-8895.

Aitkenhead, M. J., Coull, M. C., Towers, W., Hudson, G., & Black, H. I. J. (2012). Predicting soil chemical composition and other soil parameters from field observations using a neural network. Computers and Electronics in Agriculture, 82, 108-116.

Archana, K., &Saranya, K. G. (2020). Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier. *Seventh Sense Research Group (SSRG) International Journal of Computer Science and Engineering,* 7(5), 1-4.

Azizi, A., Gilandeh, Y. A., Mesri-Gundoshmian, T., Saleh-Bigdeli, A. A., &Moghaddam, H. A. (2020). Classification of soil aggregates: A novel approach based on deep learning. Soil and Tillage Research, 199, 104586.

Ballesteros J. M., Cartujano, A. R., Evaldez, D., Macutay, J., (2021). Online ordering and recommender system of combine harvester parts and equipment with 3D modelling and augmented reality brochure for BLAZE equifarm and general merchandise. *11th International Workshop on Computer Science and Engineering (WCSE 2021)*, 174-179.

Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. Biosystems Engineering, 180, 96-107.

Besalatpour, A., Hajabbasi, M. A., Ayoubi, S., Gharipour, A., &Jazi, A. Y. (2012). Prediction of soil physical properties by optimized support vector machines. International Agrophysics, 26(2).

Biswas, S., Marwaha, S., Malhotra, P. K., Wahi, S. D., Dhar, D. W., & Singh, R. (2013). Building and querying microbial ontology. *Procedia Technology, 10*, 13-19.

Cheeti, S., Kumar, G. S., Priyanka, J. S., Firdous, G., &Ranjeeva, P. R. (2021). Pest Detection and Classification Using YOLO AND CNN. Annals of the Romanian Society for Cell Biology, 15295-15300.

Chen, J. W., Lin, W. J., Cheng, H. J., Hung, C. L., Lin, C. Y., & Chen, S. P. (2021). A smartphone-based application for scale pest detection using multiple-object detection methods. Electronics, 10(4), 372.

Chen, J., Zhang, D., Nanehkaran, Y. A., & Li, D. (2020). Detection of rice plant diseases based on deep transfer learning. Journal of the Science of Food and Agriculture, 100(7), 3246-3256.

Das, B. *et al.* (2017a) “Comparison of different uni-and multi-variate techniques for monitoring leaf water status as an indicator of water-deficit stress in wheat through spectroscopy,” *Biosystems Engineering*, 160, pp. 69–83.

Deb, C. K., Marwaha, S., Malhotra, P. K., Wahi, S. D., &Pandey, R. N. (2015, March). Strengthening soil taxonomy ontology software for description and classification of USDA soil taxonomy up to soil series. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1180-1184). IEEE.

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and electronics in agriculture, 145, 311-318.

Fuentes, A., Yoon, S., Kim, S. C., & Park, D. S. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors, 17(9), 2022.

Fusco, R., Grassi, R., Granata, V., Setola, S. V., Grassi, F., Cozzi, D., ...&Petrillo, A. (2021). Artificial Intelligence and COVID-19 Using Chest CT Scan and Chest X-ray Images: Machine Learning and Deep Learning Approaches for Diagnosis and Treatment. *Journal of Personalized Medicine*, **11(10)**, 993.

Glória, A.; Cardoso, J.; Sebastião, P. Sustainable irrigation system for farming supported by machine learning and real-time sensor data. Sensors 2021, 21, 3079.

Janik, L. J., Forrester, S. T., & Rawson, A. (2009). The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. Chemometrics and Intelligent Laboratory Systems, 97(2), 179-188.

Jimenez, A.F.; Ortiz, B.V.; Bondesan, L.; Morata, G.; Damianidis, D. Long short-term memory neural network for irrigation management: A case study from southern Alabama, USA. Precis. Agric. 2021, 22, 475–492

Johannes, A., Picon, A., Alvarez-Gila, A., Echazarra, J., Rodriguez-Vaamonde, S., Navajas, A. D., & Ortiz-Barredo, A. (2017). Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. Computers and electronics in agriculture, 138, 200-209.

Kalambukattu, J. G., Kumar, S., & Raj, R. A. (2018). Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. Environmental earth sciences, 77(5), 1-14.

Kamatchi, S., B. &Parvathi, R. (2019). Improvement of crop production using recommender system by weather forecasts. *Procedia Computer Science*, 165, 724–732.

Karnik, J., &Suthar, A. (2021). Agricultural Plant Leaf Disease Detection Using Deep Learning Techniques. Available at SSRN 3917556.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3), 77-87.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, **521(7553)**, 436-444.

Li, D., Wang, R., Xie, C., Liu, L., Zhang, J., Li, R., ...& Liu, W. (2020). A recognition method for rice plant diseases and pests video detection based on deep convolutional neural network. Sensors, 20(3), 578.

Liu, J., & Wang, X. (2020). Tomato diseases and pests detection based on improved Yolo V3 convolutional neural network. Frontiers in plant science, 11, 898.

Lu, J., Hu, J., Zhao, G., Mei, F., & Zhang, C. (2017). An in-field automatic wheat disease diagnosis system. Computers and electronics in agriculture, 142, 369-379.

Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G., (2015). Recommender system application developments: a survey. *Decision Support Systems*, 74, 12- 32.

Manoranjan, D., Malhotra, P. K., Sudeep, M., &Pandey, R. N. (2012). Building and querying soil ontology. *Journal of the Indian society of agricultural statistics*, *66*(3), 459-464.

Mohanty, S. P., Hughes, D. P., &Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in plant science, 7, 1419.

Naha, S. and Marwaha, S. (2020). Context-Aware Recommender System for Maize Cultivation. *Journal of Community Mobilization and Sustainable Development*, 15(2), 485-490.

Nigam, S., Jain, R., Marwaha, S., & Arora, A. (2021). Wheat rust disease identification using deep learning. De Gruyter.

Padarian, J., Minasny, B., &McBratney, A. B. (2019). Using deep learning for digital soil mapping. Soil, 5(1), 79-89.

Pande, S. M., Ramesh, P. K., Anmol, A., Aishwarya, B. R., Rohilla, K., &Shaurya, K. (2021). Crop recommender system using machine learning approach. *In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1066-1071. IEEE.

Patil, A. P., &Deka, P. C. (2016). An extreme learning machine approach for modelling evapotranspiration using extrinsic inputs. Computers and electronics in agriculture, 121, 385-392.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3), 56- 58.

Rich, E., & Knight, K. (1991). *Artificial Intelligence. 2nd Edn*. New York, NY, United States: McGraw-Hill.

Rich, E., Knight, K., and Nair, S. B. (2009). *Artificial Intelligence*. *3rd Edn*. New Delhi, India: Tata McGraw-Hill.

Rivera, J. I., & Bonilla, C. A. (2020). Predicting soil aggregate stability using readily available soil properties and machine learning techniques. Catena, 187, 104408.

Sibiya, M., &Sumbwanyambe, M. (2019). A computational procedure for the recognition and classification of maize leaf diseases out of healthy leaves using convolutional neural networks. AgriEngineering, 1(1), 119-131.

Sirsat, M. S., Cernadas, E., Fernández-Delgado, M., & Khan, R. (2017). Classification of agricultural soil parameters in India. Computers and electronics in agriculture, 135, 269-279.

Smyth, B. (2007). Case-based recommendation. In The adaptive web. Springer, Berlin, Heidelberg, 342-376.

Taghizadeh-Mehrjardi, R., Ayoubi, S., Namazi, Z., Malone, B. P., Zolfaghari, A. A., &Sadrabadi, F. R. (2016). Prediction of soil surface salinity in arid region of central Iran using auxiliary variables and genetic programming. Arid Land Research and Management, 30(1), 49-64.

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 161, 272-279.

Vaishnavi, S., Shobana, M., Sabitha, R., &Karthik, S. (2021). Agricultural Crop Recommendations based on Productivity and Season. *In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS).* 1, 883-886. IEEE.

Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for computer science publications. *Knowledge-Based Systems*, 157, 1-9.

Zema, D.A.; Nicotra, A.; Mateos, L.; Zimbone, S.M. Improvement of the irrigation performance in water users associations integrating data envelopment analysis and multi-regression models. Agric. Water Manag. 2018, 205, 38–49.

INTRODUCTION TO R SOFTWARE

Soumen Pal, B. N. Mandal
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
E-mail: Soumen.Pal@icar.gov.in

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.R is a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages.

R environment

The R environment provides an integrated suite of software facilities for data manipulation, calculation and graphical display. It has

- a data handling and storage facility,
- a suite of operators for calculations on arrays and matrices,
- a large, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display, and
- a well-developed, simple and effective programming language (called ‘S’) which includes conditionals, loops, user defined functions and input and output facilities.

Origin

R can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus systems. Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland started the project on R in 1995 and hence the name software has been named as ‘R’.

R was introduced as an environment within which many classical and modern statistical techniques can be implemented. A few of these are built into the base R environment, but many are supplied as packages. There are a number of packages supplied with R (called “standard” and “recommended” packages) and many more are available through the CRAN family of Internet sites (via http://cran.r-project.org) and elsewhere.

Availability

Since R is an open source project, it can be obtained freely from the website www.r-project.org. One can download R from any CRAN mirror out of several CRAN

(Comprehensive R Archive Network) mirrors. Latest available version of R is R version 4.3.1 and it has been released on June 16 2023.

Installation

To install R in windows operating system, simply double click on the setup file. It will automatically install the software in the system.

Usage

R can work under Windows, UNIX and Mac OS. In this note, we consider usage of R in Windows set up only.

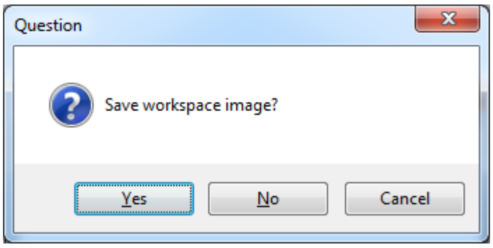
Difference with other packages

There is an important difference between R and the other statistical packages. In R, a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give large amount of output from a given analysis, R will give minimal output and store the results in an object for subsequent interrogation by further R functions.

Invoking R

If properly installed, usually R has a shortcut icon on the desktop screen and/or you can find it under Start| All Programs| R menu.

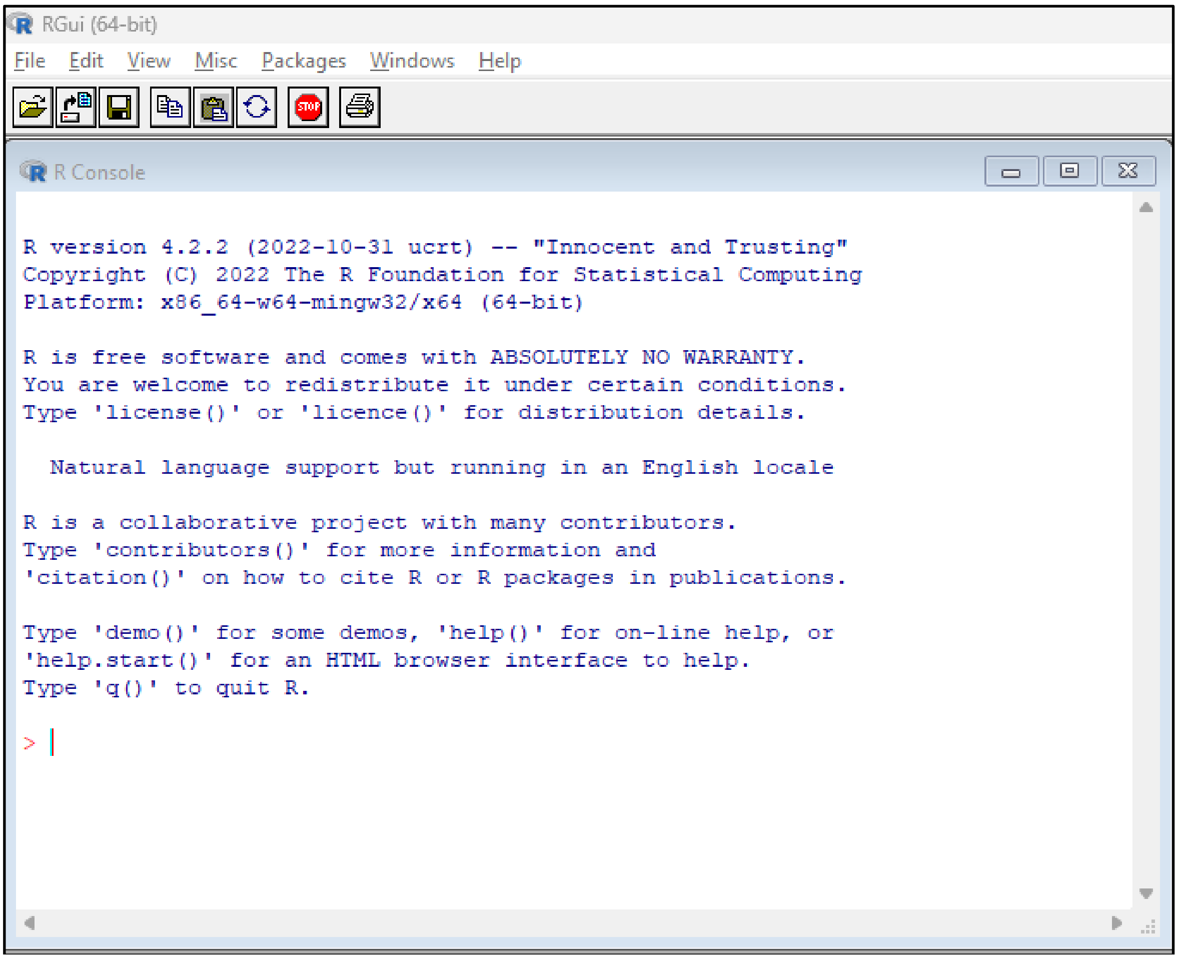
To quit R, type q() at the R prompt (>) and press Enter key. A dialog box will ask whether to save the objects you have created during the session so that they will become available next time when R will be invoked.


Question

Save workspace image?

Windows of R

R has only one window and when R is started it looks like



```
R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

R commands

i. R commands are case sensitive, so X and x are different symbols and would refer to different variables.

ii. Elementary commands consist of either expressions or assignments.

iii. If an expression is given as a command, it is evaluated, printed and the value is lost.

iv. An assignment also evaluates an expression and passes the value to a variable but the result is not automatically printed.

v. Commands are separated either by a semi-colon (';'), or by a newline.

vi. Elementary commands can be grouped together into one compound expression by braces '{' and '}'.

vii. Comments can be put almost anywhere, starting with a hashmark ('#'). Anything written after # marks to the end of the line is considered as a comment.

viii. Window can be cleared of lines by pressing Ctrl + L keys.

Executing commands from or diverting output to a file

If commands are stored in an external file, say ‘D:/commands.txt’ they may be executed at any time in an R session with the command

```
>source("d:/commands.txt")
```

For Windows Source is also available on the File menu.

The function *sink(),*

```
>sink("d:/record.txt")
```

will divert all subsequent output from the console to an external file, ‘record.txt’ in D drive. The command

```
>sink()
```

restores it to the console once again.

Simple manipulations of numbers and vectors

R operates on named data structures. The simplest such structure is the numeric vector, which is a single entity consisting of an ordered collection of numbers. To set up a vector named x, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

The function *c()* assigns the five numbers to the vector x. The assignment operator (<-) ‘points’ to the object receiving the value of the expression. Once can use the ‘=’ operator as an alternative.

A single number is taken as a vector of length one.

Assignments can also be made in the other direction, using the obvious change in the assignment operator. So the same assignment could be made using

```
>c(10.4, 5.6, 3.1, 6.4, 21.7) -> x
```

If an expression is used as a complete command, the value is printed. So now if we were to use the command

```
> 1/x
```

the reciprocals of the five values would be printed at the terminal.

The elementary arithmetic operators

+ addition

– subtraction

* multiplication

/ division

^ exponentiation

Arithmetic functions

log, exp, sin, cos, tan, sqrt,

Other basic functions

max(x) – maximum element of vector x,

min(x)- minimum element of vector x,

range (x) – range of the values of vector x ,

length(x) - the number of elements in x,

sum(x) - the total of the elements in x,

prod(x) – product of the elements in x

mean(x) – average of the elements of x

var(x) – sample variance of the elements of (x)

sort(x) – returns a vector with elements sorted in increasing order.

Logical operators

< - less than

<= less than or equal to

>greater than

>= greater than or equal to

== equal to

!= not equal to.

Other objects in R

Matrices or arrays - multi-dimensional generalizations of vectors.

Lists - a general form of vector in which the various elements need not be of the same type, and are often themselves vectors or lists.

Functions - objects in R which can be stored in the project's workspace. This provides a simple and convenient way to extend R.

Matrix facilities

A matrix is just an array with two subscripts. R provides many operators and functions those are available only for matrices. Some of the important R functions for matrices are

t(A) – transpose of the matrix A

nrow(A) – number of rows in the matrix A ncol(A) – number of columns in the matrix A

A%*% B– Cross product of two matrices A and B

A*B – element by element product of two matrices A and B

diag (A) – gives a vector of diagonal elements of the square matrix A

diag(a) – gives a matrix with diagonal elements as the elements of vector a

eigen(A) – gives eigen values and eigen vectors of a symmetric matrix A

rbind (A,B) – concatenates two matrix A and B by appending B matrix below A matrix (They should have same number of columns)

cbind(A, B) - concatenates two matrix A and B by appending B matrix in the right of A matrix (They should have same number of rows)

Data frame

Data frame is an array consisting of columns of various mode (numeric, character, etc). Small to moderate size data frame can be constructed by *data.frame()* function. For example, following is an illustration how to construct a data frame from the car data*:

Make	**Model**	**Cylinder**	**Weight**	**Mileage**	**Type**
Honda	Civic	V4	2170	33	Sporty
Chevrolet	Beretta	V4	2655	26	Compact
Ford	Escort	V4	2345	33	Small
Eagle	Summit	V4	2560	33	Small
Volkswagen	Jetta	V4	2330	26	Small
Buick	Le Sabre	V6	3325	23	Large
Mitsubishi	Galant	V4	2745	25	Compact
Dodge	Grand Caravan	V6	3735	18	Van
Chrysler	New Yorker	V6	3450	22	Medium
Acura	Legend	V6	3265	20	Medium

```
> Make<-
c("Honda","Chevrolet","Ford","Eagle","Volkswagen","Buick"
,"Mitsbusihi",
+ "Dodge","Chrysler","Acura")
>
Model=c("Civic","Beretta","Escort","Summit","Jetta","LeSa
bre","Galant",
+ "Grand Caravan","NewYorker","Legend")
```

Note that the plus sign (+) in the above commands are automatically inserted when the carriage return is pressed without completing the list. Save some typing by using *rep()* command. For example, *rep("V4",5)* instructs R to repeat V4 five times.

```
> Cylinder<-c(rep("V4",5),"V6","V4",rep("V6",3))
> Cylinder
 [1] "V4" "V4" "V4" "V4" "V4" "V6" "V4" "V6" "V6" "V6"
> Weight<-
c(2170,2655,2345,2560,2330,3325,2745,3735,3450,3265)
> Mileage<-c(33,26,33,33,26,23,25,18,22,20)
> Type<-
c("Sporty","Compact",rep("Small",3),"Large","Compact","Va
n",rep("Medium",2))
```

Now *data.frame()* function combines the six vectors into a single data frame.

```
> Car<-
data.frame(Make,Model,Cylinder,Weight,Mileage,Type)
> Car
         Make         Model Cylinder Weight
Mileage    Type
1       Honda         Civic       V4   2170   33  Sporty
2   Chevrolet       Beretta       V4   2655   26 Compact
3        Ford        Escort       V4   2345   33   Small
4       Eagle        Summit       V4   2560   33   Small
5  Volkswagen         Jetta       V4   2330   26   Small
6       Buick      LeSabre        V6   3325   23   Large
7  Mitsbusihi        Galant       V4   2745   25 Compact
8  Dodge Grand       Caravan      V6   3735   18     Van
9    Chrysler    New Yorker       V6   3450   22  Medium
10      Acura        Legend       V6   3265   20  Medium
> names(Car)
[1] "Make"     "Model"    "Cylinder"
"Weight"   "Mileage"  "Type"
```

Just as in matrix objects, partial information can be easily extracted from the data frame:

```
>Car[1,]
   Make Model Cylinder Weight Mileage   Type
1 Honda Civic       V4   2170      33 Sporty
```

In addition, individual columns can be referenced by their labels:

```
>Car$Mileage
 [1] 33 26 33 33 26 23 25 18 22 20
>Car[,5]        #equivalent expression
```

```
> mean(Car$Mileage)    #average mileage of the 10
vehicles
[1] 25.9
> min(Car$Weight)
[1] 2170
```

table() command gives a frequency table:

```
>table(Car$Type)
Compact   Large  Medium   Small  Sporty     Van
      2       1       2       3       1       1
```

If the proportion is desired, type the following command instead:

```
>table(Car$Type)/10
Compact   Large  Medium   Small  Sporty     Van
    0.2     0.1     0.2     0.3     0.1     0.1
```

Note that the values were divided by 10 because there are that many vehicles in total.

If you don't want to count them each time, the following does the trick:

```
>table(Car$Type)/length(Car$Type)
```

Cross tabulation is very easy, too:

```
>table(Car$Make, Car$Type)

             Compact Large Medium Small Sporty Van
  Acura      0       0     1      0     0      0
  Buick      0       1     0      0     0      0
  Chevrolet  1       0     0      0     0      0
  Chrysler   0       0     1      0     0      0
  Dodge      0       0     0      0     0      1
  Eagle      0       0     0      1     0      0
  Ford       0       0     0      1     0      0
  Honda      0       0     0      0     1      0
  Mitsbusihi 1       0     0      0     0      0
  Volkswagen 0       0     0      1     0      0
```

What if you want to arrange the data set by vehicle weight? *order()* gets the job done.

```
>i<-order(Car$Weight);i

 [1] 1 5 3 4 2 7 10 6 9 8

> Car[i,]

         Make         Model Cylinder Weight

Mileage    Type

1       Honda         Civic     V4   2170       33  Sporty

5  Volkswagen         Jetta     V4   2330       26   Small

3        Ford        Escort     V4   2345       33   Small

4       Eagle        Summit     V4   2560       33   Small

2   Chevrolet      Beretta     V4   2655       26 Compact

7  Mitsbusihi        Galant     V4   2745       25 Compact
```

```
10        Acura          Legend    V6    3265      20   Medium
6         Buick          LeSabre   V6    3325      23    Large
9      Chrysler        NewYorker   V6    3450      22   Medium
8     Dodge Grand        Caravan   V6    3735      18      Van
```

Creating/editing data objects

```
>y<-c(1,2,3,4,5);y
```

[1] 1 2 3 4 5

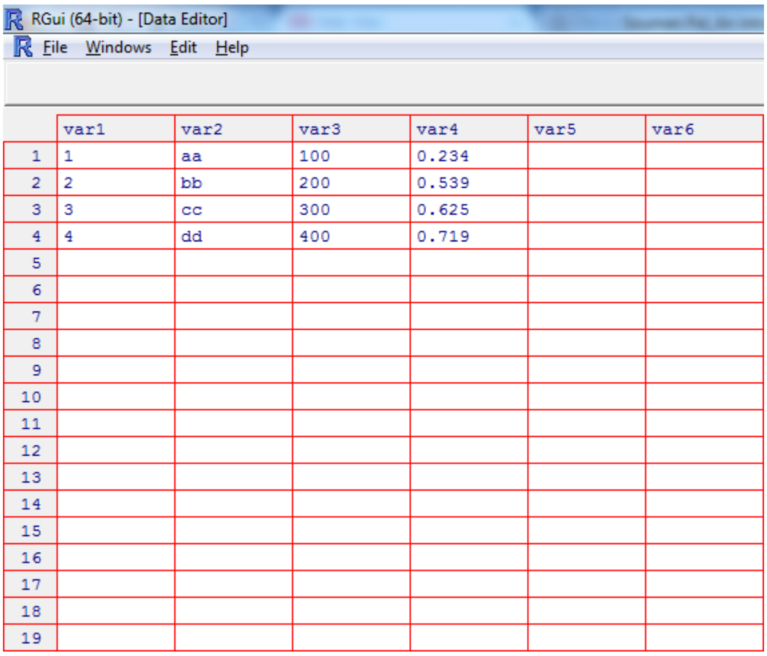
If you want to modify the data object, use *edit()* function and assign it to an object. For example, the following command opens R Editor for editing.

```
> y<-edit(y)
```

If you prefer entering the data.frame in a spreadsheet style data editor, the following command invokes the built-in editor with an empty spreadsheet.

```
> data1<-edit(data.frame())
```

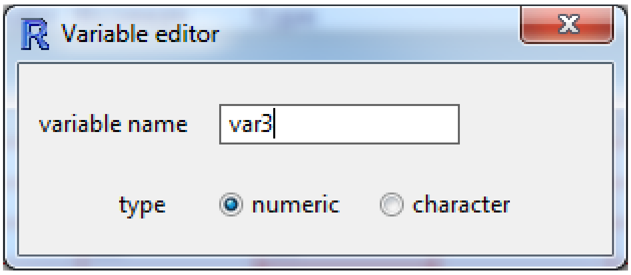
After entering a few data points, it looks like this:



	var1	var2	var3	var4	var5	var6
1	1	aa	100	0.234		
2	2	bb	200	0.539		
3	3	cc	300	0.625		
4	4	dd	400	0.719		
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

You can also change the variable name by clicking once on the cell containing it.

Doing so opens a dialog box:



variable name: var3

type ● numeric ○ character

When finished, click ☒ in the upper right corner of the dialog box to return to the

Data Editor window. Close the Data Editor to return to the R command window (R Console). Check the result by typing:

```
> data1
```

Reading data from files

When data files are large, it is better to read data from external files rather than entering data through the keyboard. To read data from an external file directly, the external file should be arranged properly.

The first line of the file should have a name for each variable. Each additional line of the file has the values for each variable.

Input file form with names and row labels:

Price	Floor	Area	Rooms	Age	isNew
52.00	111.0	830	5	6.2	no
54.75	128.0	710	5	7.5	no
57.50	101.0	1000	5	4.2	yes
57.50	131.0	690	6	8.8	no
59.75	93.0	900	5	1.9	yes

...

By default numeric items (except row labels) are read as numeric variables and non-numeric variables, such as isNew in the example, as factors. This can be changed if necessary.

The function *read.table()* can then be used to read the data frame directly

```
>HousePrice<-read.table("d:/houses.data", header = TRUE)
```

Reading comma delimited data

The following commands can be used for reading comma delimited data into R.

read.csv(filename) This command reads a .CSV file into R. You need to specify the exact filename with path.

read.csv(file.choose()) This command reads a .CSV file but the *file.choose()* part opens up an explorer type window that allows you to select a file from your computer. By default, R will take the first row as the variable names.

read.csv(file.choose(), header=T)

This reads a .CSV file, allowing you to select the file, the header is set explicitly. If you change to header=F then the first row will be treated like the rest of the data and not as a label.

Storing variable names

Through *read.csv()* or *read.table()* functions, data along with variable labels is read into R memory. However, to read the variables' names directly into R, one should use *attach(dataset)* function. For example,

```
>attach(HousePrice)
```

causes R to directly read all the variables' names eg. Price, Floor, Area etc. it is a good practice to use the *attach (datafile)* function immediately after reading the *datafile* into R.

Packages

All R functions and datasets are stored in packages. The contents of a package are available only when the package is loaded. This is done to run the codes efficiently without much memory usage. To see which packages are installed at your machine, use the command

```
>library()
```

To load a particular package, use a command like

```
>library(forecast)
```

Users connected to the Internet can use the *install.packages()* and *update.packages()* functions to install and update packages. Use *search()* to display the list of packages that are loaded.

Standard package

The standard (or base) packages are considered part of the R source code. They contain the basic functions those allow R to work with the datasets and standard statistical and graphical functions. They should be automatically available in any R installation.

Contributed packages and CRAN

There are a number of contributed packages for R, written by many authors. Various packages deal with various analyses. Most of the packages are available for download from CRAN (https://cran.r-project.org/web/packages/), and other repositories such as Bioconductor (http://www.bioconductor.org/). The collection of available packages changes frequently. As on June07, 2019, the CRAN package repository contains 14346 available packages.

Getting Help

Complete help files in HTML and PDF forms are available in R. To get help on a particular command/function etc., type *help (command name)*. For example, to get help on function 'mean', type *help(mean)* as shown below

```
>help(mean)
```

This will open the help file with the page containing the description of the function mean.

Another way to get help is to use "?" followed by function name. For example,

```
>?mean
```

will open the same window again.

In this lecture note, all R commands and corresponding outputs are given in `Courier New` font to differentiate from the normal texts. Since R is case-sensitive, i.e. typing *Help(mean),* would generate an error message,

```
>Help(mean)
Error in Help(mean) : could not find function "Help"
```

Further Readings

Various documents are available in https://cran.r-project.org/manuals.html from beginners' level to most advanced level. The following manuals are available in pdf form:

1. An Introduction to R
2. R Data Import/Export
3. R Installation and Administration
4. Writing R Extensions
5. The R language definition
6. R Internals
7. The R Reference Index

RStudio

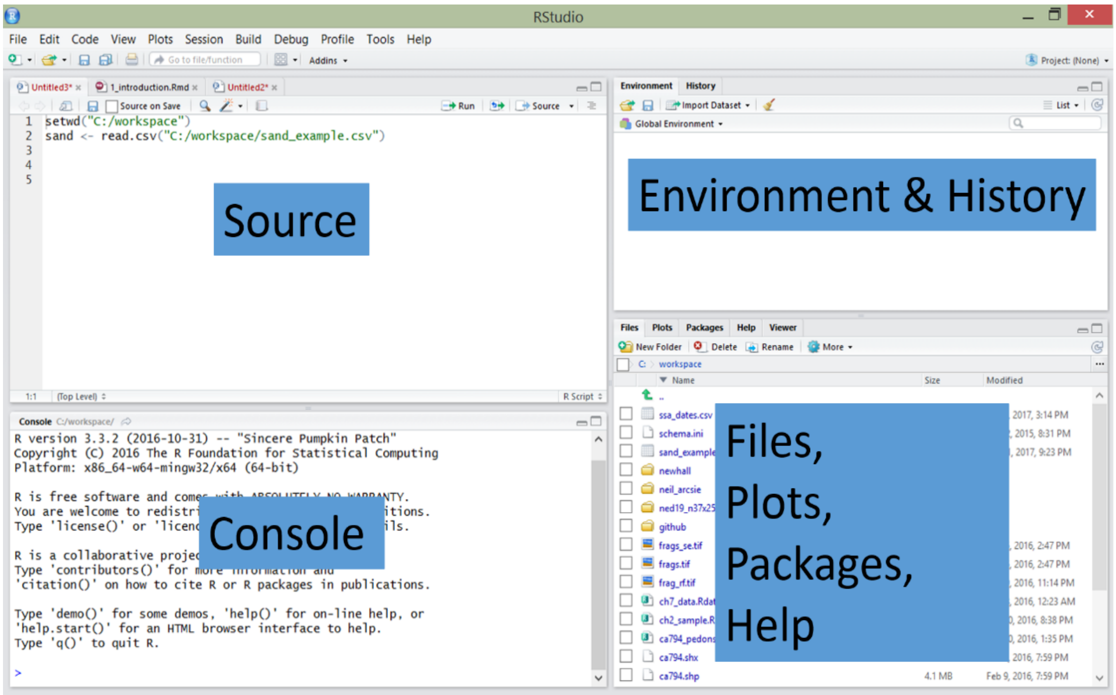
RStudio is an integrated development environment (IDE) that allows to interact with R more readily. RStudio is similar to the standard RGui, but is considerably more user friendly. It has more drop-down menus, windows with multiple tabs, and many customization options.

Installation of RStudio

RStudio requires R 3.0.1+ that means R software should be pre-installed before using RStudio.

RStudio requires a 64-bit operating system, and works exclusively with the 64 bit version of R. If you are on a 32 bit system or need the 32 bit version of R, you can use an older version of RStudio (https://www.rstudio.com/products/rstudio/older-versions/).RStudio free desktop version can be downloaded from the following link: https://www.rstudio.com/products/rstudio/download/#download

The first time RStudio is opened, three windows are seen. A forth window is hidden by default, but can be opened by clicking the **File** drop-down menu, then **New File**, and then **R Script**.

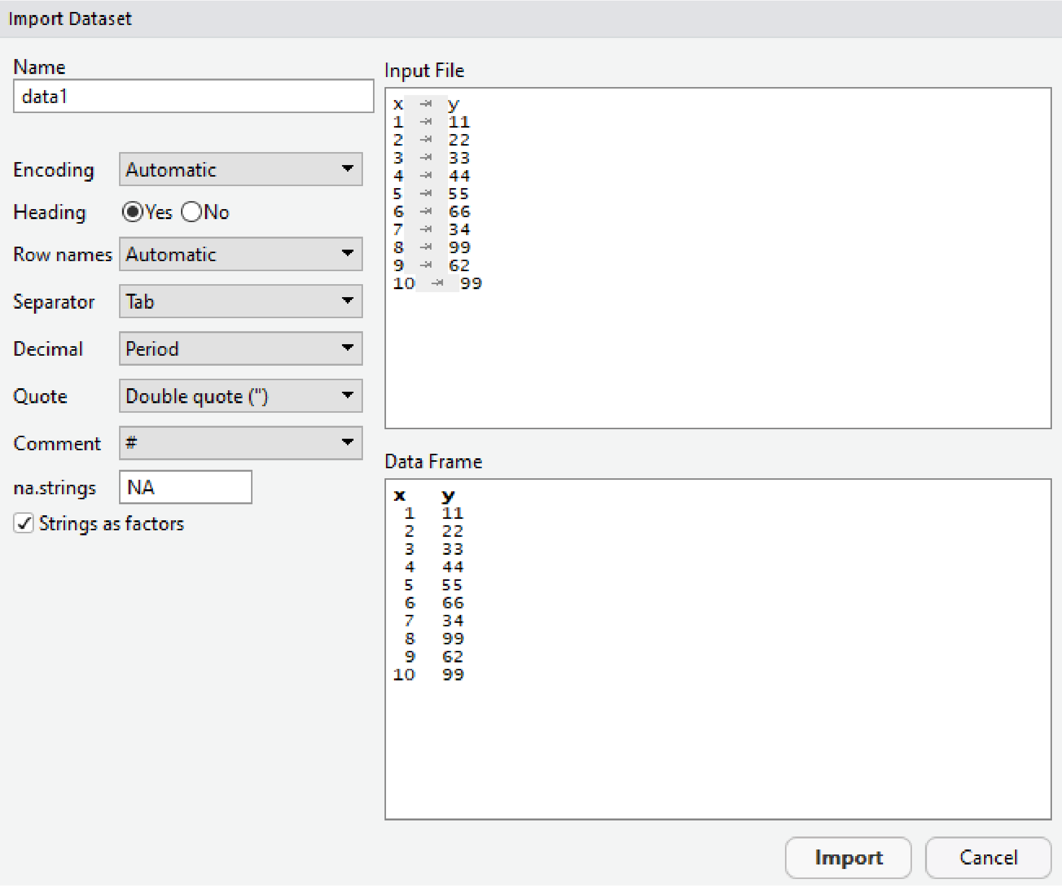


```
setwd("C:/workspace")
sand <- read.csv("C:/workspace/sand_example.csv")
```

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistri                     itions.
Type 'license()' or 'licenc                     ils.

R is a collaborative projec
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Importing Data in R Studio

1. Click on the import dataset button in the top-right section under the environment tab. Select the file you want to import and then click open. The Import Dataset dialog will appear as shown below
2. After setting up the preferences of separator, name and other parameters, click on the Import button. The dataset will be imported in R Studio and assigned to the variable name as set before.



Import Dataset

Name: data1

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Automatic

Separator: Tab

Decimal: Period

Quote: Double quote (")

Comment: #

na.strings: NA

☑ Strings as factors

Input File

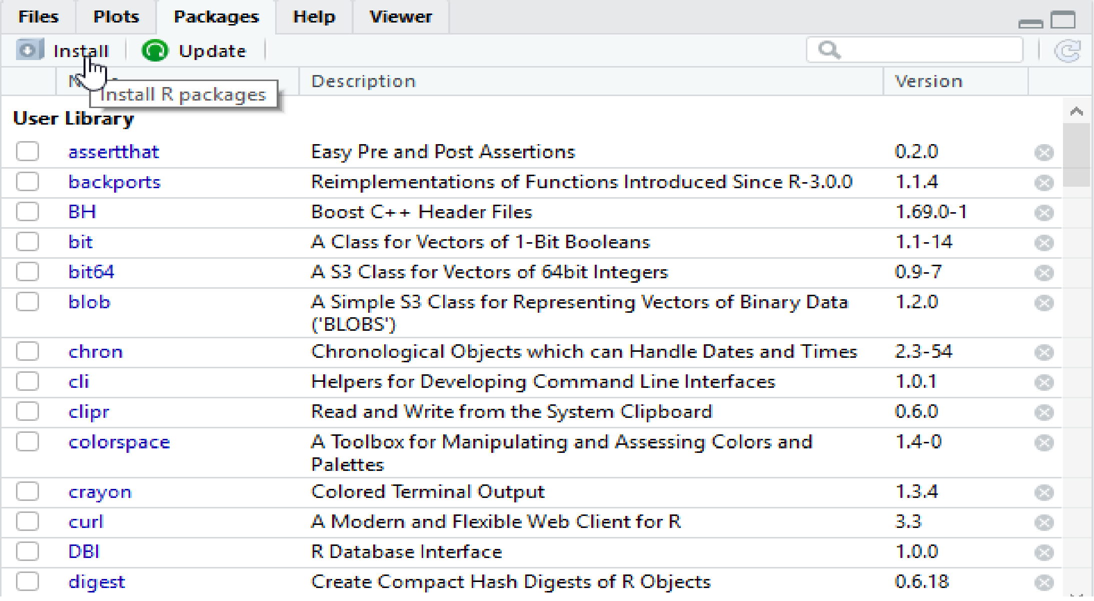
```
x	y
1	11
2	22
3	33
4	44
5	55
6	66
7	34
8	99
9	62
10	99
```

Data Frame

```
x   y
 1  11
 2  22
 3  33
 4  44
 5  55
 6  66
 7  34
 8  99
 9  62
10  99
```

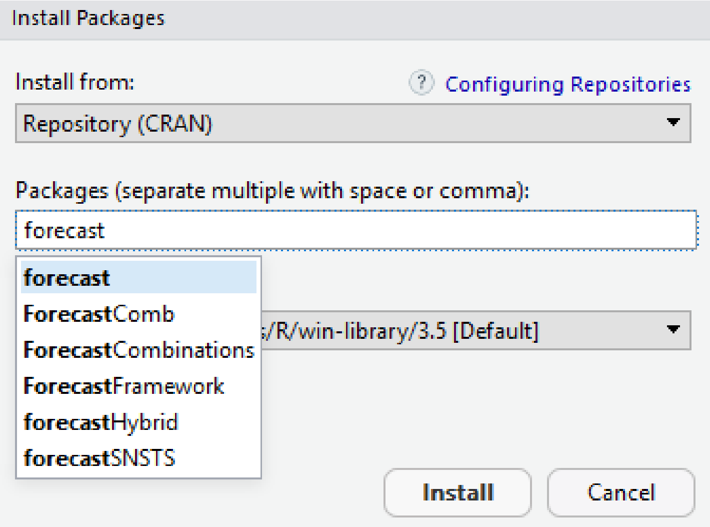
Installing Packages in RStudio

Within the **Packages** tab, a list of all the packages currently installed on the working computer and 2 buttons labeled either “Install” or “Update” are seen. To install a new package simply select the Install button. It is possible to install one or more than one packages at a time by simply separating them with a comma.



User Library

	Name	Description	Version
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.4
□	BH	Boost C++ Header Files	1.69.0-1
□	bit	A Class for Vectors of 1-Bit Booleans	1.1-14
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.0
□	chron	Chronological Objects which can Handle Dates and Times	2.3-54
□	cli	Helpers for Developing Command Line Interfaces	1.0.1
□	clipr	Read and Write from the System Clipboard	0.6.0
□	colorspace	A Toolbox for Manipulating and Assessing Colors and Palettes	1.4-0
□	crayon	Colored Terminal Output	1.3.4
□	curl	A Modern and Flexible Web Client for R	3.3
□	DBI	R Database Interface	1.0.0
□	digest	Create Compact Hash Digests of R Objects	0.6.18



Install from: Repository (CRAN)

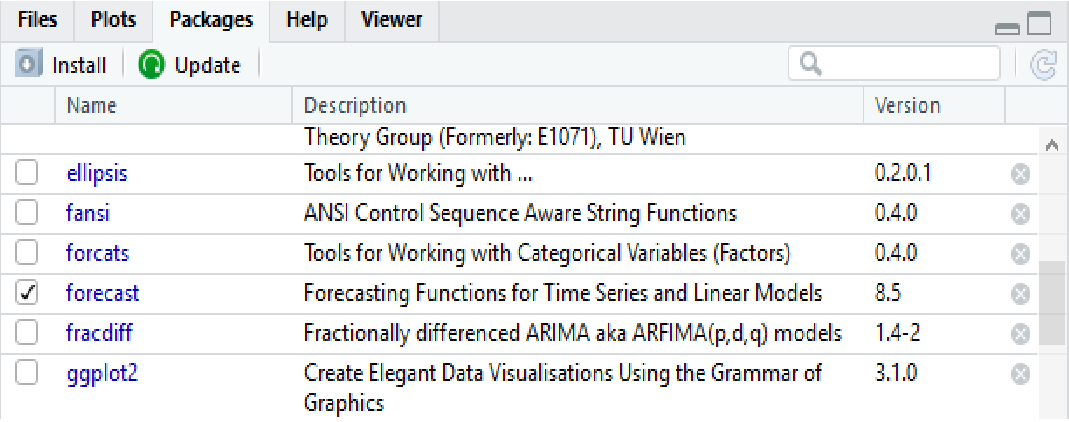
Packages (separate multiple with space or comma): forecast

- forecast
- ForecastComb
- ForecastCombinations
- ForecastFramework
- forecastHybrid
- forecastSNSTS

:/R/win-library/3.5 [Default]

Loading Packages in RStudio

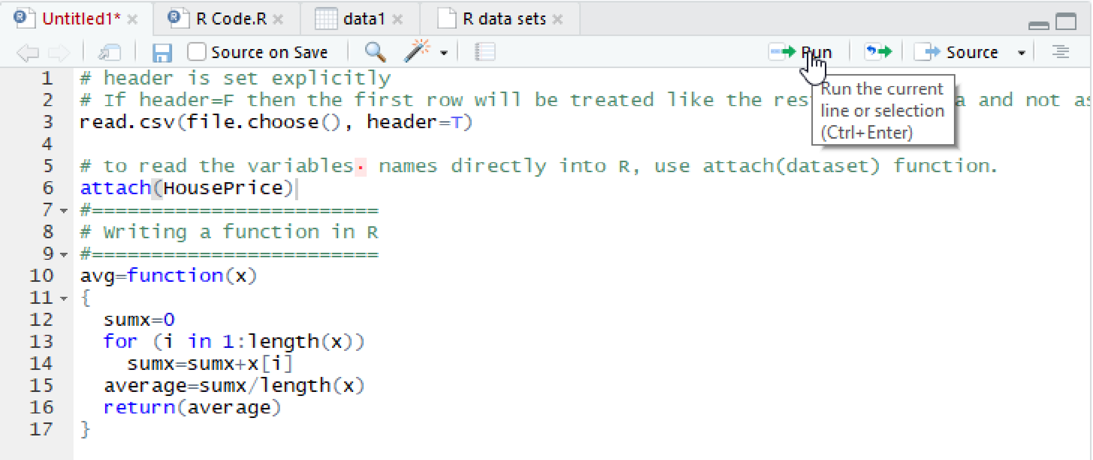
Once a package is installed, it must be loaded into the R session to be used.



	Name	Description	Version
		Theory Group (Formerly: E1071), TU Wien	
□	ellipsis	Tools for Working with ...	0.2.0.1
□	fansi	ANSI Control Sequence Aware String Functions	0.4.0
□	forcats	Tools for Working with Categorical Variables (Factors)	0.4.0
☑	forecast	Forecasting Functions for Time Series and Linear Models	8.5
□	fracdiff	Fractionally differenced ARIMA aka ARFIMA(p,d,q) models	1.4-2
□	ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	3.1.0

Writing Scripts in RStudio

RStudio's Source Tabs serve as a built-in text editor. Prior to executing R functions at the Console, commands are typically written down (or scripted).To write a script, simply open a new R script file by clicking File>New File>R Script.



Run the current line or selection (Ctrl+Enter)

```
# header is set explicitly
# If header=F then the first row will be treated like the res          a and not as
read.csv(file.choose(), header=T)

# to read the variables' names directly into R, use attach(dataset) function.
attach(HousePrice)
#========================
# Writing a function in R
#========================
avg=function(x)
{
  sumx=0
  for (i in 1:length(x))
    sumx=sumx+x[i]
  average=sumx/length(x)
  return(average)
}
```

Within the text editor type out a sequence of functions.

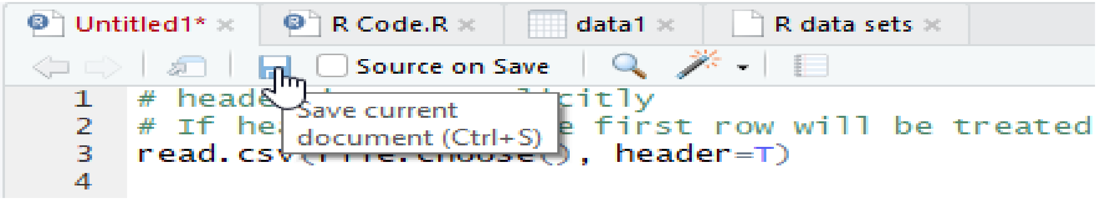
- Place each function (e.g. read.csv()) on a separate line.
- If a function has a long list of arguments, place each argument on a separate line.
- A command can be executed from the text editor by placing the cursor on a line and typing Crtl + Enter, or by clicking the Run button.
- An entire R script file can be executed by clicking the Source button.

Saving R files in RStudio

In R, several types of files can be saved to keep track of the work performed. The file types include: script, workspace, history and graphics.

R script (.R)

An R script is a text file of R commands that have beentyped. To save R scripts in RStudio, click the save button from R script tab. Save scripts with the .R extension.

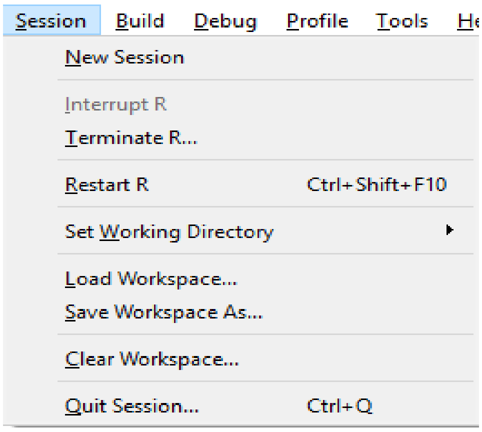


```
# heade                  citly
# If he                e first row will be treated
read.csv                   , header=T)
```

To open an R script, click the file icon.

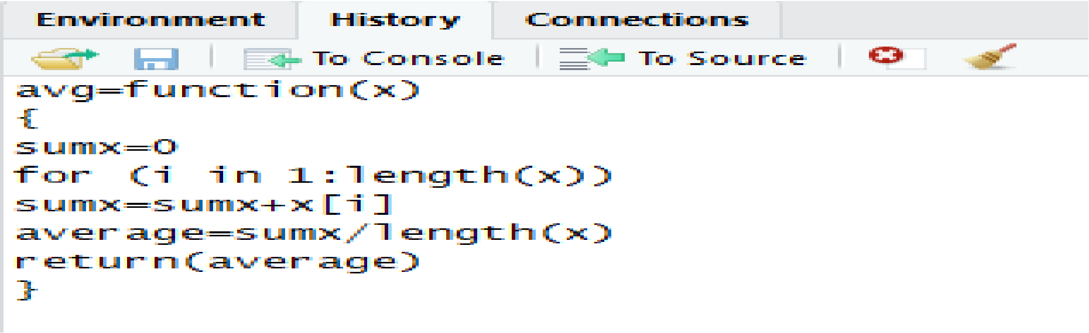
Workspace (.Rdata)

The R workspace consists of all the data objects created or loaded during the R session. It is possible to save or load the workspace at any time during the R session from the menu by clicking Session>Save Workspace As.., or the save button on the Environment Tab.



R history (.Rhistory)

Rhistory file is a text file that lists all of the commands that have been executed. It does not keep a record of the results. To load or save R history from the History Tab click the **Open File** or **Save** button.

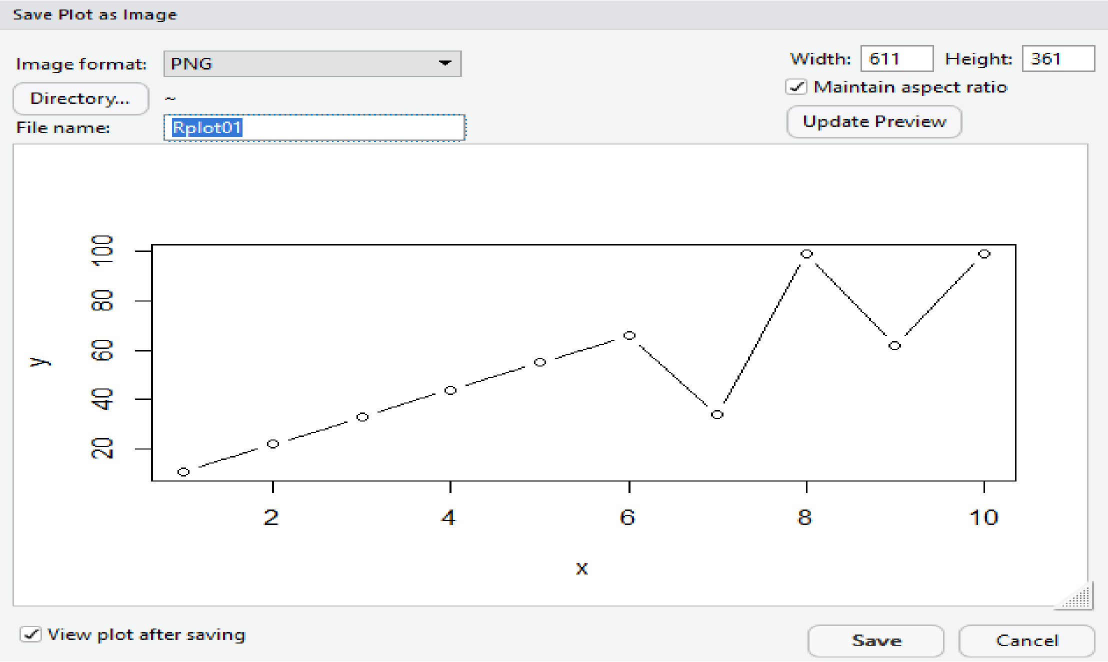


```
avg=function(x)
{
sumx=0
for (i in 1:length(x))
sumx=sumx+x[i]
average=sumx/length(x)
return(average)
}
```

R Graphics

Graphic outputs can be saved in various formats like pdf, png, jpeg, bmp etc.

To save a graphic: (1) Click the **Plots** Tab window, (2) click the **Export** button, (3) **Choose** desired format, (4) **Modify** the export settings as desired and (4) click **Save**.



Save Plot as Image

Image format: PNG

Directory... ~

File name: Rplot01

Width: 611 Height: 361

☑ Maintain aspect ratio

☑ View plot after saving

References

1. http://ncss-tech.github.io/stats_for_soil_survey/chapters/1_introduction/1_introduction.html
2. http://web.cs.ucla.edu/~gulzar/rstudio/basic-tutorial.html
3. http://www.gardenersown.co.uk/Education/Lectures/R/index.htm
4. https://www.cran.r-project.org
5. https://www.rstudio.com
6. Matloff, N. (2011). *The art of R programming: A tour of statistical software design*. No Starch Press.
7. Venables, W. N., Smith, D. M. and R Development Core Team (2009). An introduction to R: Notes on R: A programming Environment for Data Analysis and Graphics, version 1.7. 1.

DESCRIPTIVE STATISTICS AND EXPLORATORY DATA ANALYSIS

Md Yeasin
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
E-mail: yeasin.iasri@gmail.com

1. Introduction

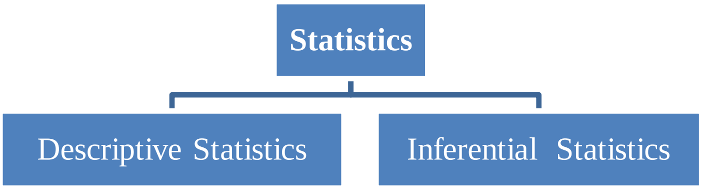
The word 'Statistics' has been derived from the Latin word '**Status**' or the Italian word '**Statista**' or the German word '**Statistik**' each of which means 'political state'. Statistics is a broad concept featuring applications in a wide range of areas. Statistics, in general, can be defined as the process for collecting, analyzing, interpreting, and making conclusions from data. In other terms, statistics is the approach established by scientists and mathematicians for analyzing and deriving conclusions from acquired data. Everything that has anything to do with the collection, processing, interpretation, and presentation of data falls within the scope of statistics.

Definition of statistics: Statistics is a branch of mathematics that deals with collecting, organizing, summarizing, presenting, and analyzing data as well as providing valid results and interpreting towards reasonable decisions.

Statisticians, in other words, give methodologies for

- **Design:** Planning and conducting out research projects.
- **Description:** Data summarization and exploration.
- **Inference:** Making predictions and inferences about the data

Statistics can be divided into two sections; one is descriptive statistics and another is inferential statistics.



Statistics

Descriptive Statistics

Inferential Statistics

Descriptive statistics helps describe, show or summarize data in a meaningful way. Descriptive statistics provides us with tools, tables, graphs, averages, ranges, correlations for organizing and summarizing data. Examples: measures of central tendency, measures of dispersion, skewness, kurtosis etc.

Inferential statistics helps to understand the properties of the population by observing the sample values. Inferential statistics deals with the estimation of parameters and test of hypothesis.

In this section we briefly discussed the descriptive statistics such as measures of central tendency, measures of dispersion, skewness, and kurtosis

2. Measures of central tendency

Central tendency is a statistical measure that determines a single value that accurately describes the center of the distribution. The objective of central tendency is to identify the single value that is the best representative for the entire set of data.

Different measure of central tendency are:

- Mean
    - Arithmetic mean
    - Geometric mean
    - Harmonic mean
- Median
- Mode
- Quartiles
- Deciles
- Percentiles

1.1.Mean (Arithmetic mean: A.M.):

The mean is the most commonly used measure of central tendency. For computation of the mean data should be numerical values measured on an interval or ratio scale. To compute the mean, we add the observation of data sets and then divide by the number of observation.

$$Mean = \frac{Sum\ of\ all\ observation}{Total\ number\ of\ observationa}$$

1.1.1. Simple mean: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The arithmetic mean is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Mean for frequency distribution: Let $X_1, X_2, \dots, X_n$ are observations with correspondingfrequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The arithmetic mean is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$$

Properties of mean:

- It depends on change of origin as well as the change of scale.

$$U = a + hX$$

Where a is origin and h is scale

Then $\bar{U} = a + h\bar{X}$.

- If are$\bar{X}_1$and $\bar{X}_2$ the means of two sets of values with $n_1$and n_2 observations respectively, then their combined mean is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

- Algebraic sum of deviations of set of values from their mean is zero.

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

- The sum of squares of deviation of set of values about its mean is minimum

$$\textstyle\sum_{i=1}^{n}(X_i - A)^2 \text{ is minimum when } A = \bar{X}$$

Merits of mean:

- Easy to understand
- Easy to calculate.
- It is rigidly defined.
- It is based on all observations.
- It is least affected by sampling fluctuations.
- It is capable of further mathematical treatment.

Demerits of mean:

- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.
- It cannot be calculated if any observations are missing in the data series.
- It is not suitable for highly skewed distribution.

1.1.2. Geometric mean (G.M.):

For n observations, Geometric mean is the nth root of their product.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The geometric mean is defined as

$$G = (X_1 * X_2 * \dots * X_n)^{1/n}$$

For frequency distribution: Let $X_1, X_2, \dots, X_n$ are observations with correspondingfrequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The geometric mean is defined as

$$G = ({X_1}^{f_1} * {X_2}^{f_2} * \dots * {X_n}^{f_n})^{1/N}$$

Use of geometric mean:

- Measure average relative changes, averaging ratios and percentages
- Best average for construction of index number

Merits of geometric mean:

- It is based on all observations.
- It is not affected by sampling fluctuations.
- It is capable of further mathematical treatment.

Demerits of geometric mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.
- It cannot be calculated if any observations are missing in the data series.

1.1.3. Harmonic mean (H.M.):

Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the observations of the sets.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The harmonic mean is defined as

$$H = \frac{n}{\sum_{i=1}^{n} 1/X_i}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The harmonic mean is defined as

$$H = \frac{N}{\sum_{i=1}^{n} f_i/X_i}$$

Use of harmonic mean:

- Measure the change where the values of a variable are compared with a constant quantity of another variable like time, distance travelled within a given time, quantities purchased or sold over a unit.

Merits of harmonic mean:

- It gives more weight to the small item and less weight to large values.
- It is based on all observations.
- It is not affected by sampling fluctuations.

- It is capable of further mathematical treatment.

Demerits of harmonic mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristics.
- It cannot be calculated if any observations are missing in the data series.

Relation between A.M., G.M. and H.M.:

- For given two observations, $A.M. \geq G.M. \geq H.M.$
- $G.M. = \sqrt{A.M. * H.M.}$
- $A.M. = \frac{G.M.^2}{H.M.}$
- $H.M. = \frac{G.M.^2}{A.M.}$

1.2. Median:

Median is the value situated in the middle position when all the observations are arranged in an ascending/descending order. The median is the central value of an ordered data series. It divides the data sets exactly into two parts. Fifty percent of observations are below the median and 50% are above the median. Median is also known as 'positional average'. The Median is the 50th percentiles, 10th deciles, and 2nd quartiles. Median is also the intersect point of less than and more than ogive curve.

Median for non-frequency data:

Step 1 Order the data from smallest to largest.

Step2 If the number of observations is odd, then (n + 1)/2th observation (in the ordered set) is the median. When the total number of observations is even, the median is given by the mean of n/2th and (n/2 + 1)th observation.

Median for group frequency data:

Step 1 Obtain the cumulative frequencies for the data.

Step 2 Markthe class corresponding to which a cumulative frequency is greater than N/2. That class is the median class.

Step 3 Then median is evaluated by an interpolation formula

$$Median = l + \frac{h}{f}(\frac{N}{2} - C)$$

Where, l = lower limit of the median class

N= Number of observations

C = cumulative frequency of the class proceeding to the median class

f = frequency of the median class

h= magnitude of the median class

Note: Graphically, we can find the median by histogram.

Use of median:

- Qualitative data can be arranged in ascending or descending order of magnitude.
- Find average intelligence, honesty, etc.

Merits of median:

- It is rigidly defined.
- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on an ordinal scale.

Demerits of median:

- It is not based on all observations.
- The calculation is more complex than the mean.
- It is not capable of further mathematical treatment.
- As compared to the mean, it is much affected by sampling fluctuations.

1.3 Mode:

Mode is defined as the value that occurs most frequently in the data. If in the data sets each observation occurs only once, then it does not have mode. When the data set has two or more values equal to the highest frequency than two or more mode are present in the datasets.

Mode for ungroup frequency data: The observation which has the highest frequency in the data sets.

Mode for group (equal width) frequency data:

Step 1 Identify the modal class. Modal class is the class with the largest frequency.

Step 2 Find mode by using interpolated formula.

$$mode = l + \frac{h(f_0 - f_{-1})}{(f_0 - f_{-1}) - (f_1 - f_0)}$$

Where, l = lower limit of the modal class

f_0 = frequency of the modal class

f_{-1}=frequency of the preceding modal class

f_1=frequency of the succeeding modal class

h= magnitude of the modal class

Note: Graphically, we can find mode by histogram.

Use of mode:

- To find ideal consumer preferences for different kinds of products.
- The best measure for the average size of shoes or shirts.

Merits of mode:

- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on a nominal scale.

Demerits of mode:

- It is ill-defined.
- It is not based on all observations.
- The calculation is more complex than the mean.
- It is not capable of further mathematical treatment.
- As compare to the mean, it is much affected by sampling fluctuations.

Quartiles: Quartiles are the three points that divide the whole data into four equal parts.

$$Q_i = l + \frac{h}{f}(\frac{iN}{4} - C)$$

Deciles: Deciles are the nine points that divide the whole data into ten equal parts.

$$D_i = l + \frac{h}{f}(\frac{iN}{10} - C)$$

Percentiles: Percentiles are the ninety-nine point that divides the whole data into hundreds of equal parts.

$$P_i = l + \frac{h}{f}(\frac{iN}{100} - C)$$

Note: $\boldsymbol{Median = 2nd\ Quartles = 5th\ Deciles = 50th\ Percentiles}$

Empirical formula between mean median and mode: If the data sets area symmetric in nature, then

$$\boldsymbol{Mean - \ Mode\ = 3(Mean - \ Median)}$$

The best measure of central tendency:

According to proof. Yule, Mean is the best measure of central tendency. But there are some situations where the other measures of central tendency are preferred.

Scale	Use measure	Best measure
Interval	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ratio	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ordinal	Median, Mode	Median
Nominal	Mode	Mode

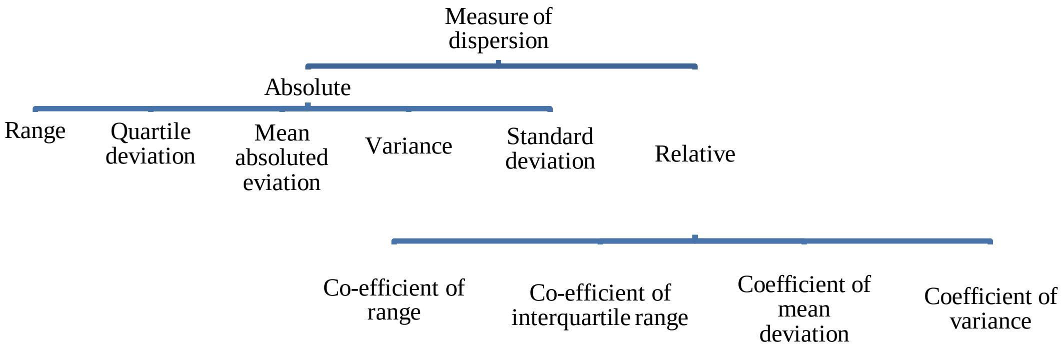
2. Measure of Dispersion

The measure of central tendency such as mean, median, and mode only locate the center of the data. It does not infer anything about the spread of the data. Two data sets can have the same mean but they can be entirely different.

Data 1	38	42	41	44	45
Data 2	50	53	41	35	31

In the above example, two datasets have the same mean. So measures of central tendency are not adequate to describe data. Thus to describe data, one needs to know the measure of scatterness of observations. Dispersion is defined as deviation or scatterness of observations from their central values.

Various measure of dispersion are:



1.2 Range (R):

Range is the simplest measure of dispersion. It is defined as the difference between the highest value and lowest value of the variable. It is a crude measure of dispersion.

$$\boldsymbol{Range = highest\ value\ (H) - lowest\ value\ (L)}$$

Merits of range:

- It is easy to understand and calculate.
- It is not affected by frequency of the data.

Demerits of range:

- It does not depend on all observations.
- It is very much affected by the extreme items.
- It cannot be calculated from open-end class intervals.
- It is not suitable for further mathematical treatment.
- It is the most unreliable measure of dispersion.

1.3 Quartile deviation (Q.D.):

Interquartile range is the difference between the first and third quartile. Hence the interquartile range describes the middle 50% of observations.

$$\boldsymbol{Inter\ quartile\ range = \ Q3 - \ Q1}$$

Where,

Q^3=first quartile of the data

Q^1=third quartile of the data

Quartile deviation (Q.D.) is the half of the inter quartile range.

$$\boldsymbol{Quartile\ deviation\ (Q.D.)\ = \frac{Q3 - \ Q1}{2}}$$

Merits of Quartile deviation:

- It is easy to understand and calculate.
- It is not affected by extreme values
- It can be calculated for open end frequency data

Demerits of Quartile deviation:

- It does not depend on all observations.
- It is not suitable for further mathematical treatment.
- It is very much affected by sampling fluctuations.

1.4 Mean absolute deviation (MAD):

The absolute deviation of each value from the central value (mean is preferable) is calculated and the arithmetic mean of these deviations is called mean absolute deviation.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observations of a data set. The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^{n} |X_i - A|}{n}$$

The mean absolute deviation (MAD) about mean is given by

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^{n} f_i |X_i - A|}{N}$$

The mean absolute deviation (MAD) about mean is given by

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} f_i |X_i - \bar{X}|}{N}$$

Merits of mean absolute deviation about mean:

- It is easy to understand and calculate.
- It is based on all observations.

Demerits of mean absolute deviation about mean:

- It is not suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.

1.5 Standard deviation (S.D.):

It is the best measure and the most commonly used measure of dispersion. It is defined as the positive square-root of the arithmetic mean of the square of the deviations of the given observation from their arithmetic mean. It takes into consideration the magnitude of all the observations and gives the minimum value of dispersion possible. It is also known as Root Mean Square Deviation about mean.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The standard deviation A is given by

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The standard deviation is given by

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^2}{N}}$$

Properties of standard deviation:

- It is the independent of the change of origin but dependent on the change of scale

  Let $U = a + hX$, then $sd(U) = |h| * sd(x)$
- If all observations are equal standard deviation is zero.
- It is never less than the quartile deviation and mean absolute deviation.

Merits of standard deviation:

- It is based on all observations.
- It is less affected by extreme values.
- It is suitable for further mathematical treatment.

Demerits of standard deviation:

- It is suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.
- It cannot be computed for open-end class data.

1.6 Variance

It is defined as the square of the standard deviation. Unit of the variance is the square of the actual observations, whereas unit of the standard deviation is same as actual observations.

Relations between R, Q.D., M.D. and S.D.

$$\mathbf{9QD = \frac{15}{2}MD = 6SD = R}$$

1.7 Coefficient of Variation (CV):

The Coefficient of variation for a data set defined as the ratio of the standard deviation to the mean and expressed in percentage.

$$\boldsymbol{CV = \frac{SD}{mean} * 100\%}$$

C.V is the relative measure of dispersion. It is the best measure among all the relative measure of dispersion. C.V is used to compare variability or consistency between two or more data series. If C.V. is greater indicate that the group is more variable, less stable, less uniform and less consistent. If the C.V. is less, it indicates that the group is less variable or more stable or more uniform and more consistent.

Example: Consider the data on score of Kohli and Smith in ODI cricket. The mean and standard deviation for Kohli are 55 and 5 respectively. The mean and standard deviation for Smith are 50 and 10 respectively. Find C.V. value for both the data and make compare them.

Solution:

For Kohli, $CV = \frac{5}{55} * 100 = 9\%$

For Smith, $CV = \frac{10}{50} * 100 = 20\%$

The Smith is subject to more variation in score than Kohli. So Kohli is more consistent than Smith.

$$\mathbf{3.6. Coefficient\ of\ range} = \frac{\boldsymbol{H-L}}{\boldsymbol{H+L}} * \mathbf{100\%}$$

$$\mathbf{3.7.\ Coefficient\ of\ inter\ quartile\ range} = \frac{\boldsymbol{Q3-Q1}}{\boldsymbol{Q3+Q1}} * \mathbf{100\%}$$

$$\mathbf{3.8. Coefficient\ of\ mean\ deviation}$$

$$= \frac{\boldsymbol{MAD}}{\boldsymbol{averave\ from\ which\ it\ is\ calculated}} * \mathbf{100\%}$$

Numerical Examples: The marks of 10 students in statistics examination are as follows:

$$10,12,15,12,16,20,13,17,15,15$$

Find mean, median, mode, range and standard deviation.

Solution:

X_i	f_i	$f_i\ X_i$	$f_i(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$f_i(X_i - \overline{X})^2$
10	1	10	-4.5	20.25	20.25
12	2	24	-5	6.25	12.5
13	1	13	-1.5	2.25	2.25
15	3	45	1.5	0.25	0.75
16	1	16	1.5	2.25	2.25
17	1	17	2.5	6.25	6.25
20	1	20	5.5	30.25	30.25
Total	10	145		67.75	74.5

$$mean = \frac{145}{10} = 14.5$$
$$median = 15$$
$$mode = 15$$
$$range = 20 - 10 = 10$$
$$SD = \frac{74.5}{10} = 7.45$$

2 Skewness and kurtosis:

We have discussed measures of central tendency and measure of dispersion which describe the location and scale parameter of the data sets. They do not give any idea about the shape of the data structure. The measure of skewness and kurtosis illustrate the shape of the data sets. The measure of skewness gives the direction and the magnitude of the lack of symmetry and the measure of kurtosis gives the idea of the flatness of the curve.

2.2 Skewness

Skewness measures the degree of asymmetry of the data. Skewness refers to the lack of symmetry.

Skewness is mainly three types: Positive skewness, Negative skewness, and Symmetric data.

Positive Skewness:

A data is said to be positive skew if the long tail is on the right side of the peak. The mean is on the right of the peak value. Here Mean > Median > Mode.

Negative Skewness:

A data is said to be negative skew if the long tail is on the left side of the peak. The mean is on the left of the peak value. Here Mean < Median < Mode.

Symmetric

The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle. When data is symmetrically distributed, the left-hand side and right-hand side, contain the same number of observations. Here Mean = Median = Mode.

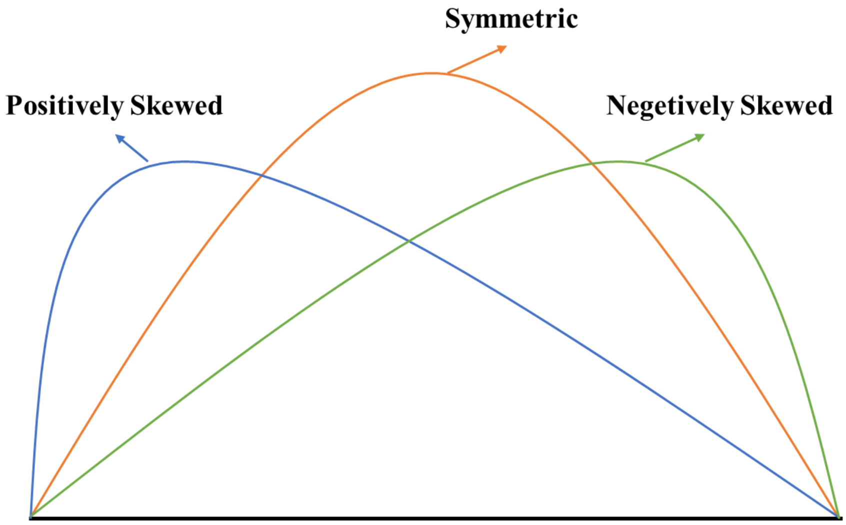


Figure 1. Skewness

The measure of Skewness:

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Interpretation:

1. If $S_k = 0$, then the frequency distribution is normal and symmetrical.

2. If $S_k > 0$, then the frequency distribution is positively skewed.

3. If $S_k < 0$, then the frequency distribution is negatively skewed.

2.3 Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails or outliers. Data sets with low kurtosis tend to have light tails or lack of outliers. A uniform distribution would be the extreme case.

Types of kurtosis: Leptokurtic or heavy-tailed distribution, Mesokurtic, Platykurtic or short-tailed distribution

Leptokurtic

Leptokurtic indicates that distribution is peaked and possesses thick tails.

Platykurtic

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is a flatter (less peaked) when compared with the normal distribution.

Mesokurtic

Mesokurtic is the same as the normal distribution. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

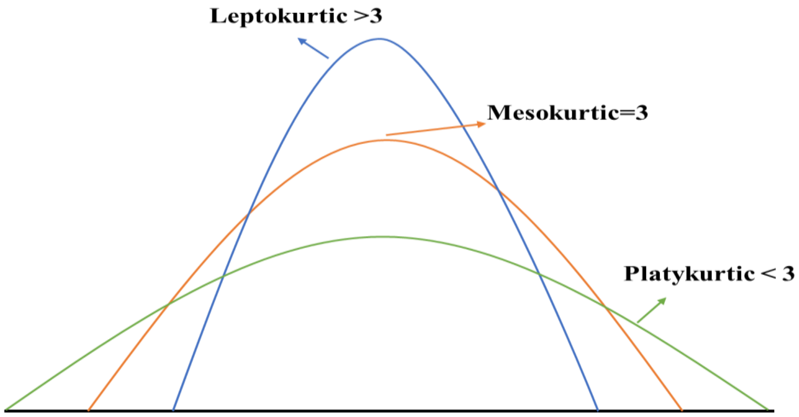


Figure 2. Kurtosis

Measurement of Kurtosis $(\beta_2) = \frac{1}{N-1}\frac{\sum(y_i-\bar{y})^4}{s^4}$

$\gamma_2 = \beta_2 - 3$

Data presentation

Non dimensional diagram	Pictograms
Two dimensional diagram	Bar diagram, Pie diagrams, Histograms, Box Plot
Three dimensional diagram	Cubes, Cylinders diagrams

There are three broad ways of presenting data. These are Textual presentation, Tabular presentation, and Graphic or diagrammatic presentation. We discussed only a few important diagrammatic presentations of data.

2.4 Bar Diagram

2.4.1 Simple Bar Diagram

If the classification is based on attributes and if the attributes are to be compared with respect to a single character we use a simple bar diagram. Simple bar diagrams consist of vertical bars of equal width. The heights of these bars are proportional to the volume or magnitude of the attribute. All bars stand on the same baseline. The bars are separated from each other by equal intervals. The bars may be colored or marked.

2.4.2 Multiple bar diagram

If the data is classified by attributes and if two or more characters or groups are to be compared within each attribute we use multiple bar diagrams. If only two characters are to be compared within each attribute, then the resultant bar diagram used is known as the double bar diagram. The multiple bar diagram is simply the extension of a simple bar diagram. For each attribute, two or more bars representing separate characters or groups are to be placed side by side. Each bar within an attribute will be marked or colored differently in order to distinguish them. The same type of marking or coloring should be done under each attribute. A footnote has to be given explaining the markings or colorings.

2.4.3 Component bar diagram

This is also called a subdivided bar diagram. Instead of placing the bars for each component side by side, we may place this one on top of the other. This will result in a component bar diagram.

2.5 Histogram

Histograms is suitable for continuous class frequency distribution. We mark off class intervals along the x-axis and frequencies (frequency density for unequal frequency data)along the y-axis.

- For equal class intervals, the heights of the rectangles will be proportional to the frequencies, while for unequal class intervals, the heights will be equal (or proportional) to the frequency densities.
- A frequency polygon is a line graph obtained by connecting the midpoints of the tops of the rectangles in the histogram.

Table 1. Differences between bar diagrams and histograms

Characteristics	**Bar Diagrams**	**Histograms**
Frequency is measured by	Height of the bar	Area of the bar
Gaps between the bars	Yes	No
Width of the bar	Equal	May not be equal
Data types	Discrete and Continuous	Continuous only

2.6 Pie diagrams

When we are interested in the relative importance of the different components of a single factor, we use pie diagrams. For the pie diagram, one circle is used and the area enclosed by it being taken as 100. Itis then divided into a number of sectors by drawing angles at the center, the area of each sector representing the corresponding percentage.

2.7 Box Plot

Minimum, maximum, and quartiles ($Q_{1,}$ Median, Q_3) together provide information on the center and variation of the variable in a nice compact way. Written in increasing order, they comprise what is called the five-number summary of the variable. A box plot is based on the five-number summary and can be used to provide a graphical display of the center and variation of the observed values of the variable in a data set. It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

N.B: Examples of graphical presentation have been given in our basic statistics with excel manual.

3 Robust Estimate of Mean and Standard Deviation

The mean and standard deviation provides a correct estimation only if the variable is normally distributed and without outliers. If the variable is skewed and/or has outliers, the mean and standard deviation will be excessively influenced by the extreme observations and provide faulty statistics of data. There are many alternatives to the mean and standard deviation. Alternatives to the mean include the well-known median and trimmed mean, Winsorized mean, and M-estimators and for standard deviation, the alternatives include the Inter-Quartile Range (IQR) and the Median Absolute Deviation (MAD), Trimmed standard deviation, the Winsorized standard deviation, and M-estimators. Median, IQR, MAD are already discussed in the previous section in detail. Here we only discussed the trimmed, Winsorized, and M estimators for mean and standard deviation.

3.2 Trimmed Mean and Standard Deviation

A trimmed mean and standard deviation is similar to a "regular" mean but it trims any outliers from both the side. To obtain the 20% trimmed mean, the 20% lowest and 20 % highest values are removed and the mean is computed on the remaining observations. In our example, these values will be: 4, 4, 5, 5, 6, 6, and the 20% trimmed mean will be equal to 5.

3.3 Winsorized Mean and Standard Deviation

The Winsorized technique is similar to the trimmed technique but the lowest (resp. highest) values are not removed but replaced by the lowest (resp. highest) untrimmed score. In our example, the values of the variables, also called Winsorized scores, will then be: 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, and the 20% Winsorized mean will be equal to 5.

3.4 M estimators

The trimmed mean all either take or drop observations. As for the Winsorized mean, it replaces values with less extreme values. In contrast, the M-estimators, weight each observation according to a function selected for its special properties. The weights depend on a constant that can be chosen by the researcher. The M-estimator solves this problem of assigning a zero value to many observations by down weighting the observations progressively. The only aspect of the M-estimator that could worry substantive researchers is that one must choose the degree of down weighting of the observations.

STATISTICAL DATA ANALYSIS USING MICROSOFT EXCEL

Sanchita Naha
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-12
Sanchita.naha@icar.gov.in

Statistics is the study of collection, analysis, interpretation, presentation, and organization of data. Broadly, two statistical methodologies are used for data analysis, descriptive statistics, and inferential statistics. Statistical analysis can be done using software like MS Excel, SPSS, R but this tutorial is restricted to major statistical analysis methods using Microsoft Excel. Statistical analysis mainly encompasses descriptive statistics and inferential statistics.

1. **Descriptive Statistics:** Descriptive statistics is used to describe or summarize data in a meaningful way. Descriptive statistics provides us with tools, tables, graphs, averages, ranges, correlations for organizing and summarizing data. In descriptive statistics data is summarized with the following major numerical descriptors like

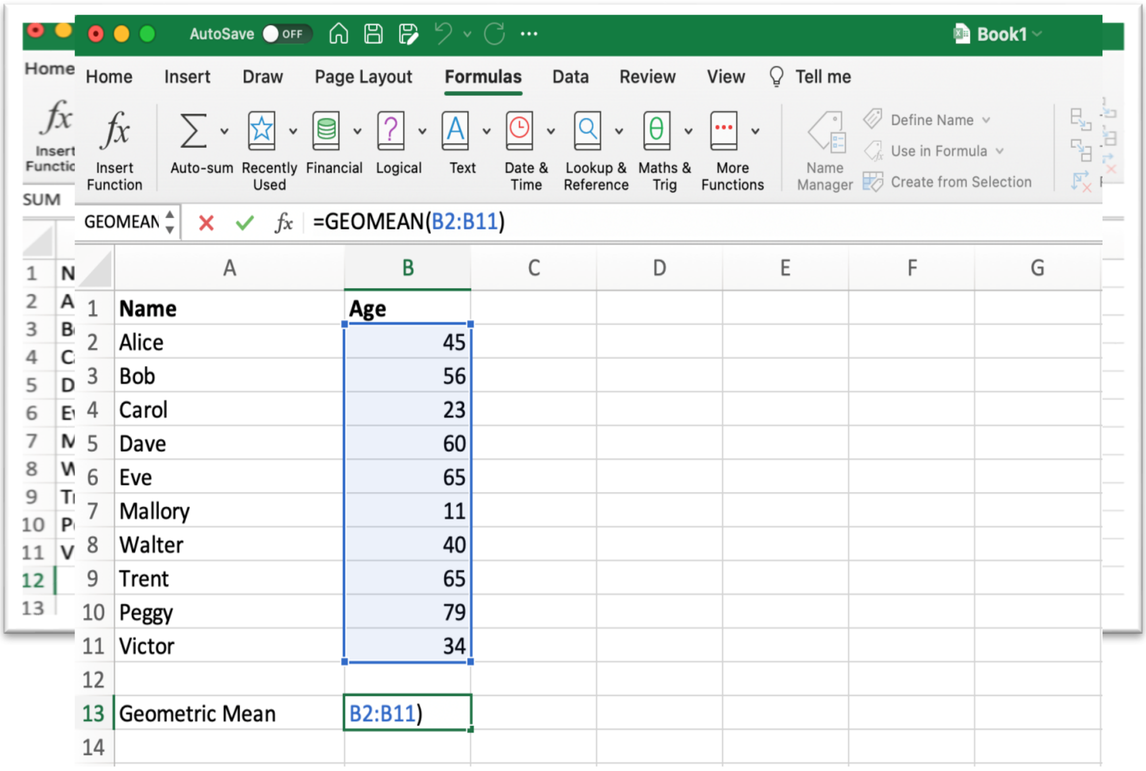
- **Arithmetic Mean:** It is defined as the average of the data values. For mean computation, data must be in numeric form.

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n}$$

n = number of observations

Steps to compute mean in Excel:

Select data points > Click formulas > Expand auto-sum drop down menu >



=GEOMEAN(B2:B11)

	A	B	C	D	E	F	G
1	Name	Age					
2	Alice	45					
3	Bob	56					
4	Carol	23					
5	Dave	60					
6	Eve	65					
7	Mallory	11					
8	Walter	40					
9	Trent	65					
10	Peggy	79					
11	Victor	34					
12							
13	Geometric Mean	B2:B11)					
14							

select Average > Enter.

- **Geometric Mean:** It is the n^{th} root of the product of individual data points. Let X_1, X_2, ... , X_n be the n^{th} observation of a data set. The geometric mean is defined as

$$GM = (x_1 * x_2 * \ldots * x_n)^{1/n}$$

Steps to compute geometric mean in Excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'GEOMEAN' > click 'Insert Function' > Enter.

e.g., GEOMEAN (B2:B11)

The geometric mean is used in finance to calculate average growth rates and is referred to as the compounded annual growth rate.

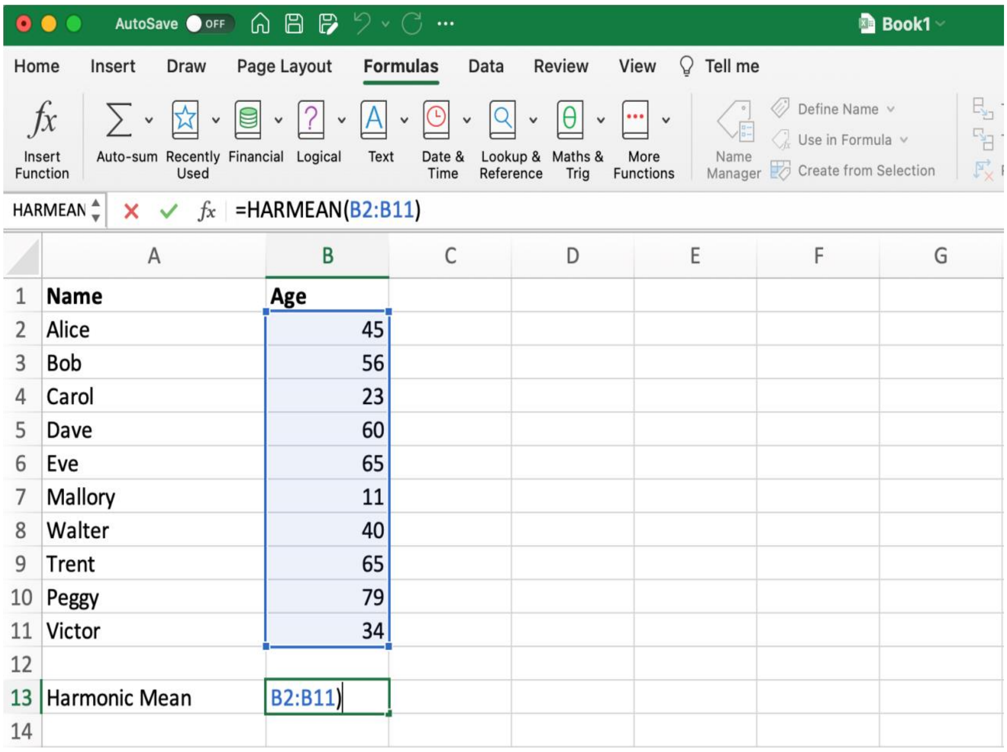
- **Harmonic Mean:** Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the observations of the datasets.

$$HM = \frac{n}{\sum_{i=1}^{n} {}^{1}/_{x_i}}$$

Steps to compute harmonic mean in Excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'HARMEAN' > click 'Insert Function' > Enter.

e.g., HARMEAN (B2:B11)



=HARMEAN(B2:B11)

	A	B	C	D	E	F	G
1	Name	Age					
2	Alice	45					
3	Bob	56					
4	Carol	23					
5	Dave	60					
6	Eve	65					
7	Mallory	11					
8	Walter	40					
9	Trent	65					
10	Peggy	79					
11	Victor	34					
12							
13	Harmonic Mean	B2:B11)					
14							

- **Median:** Median is the value in the middlemost position of all the observations when arranged in an ascending/descending order. The median is the central value of an ordered data series. It divides the data sets exactly into two parts. Fifty percent of observations are below the median value and 50% are above the median. Median is also known as 'positional average'.

  Steps to compute median in Excel:

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'MEDIAN' > click 'Insert Function' > Enter.

- **Mode:** Mode is defined as the value that occurs most frequently in the data. If in the data sets each observation occurs only once, then it does not have mode. When the data set has two or more values equal to the highest frequency than two or more mode are present in the datasets.

  Steps to compute median in Excel:

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'MODE' > click 'Insert Function' > Enter.

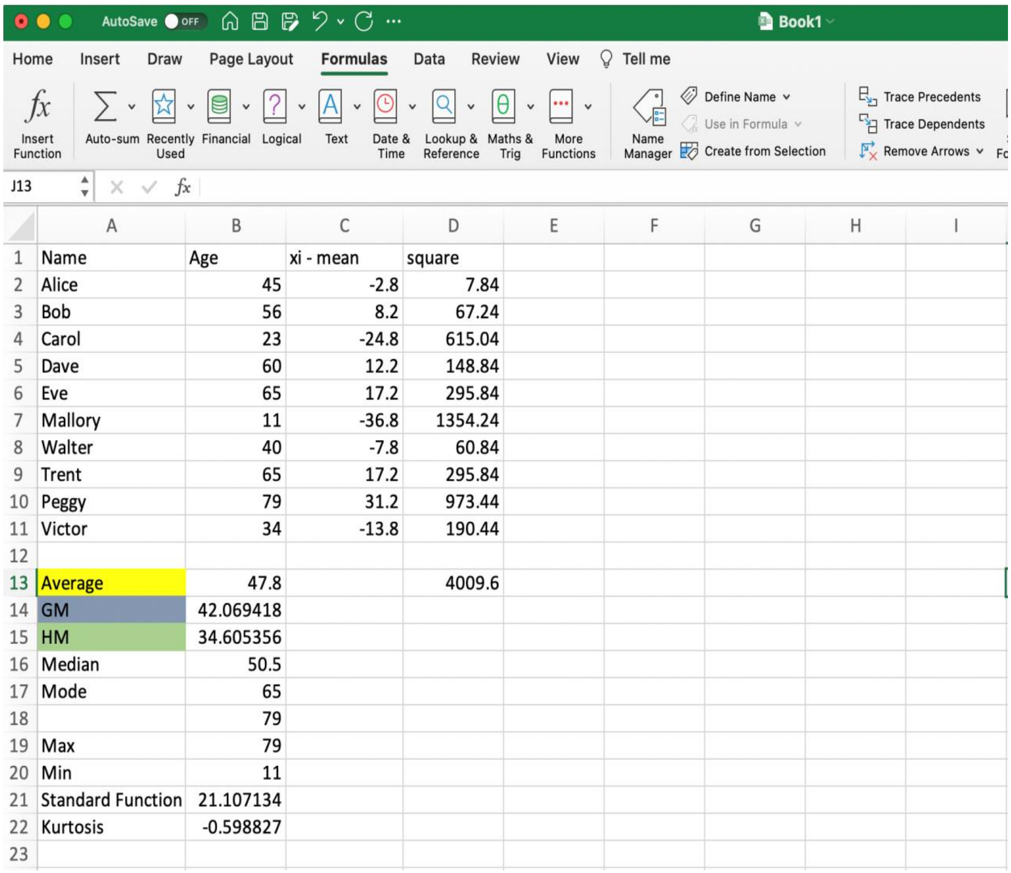
- **Range:** It is defined as the difference between the highest value and lowest value of the variable.

  $$Range = Maximum\ value - Minimum\ value$$

  Steps to compute range in Excel:

  Compute the maximum and minimum value among the data values. Then compute the difference between them to get the range of observations.
  Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'MAX' > click 'Insert Function' > Enter.
  Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'MIN' > click 'Insert Function' > Enter.

  Select a cell > write "=(specify the cell where maximum value is stored - specify the cell where minimum value is stored)" in the formula bar > Enter.



	A	B	C	D	E	F	G	H	I
1	Name	Age	xi - mean	square					
2	Alice	45	-2.8	7.84					
3	Bob	56	8.2	67.24					
4	Carol	23	-24.8	615.04					
5	Dave	60	12.2	148.84					
6	Eve	65	17.2	295.84					
7	Mallory	11	-36.8	1354.24					
8	Walter	40	-7.8	60.84					
9	Trent	65	17.2	295.84					
10	Peggy	79	31.2	973.44					
11	Victor	34	-13.8	190.44					
12									
13	Average	47.8		4009.6					
14	GM	42.069418							
15	HM	34.605356							
16	Median	50.5							
17	Mode	65							
18		79							
19	Max	79							
20	Min	11							
21	Standard Function	21.107134							
22	Kurtosis	-0.598827							
23									

- **Standard Deviation:** It is defined as the positive square-root of the arithmetic mean of the square of the deviations of the given observations from their arithmetic mean. It takes into consideration the magnitude of all the observations and gives the minimum value of dispersion possible. It is also known as Root Mean Square Deviation about mean.

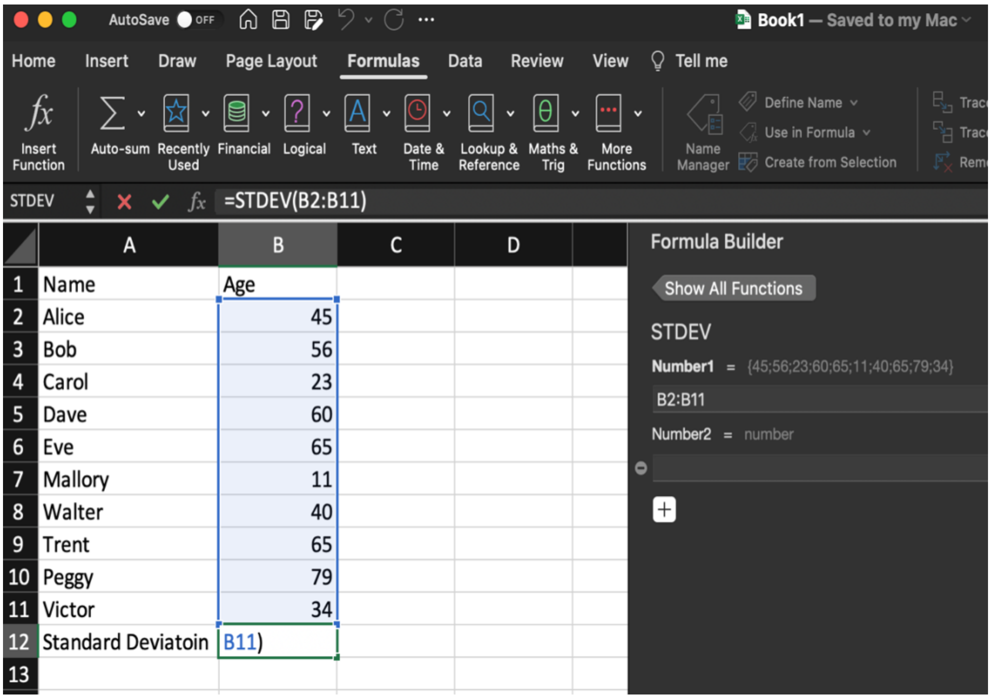
  Let x_1, x_2, …, x_n are the *n* observations in a data set. The standard deviation S.D. is given by,

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n}}$$

- **Variance:** It is defined as the square of the standard deviation. Unit of the variance is the square of the actual observations, whereas unit of the standard deviation is same as the actual observations.

  Steps to calculate Standard Deviation in excel:

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'STDEV' > click 'Insert Function' > Enter.



STDEV =STDEV(B2:B11)

	A	B	C	D
1	Name	Age		
2	Alice	45		
3	Bob	56		
4	Carol	23		
5	Dave	60		
6	Eve	65		
7	Mallory	11		
8	Walter	40		
9	Trent	65		
10	Peggy	79		
11	Victor	34		
12	Standard Deviatoin	B11)		
13				

STDEV

Number1 = {45;56;23;60;65;11;40;65;79;34}

B2:B11

Number2 = number

There are 6 versions of standard deviation formula available which are as following:

STDEV.S: This formula calculates the sample standard deviation based on numeric information alone. It ignores text and logical (TRUE or FALSE) values in the spreadsheet. The denominator in this case is *(n-1)*.

STDEV.P: This formula calculates the standard deviation for an entire population based on numeric information alone. It ignores text and logical values in the spreadsheet. The denominator in this case is *n*.

STDEVA: This formula calculates the sample standard deviation of a dataset but includes text and logical values in the calculation. All FALSE values are represented by 0, and TRUE values are represented by 1.

STDEVPA: This formula calculates the standard deviation for an entire population and includes text and logical values in the calculation. Like STDEVA, all FALSE values are represented by 0, and TRUE values are represented by 1.

STDEV: This is an older version of the STDEV.S formula that Excel used to calculate sample standard deviation before 2007. It still exists for compatibility purposes. This formula acts as the same as STDEV.S

STDEVP: This is an older version of the STDEV.P formula that still exists for compatibility.

Steps to calculate Variance in excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘STDEV’ > click ‘Insert Function’ > Enter.

- **Coefficient of Variation (CV):**The Coefficient of Variation (CV) is defined as the ratio of the standard deviation to the mean, and expressed in percentages,

$$CV = \frac{Standard\ Deviation}{Mean} * 100$$

CV is calculated to have an idea about the consistency/ variability of the series. Higher the CV means the series is more variable, less stable, less uniform, and less consistent. Lesser CV indicates that the series is less variable or more stable or more uniform and more consistent.

- **Skewness and Kurtosis:** Skewness is used to detect outliers in a data set**.** It characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.A data series is said to be positively skewed if the Mean of the data series is greater than Median and is greater than Mode. On the other hand data is said to be negatively skewed if Mean < Median < Mode. Data series is said to be symmetric if Mean = Median = Mode.

$$Pearson's\ Coefficient\ of\ Skewness = \frac{(Mean - Mode)}{Standard\ Deviation}$$

Alternate formula for computing Skewness Coefficient,

$$Coefficient\ of\ Skewness = \frac{3\ (Mean - Median)}{Standard\ Deviation}$$

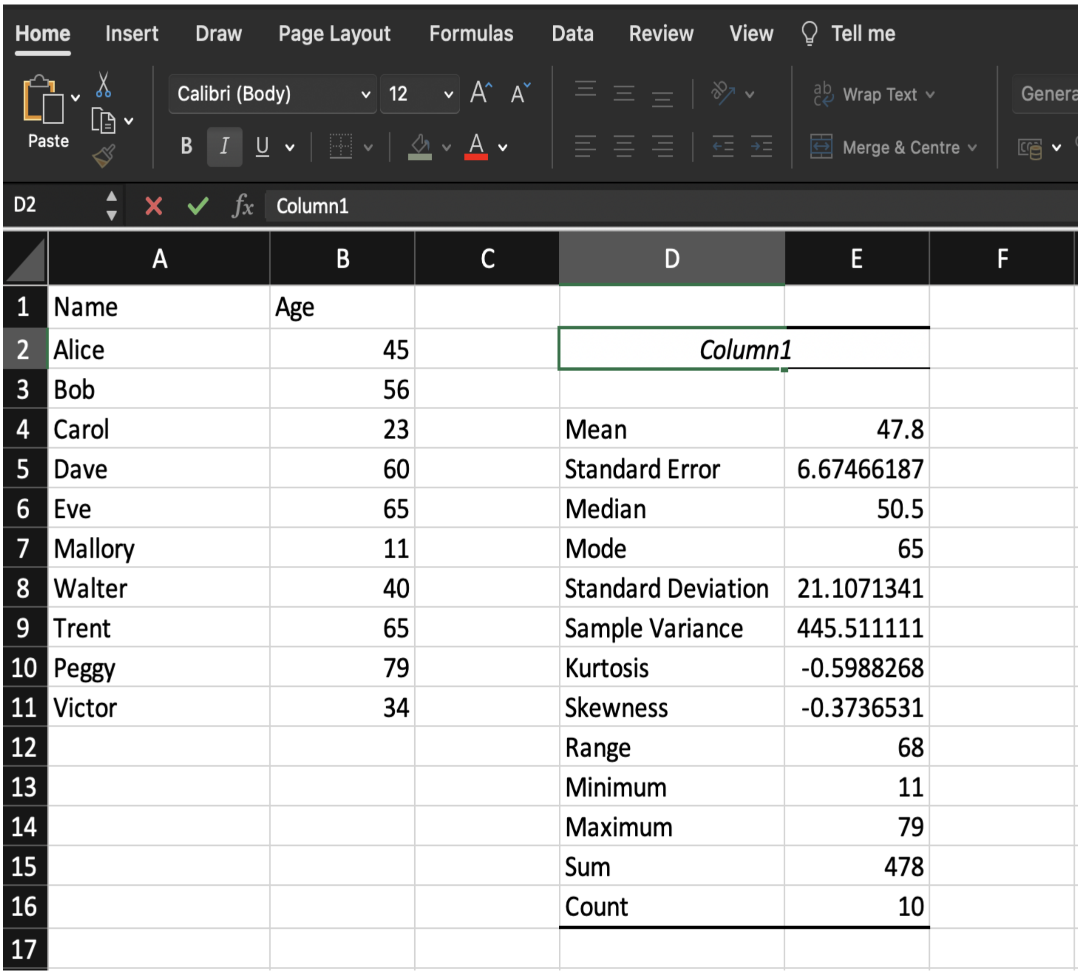
If Skewness coefficient = 0, then the distribution is normal and symmetrical.

If Skewness coefficient > 0, then the frequency distribution is positively skewed.

If Skewness coefficient < 0, then the frequency distribution is negatively skewed.

Steps to calculate Skewness Coefficient in excel:

Select data points > Click formulas > Expand auto-sum drop down menu > Find function ‘SKEW’ > click ‘Insert Function’ > Enter.



D2 Column1

	A	B	C	D	E	F
1	Name	Age				
2	Alice	45		*Column1*		
3	Bob	56				
4	Carol	23		Mean	47.8	
5	Dave	60		Standard Error	6.67466187	
6	Eve	65		Median	50.5	
7	Mallory	11		Mode	65	
8	Walter	40		Standard Deviation	21.1071341	
9	Trent	65		Sample Variance	445.511111	
10	Peggy	79		Kurtosis	-0.5988268	
11	Victor	34		Skewness	-0.3736531	
12				Range	68	
13				Minimum	11	
14				Maximum	79	
15				Sum	478	
16				Count	10	
17						

- **Kurtosis:** Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. It is the tailedness of a distribution relative to a normal distribution. Distributions with medium kurtosis (medium tails) are mesokurtic, with low kurtosis are called platykurtic, and distributions with high kurtosis are leptokurtic.

  $$\text{Measure of kurtosis, } \gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

  Kurtosis value equals to 3.0 indicates, the data distribution is mesokurtic, for kurtosis value greater than 3.0, it is called leptokurtic and for a lesser value than 3.0 the distribution is called platykurtic.

  Steps to calculate Kurtosis in excel:

  Select data points > Click formulas > Expand auto-sum drop down menu > Find function 'KURT' > click 'Insert Function' > Enter.

  Excel provides an "*Analysis Tool Pak*" add-in under the *Data* tab to generate a report of the Descriptive Statistics on the desired data.

For example, we have examination scores of 10 students in a class like the following. To generate descriptive statistics for these scores, follow the steps below.

Step 1: On the Data tab, in the Analysis group, click Data Analysis.

Step 2: Select Descriptive Statistics and click OK.

Step 3: Select the range B2:B11 as the Input Range.

Step 4: Select cell C1 as the Output Range.

Step 5: Make sure Summary statistics is checked.

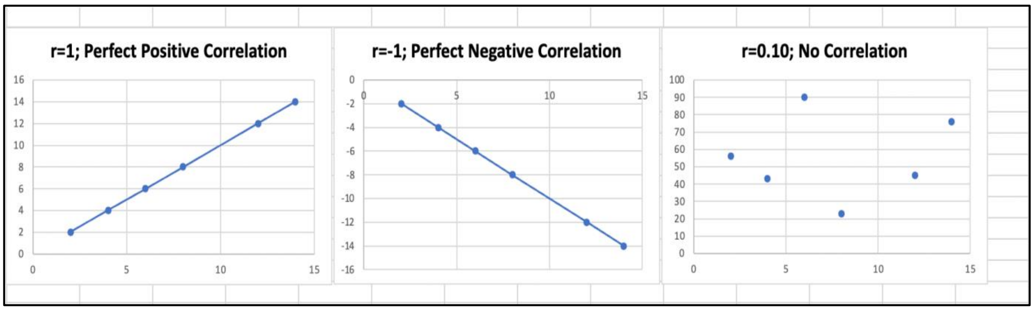
Step 6: Click ok.

2. **Correlation and Regression Analysis:** Correlation is the measurement of linear association between two variables. It is a measure that describes the strength and direction of a relationship between two variables. It is a commonly used measure in statistics, economics and social sciences for budgets, business plans etc. The correlation coefficient is used to measure the correlation between bivariate data which basically denotes the degree of linear association between two random variables.

   In statistics, there are several types of correlation measures depending on the type of data you are working with. Here, we will focus on the most common one.

   Pearson Product Moment Correlation (PPMC), popularly called as Pearson Correlation is used to evaluate linear relationships between data when a change in one variable is associated with a proportional change in the other variable.

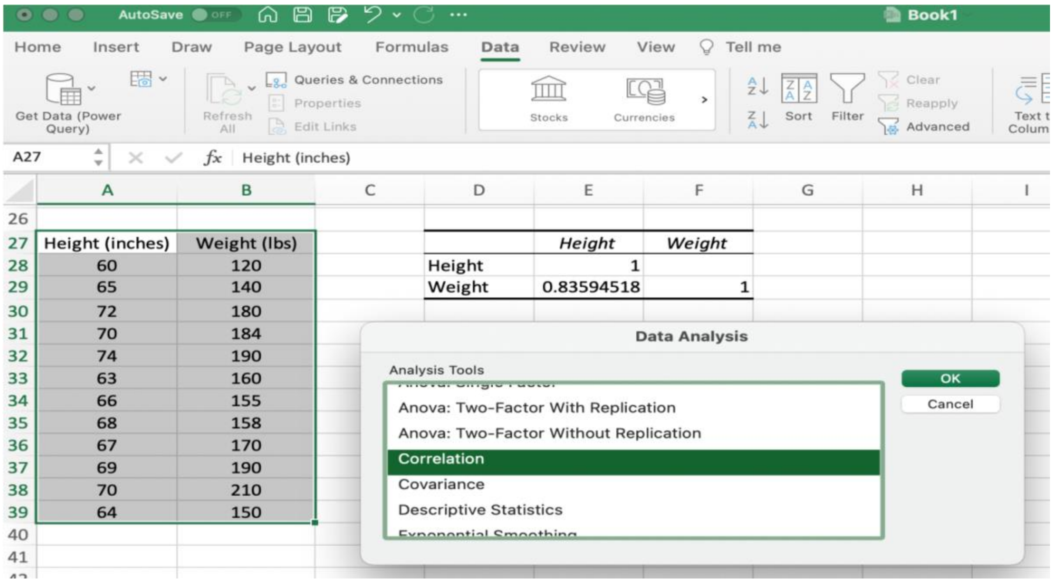
$$\textbf{Pearson Correlation Coefficient, } r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 * \sum(y_i-\bar{y})^2}}$$



The correlation coefficient value always lies between -1 and 1 and it measures both the strength and direction of the linear relationship between the variables. Correlation coefficient of +1 means a perfect positive relationship, as value of one variable increases, value of other variable increases proportionally. Correlation coefficient value of -1 means a perfect negative relationship, with increase in the value of one variable, the other one decreases proportionally. A coefficient of 0 means no linear relationship between the two variables the data points are scattered all over the graph.

Steps to calculate Pearson Correlation Coefficient in Excel:

Select '*Data'* tab > click '*Data Analysis*' > Find Correlation from the given menus > Click ok > Select the input range > select output cell > Grouped by columns > click ok.



Height (inches)	Weight (lbs)
60	120
65	140
72	180
70	184
74	190
63	160
66	155
68	158
67	170
69	190
70	210
64	150

	Height	*Weight*
Height	1	
Weight	0.83594518	1

Regression analysis is used to estimate the relationship between two or more variables. Dependent variable is the main factor you want to study, understand, or predict. Independent variables are the factors that might influence the dependent variable. Regression analysis helps to understand how the dependent variable changes when one of the independent variables vary. Regression analysis can make it easier to predict future variable trends by analyzing the trajectory of the regression line. Simple linear regression model tries to establish a linear association between the dependent and the independent variable so that the outcome of the dependent variable can be predicted using the independent variables. The simple linear regression model uses the following equation:

$$Y = a + bX + \epsilon$$

where, Y = value of the dependent variable

X = value of the independent variable

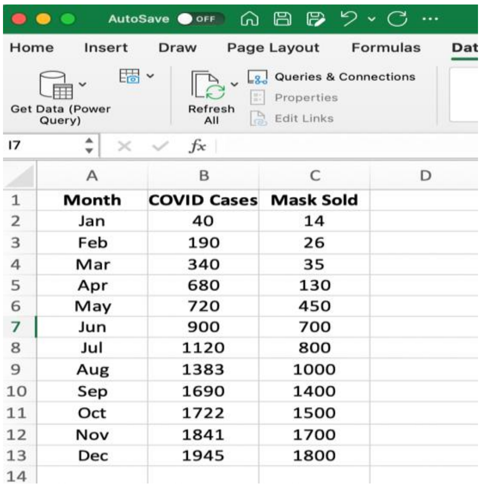
a = intercept

b = slope (regression line steepness)

ϵ = error component

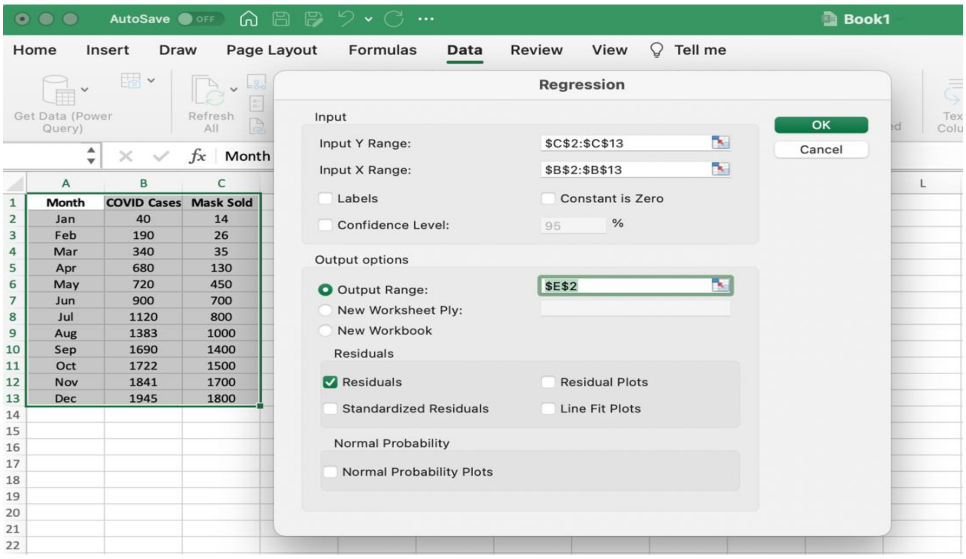
Steps to perform Regression Analysis in Excel:

Step1: Let us consider the data values for the following two variables, COVID cases and masks sold and perform a simple linear regression analysis in Excel considering number of Masks sold as the Y variable and number of COVID cases as X variable on which Y is dependent.



	A	B	C	D
1	Month	COVID Cases	Mask Sold	
2	Jan	40	14	
3	Feb	190	26	
4	Mar	340	35	
5	Apr	680	130	
6	May	720	450	
7	Jun	900	700	
8	Jul	1120	800	
9	Aug	1383	1000	
10	Sep	1690	1400	
11	Oct	1722	1500	
12	Nov	1841	1700	
13	Dec	1945	1800	
14				

Step2: Click on the 'Data' tab > Data Analysis > Select 'Regression' >click 'Ok'.



Month

	A	B	C
1	Month	COVID Cases	Mask Sold
2	Jan	40	14
3	Feb	190	26
4	Mar	340	35
5	Apr	680	130
6	May	720	450
7	Jun	900	700
8	Jul	1120	800
9	Aug	1383	1000
10	Sep	1690	1400
11	Oct	1722	1500
12	Nov	1841	1700
13	Dec	1945	1800

Regression

Input

Input Y Range: $C$2:$C$13

Input X Range: $B$2:$B$13

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $E$2

○ New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

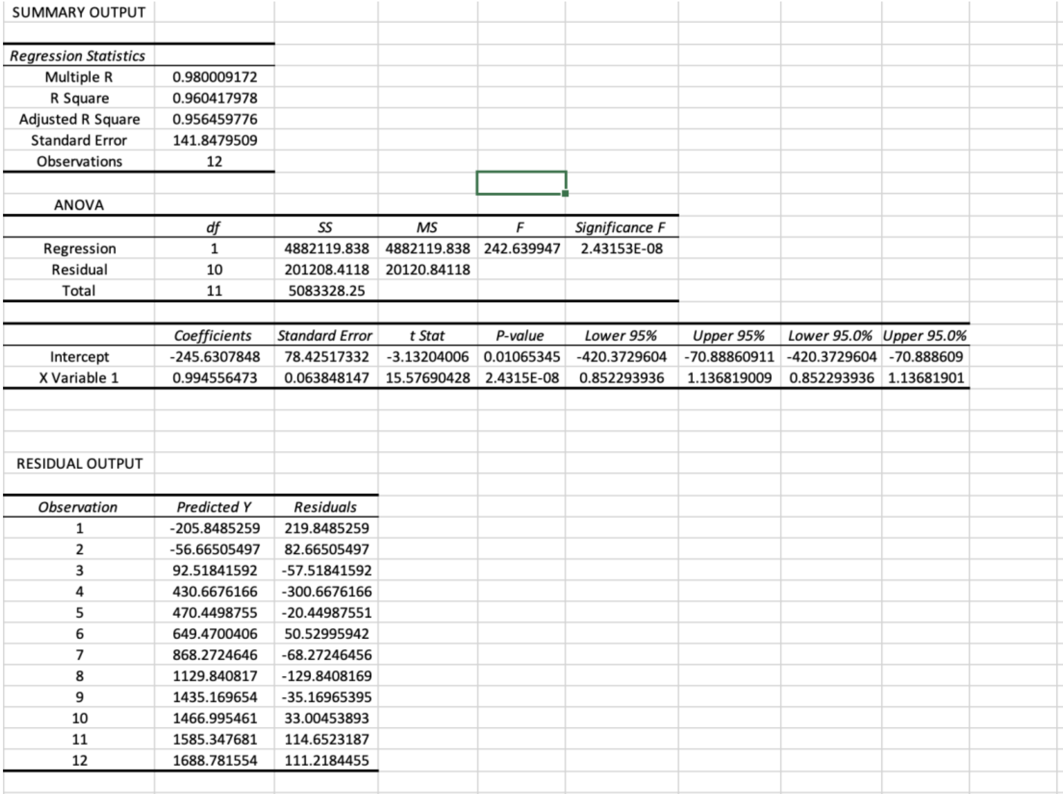
Normal Probability

□ Normal Probability Plots

Step3: In the Regression dialog box select the Input Y Range, which is our dependent variable. In this case it is (C2:C13). Then select the Input X Range, independent variable. In this example, it is the number of COVID cases (B2:B13). Select the desired output range, here E2.

Click ok.

You get the following Output:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980009172
R Square	0.960417978
Adjusted R Square	0.956459776
Standard Error	141.8479509
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4882119.838	4882119.838	242.639947	2.43153E-08
Residual	10	201208.4118	20120.84118		
Total	11	5083328.25			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-245.6307848	78.42517332	-3.13204006	0.01065345	-420.3729604	-70.88860911	-420.3729604	-70.888609
X Variable 1	0.994556473	0.063848147	15.57690428	2.4315E-08	0.852293936	1.136819009	0.852293936	1.13681901

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	-205.8485259	219.8485259
2	-56.66505497	82.66505497
3	92.51841592	-57.51841592
4	430.6676166	-300.6676166
5	470.4498755	-20.44987551
6	649.4700406	50.52995942
7	868.2724646	-68.27246456
8	1129.840817	-129.8408169
9	1435.169654	-35.16965395
10	1466.995461	33.00453893
11	1585.347681	114.6523187
12	1688.781554	111.2184455

- **Interpreting the Out putof Regression Analysis:**
  SUMMARY STATISTICS

  Multiple R is the value of the Correlation Coefficient that measures the strength of a linear relationship between two variables. The larger the absolute value, the stronger the relationship.

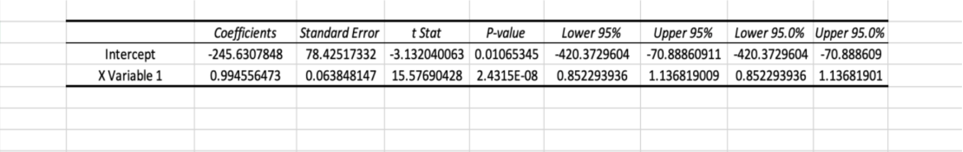
  R Square gives the Coefficient of Determination, which is used as an indicator of the goodness of fit. It shows how many points fall on the regression line. The R2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared deviations of the original data from the mean. In this example, R2 is 0.96, which is very good. It means that 96% of our values fit the regression analysis model. In other words, 96% of the dependent variables (y-values) are explained by the independent variables (x-values). Generally, R Squared of 95% or more is considered a good fit.

Adjusted R Square gives the R square adjusted for the number of independent variables in the model. For multiple regression analysis, adjusted R square value is used instead of R square.

Standard Error is another goodness-of-fit measure that shows the precision of the fitted regression model. The smaller the number, the more certain one can be about the regression equation. It is an absolute measure that shows the average distance that the data points fall from the regression line.

Observations simply provides the total number of observations used to fir the model.

COEFFICIENTS

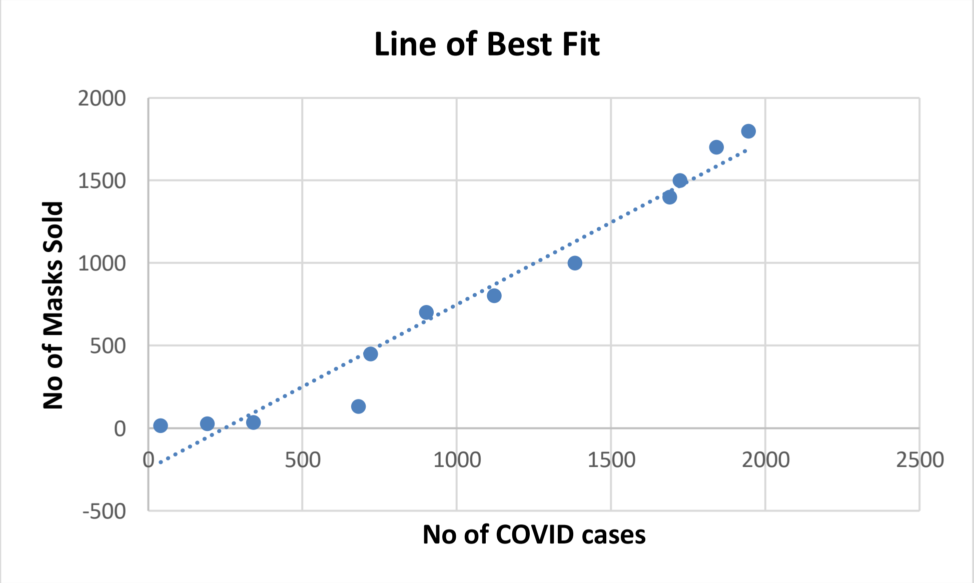


	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-245.6307848	78.42517332	-3.132040063	0.01065345	-420.3729604	-70.88860911	-420.3729604	-70.888609
X Variable 1	0.994556473	0.063848147	15.57690428	2.4315E-08	0.852293936	1.136819009	0.852293936	1.13681901

Linear regression equation fitted was, Y = b*X + a

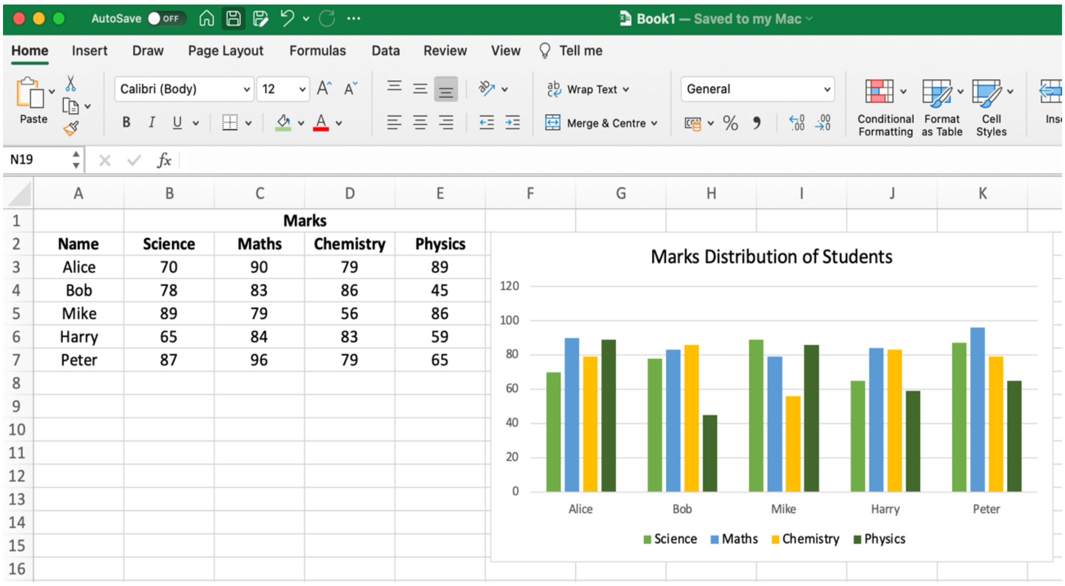
Here, Y = Mask sold; X = COVID cases; b = 0.99; a = -245.63

Therefore, 0.99 * 190 – 245.63 = -57.53



3. Create Charts/ Graphs in MS Excel:

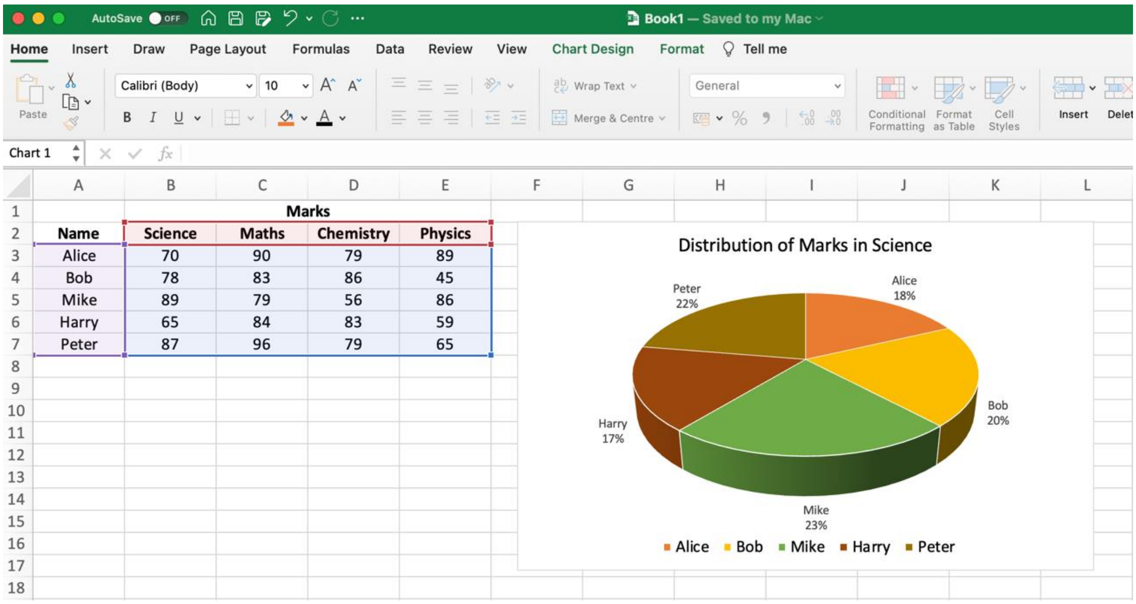
Line Diagram: Select the data for which you want to plot the graph. Click ‘Insert’ tab > go to insert column chart > pick any chart of your preference. Excel will create the graphical representation as following.



	Marks			
Name	Science	Maths	Chemistry	Physics
Alice	70	90	79	89
Bob	78	83	86	45
Mike	89	79	56	86
Harry	65	84	83	59
Peter	87	96	79	65

Pie chart: Pie chart represents the data in slices of a circle. Each slice represents the percentage contribution of each data section among the sum of individual data values.

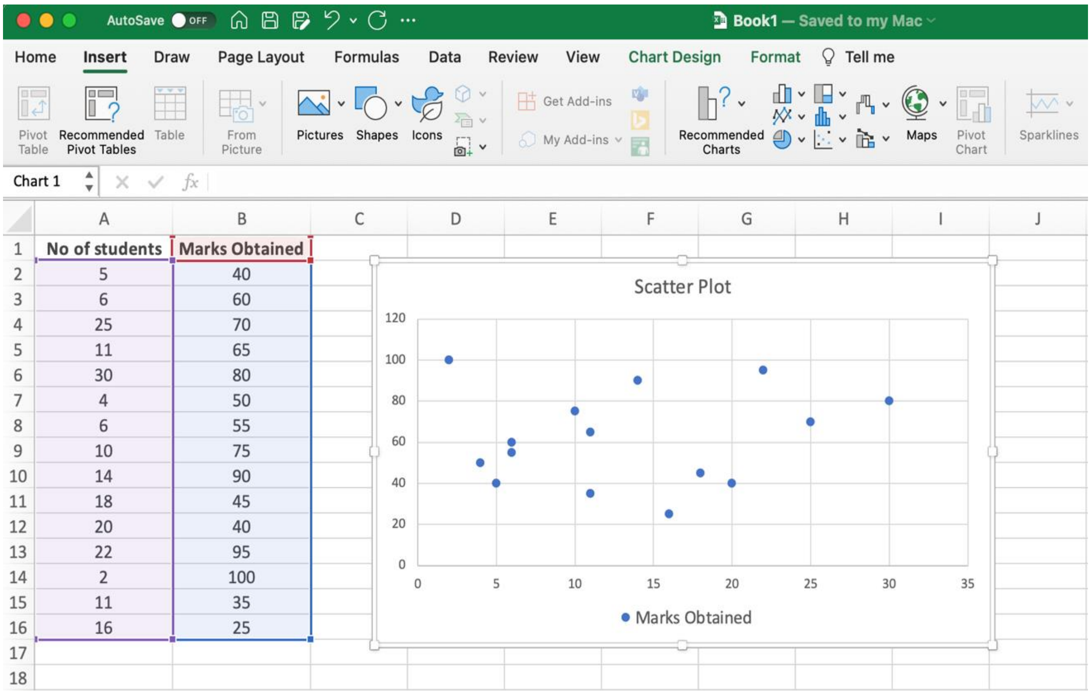
Select the data for which you want to plot the pie chart. Click insert tab > go to insert pie or doughnut chart > pick any chart of your preference. Excel will create the graphical representation as following:



	Marks			
Name	Science	Maths	Chemistry	Physics
Alice	70	90	79	89
Bob	78	83	86	45
Mike	89	79	56	86
Harry	65	84	83	59
Peter	87	96	79	65

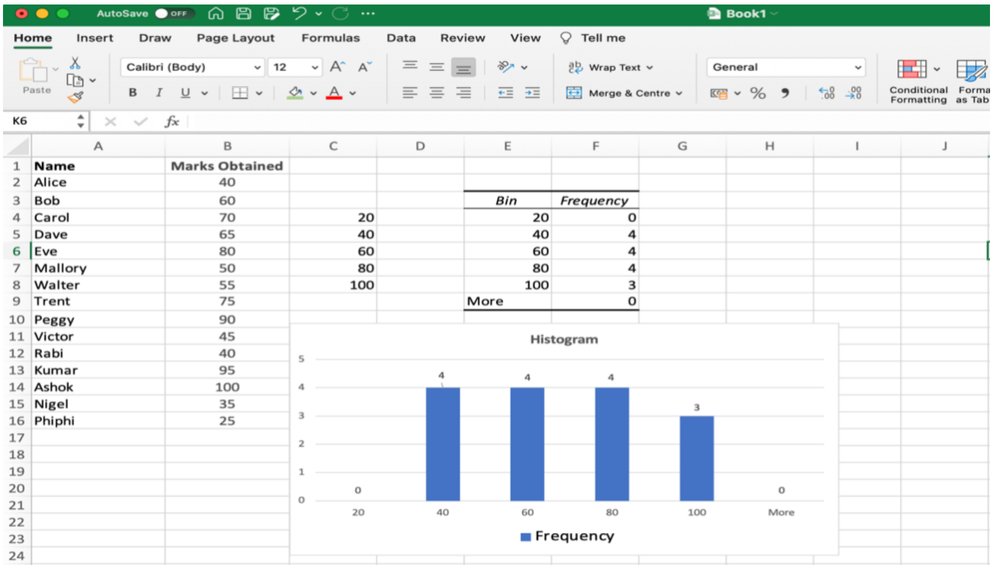
Scatter Diagram: Scatter charts are specifically used to show how one variable is related to another. There are seven scatter chart options: scatter, scatter with smooth lines and markers, scatter with smooth lines, scatter with straight lines and markers, scatter with straight lines, bubble, and 3-D bubble. For plotting a scatter chart, one needs data points for two or more variables.

Select the data> click insert tab > go to X Y Scatter chart > pick any chart of your preference. Excel will create the graphical representation as following:



	A	B
1	No of students	Marks Obtained
2	5	40
3	6	60
4	25	70
5	11	65
6	30	80
7	4	50
8	6	55
9	10	75
10	14	90
11	18	45
12	20	40
13	22	95
14	2	100
15	11	35
16	16	25
17		
18		

Histogram: Select data >click on data tab > select data Analysis >click histogram > select



	A	B	C	D	E	F
1	Name	Marks Obtained				
2	Alice	40				
3	Bob	60			*Bin*	*Frequency*
4	Carol	70	20		20	0
5	Dave	65	40		40	4
6	Eve	80	60		60	4
7	Mallory	50	80		80	4
8	Walter	55	100		100	3
9	Trent	75			More	0
10	Peggy	90				
11	Victor	45				
12	Rabi	40				
13	Kumar	95				
14	Ashok	100				
15	Nigel	35				
16	Phiphi	25				

input range (B2:B16)> select bin (class intervals, here it is C4:C8) > check Chart Output > click ok. Excel will produce the frequency table against the specified bin value and also will create a histogram diagram like following.

4. **Inferential Statistics:**

Inferential statistics is used for estimating the population data by analysing the samples obtained from it. It helps in making generalizations about the population by using different analytical tests and tools. Various sampling techniques are usedto select random samples that will represent the population accurately. Some of the important methods are simple random sampling, stratified sampling, cluster sampling, and systematic sampling techniques.

Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. In inferential statistics, a statistic is taken from the sample data (e.g., sample mean) that used to make inferences about the population parameter (e.g., the population mean). One sample t-test is the most commonly used one and sets a basic understanding of all other kinds of hypothesis testing methods.

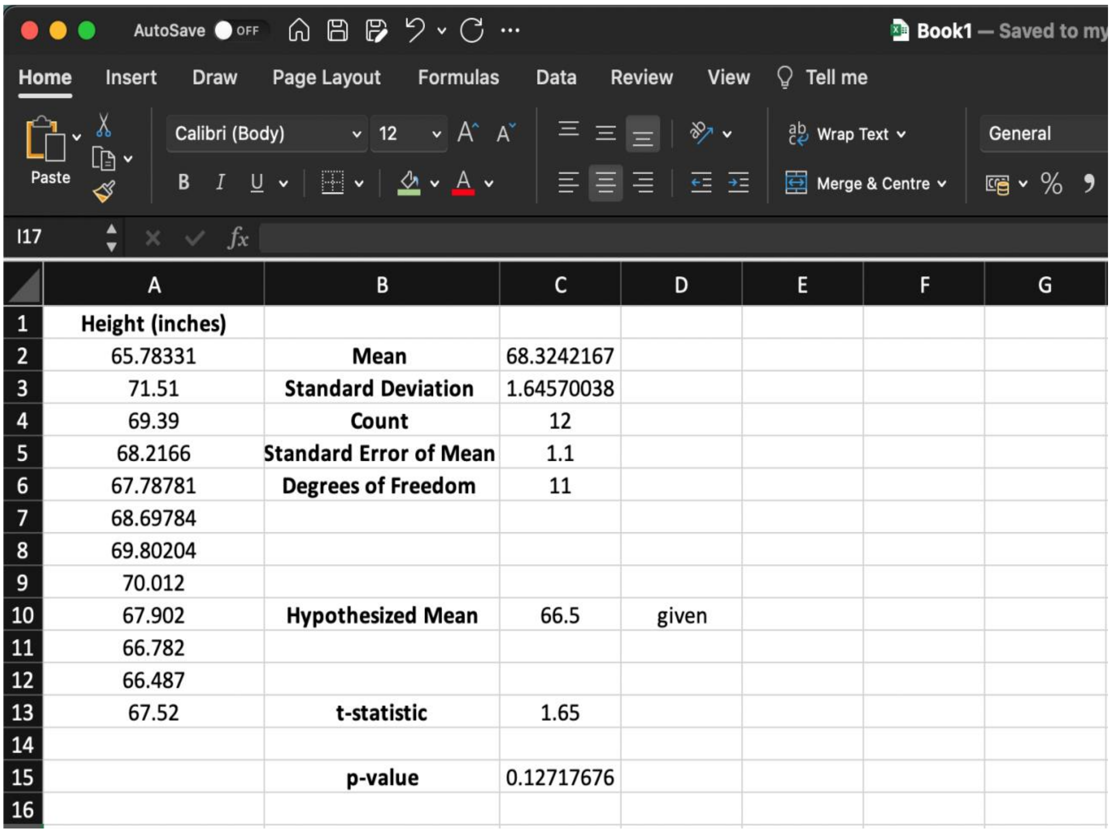
One sample *t*-test:

The one-sample t test compares a given sample mean $\bar{X}$ to a known or hypothesized value of the population mean $\mu_0$provided the population standard deviation σ is unknown. Excel does not have a built-in one-sample t test. However, the use of Excel functions and formulas makes the computations quite simple. The value of *t*-statistic can be calculated from the given formula:

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{x}}}$$

where, $\bar{X}$ is the sample mean, μ_0 is the known or hypothesized population mean and $s_{\bar{x}}$ isthe standard error of mean.To calculate the *t*-statistic in excel we need to first find the following values.

Consider a sample of 12 young female adults, we have the measurement of their heights in inches. Let us assume the national average height of 18-year-old girls is 66.5 inches. We want to perform a one-sample T-test in Excel to determine if there is any significant difference between the heights of the sample data compared with the national average height (66.5 inches).



	A	B	C	D	E	F	G
1	Height (inches)						
2	65.78331	Mean	68.3242167				
3	71.51	Standard Deviation	1.64570038				
4	69.39	Count	12				
5	68.2166	Standard Error of Mean	1.1				
6	67.78781	Degrees of Freedom	11				
7	68.69784						
8	69.80204						
9	70.012						
10	67.902	Hypothesized Mean	66.5	given			
11	66.782						
12	66.487						
13	67.52	t-statistic	1.65				
14							
15		p-value	0.12717676				
16							

The null hypothesis and alternative hypothesis for this test are:

Null hypothesis: There is no significant difference between the heights of the sample, compared with the national average.

Alternative hypothesis: There is significant difference between the heights of the sample, compared with the national average.

First of all, compute mean, standard deviation, standard error, degrees of freedom to calculate the value of the *t*-statistic as shown in the above screenshot then in an empty cell, enter =TDIST (t, df, tails) to compute the p-value.

t – the cell containing the t-statistic

df – The cell containing the degrees of freedom.

tails –1if you want to perform a one-tailed analysis, or 2 if you want to do a two-tailed analysis.p-value for this example is 0.127.

Let us assume alpha level is set at 0.05, then since the p-value is above the alpha level, we will accept the null hypothesis and reject the alternative hypothesis.In other words, there is no significant difference between the heights of the sample, compared with the national average.

TESTS OF SIGNIFICANCE AND NON-PARAMETRIC TEST

Rajeev Ranjan Kumar
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
Rajeev.kumar4@icar.gov.in

In the realm of statistics, the test of significance, also known as hypothesis testing, is a powerful tool used to make informed decisions about population parameters based on sample data. It enables researchers and analysts to assess the validity of assumptions, draw conclusions, and determine the level of confidence in their findings.

The fundamental idea behind the test of significance is to evaluate whether the observed data is strong enough to support or reject a particular hypothesis about a population characteristic. This hypothesis is typically formulated in terms of a null hypothesis (H_0), which assumes no significant difference or relationship, and an alternative hypothesis (H_1), which posits the existence of a meaningful difference or relationship.

To conduct a test of significance, a sample is collected from the population of interest, and relevant statistical techniques are employed to analyze the data. The results are then used to evaluate the likelihood of observing the sample data under the assumption that the null hypothesis is true. If the observed data is highly improbable under this assumption, it provides evidence to reject the null hypothesis in favour of the alternative hypothesis.

The test of significance involves determining a test statistic, which summarizes the data and allows for comparison against a theoretical distribution. The choice of the appropriate test statistic depends on the nature of the research question and the type of data being analyzed. Commonly used test statistics include the z-score, t-statistic, chi-square statistic, and F-statistic, among others.

1.Types of Hypotheses

In scientific research, a hypothesis is a proposed explanation or prediction about a phenomenon or relationship between variables. Hypotheses play a crucial role in guiding research and formulating testable statements that can be supported or refuted by empirical evidence. Depending on the nature of the research question and the specific objectives of the study, different types of hypotheses can be formulated. Here are some common types of hypotheses:

Null Hypothesis (H_0): The null hypothesis represents the absence of an effect, relationship, or difference between variables. It assumes that there is no statistically significant relationship or change in the population being studied. Researchers generally aim to reject the null hypothesis in favour of an alternative hypothesis. For example, the null hypothesis could state that there is no difference in test scores between two groups of students.

Alternative Hypothesis (H_1): The alternative hypothesis is the opposite of the null hypothesis. It suggests that there is a significant effect, relationship, or difference between variables in the population. Researchers seek to gather evidence to support the alternative hypothesis. Building upon the previous example, the alternative hypothesis could state that there is a difference in test scores between the two groups of students.

Directional Hypothesis: A directional hypothesis predicts the specific direction of the relationship or difference between variables. It specifies whether the effect will be positive or negative. For instance, a directional hypothesis may state that Group A will have higher test scores than Group B or that an increase in temperature will lead to a decrease in plant growth. Directional hypotheses are often used when previous research or theoretical considerations provide a basis for predicting the direction of the effect.

Non-Directional Hypothesis: Also known as a two-tailed hypothesis, a non-directional hypothesis does not predict a specific direction of the relationship or difference. It simply states that there is a significant difference or relationship between variables without specifying the direction. Researchers use non-directional hypotheses when they do not have a clear theoretical basis or prior evidence to suggest a specific direction. For example, a non-directional hypothesis may state that there is a difference in test scores between two groups of students, without specifying which group will perform better.

Composite Hypothesis: A composite hypothesis consists of multiple statements or conditions. It encompasses more than one possibility and allows for different outcomes. Composite hypotheses are often used when there are multiple factors or variables involved in the research question. For instance, a composite hypothesis could state that the effect of a particular treatment on patient outcomes varies depending on age, gender, and socioeconomic status.

Simple Hypothesis: In contrast to composite hypotheses, simple hypotheses involve a single statement or condition. They are straightforward and make specific predictions about a single variable or relationship. Simple hypotheses are commonly used when the research question focuses on a single factor or variable. For example, a simple hypothesis could state that there is a positive correlation between study time and exam scores.

2. Types of Errors

Errors can occur due to various sources of uncertainty and can impact the validity and reliability of research findings. Understanding the types of errors is essential for researchers and analysts to properly interpret and draw accurate conclusions from their data. Here are the two primary types of errors in statistics:

(A) Type I Error

Type I error, also known as a false positive, occurs when the null hypothesis (H_0) is mistakenly rejected, indicating the presence of a significant effect or relationship when, in fact, none exists in the population. It represents the probability of observing a statistically significant result due to random chance alone. Type I error is typically denoted by the symbol α (alpha) and is related to the significance level chosen for the hypothesis test.

For example, let's say a researcher conducts a study to determine if a new drug is effective in reducing blood pressure. The null hypothesis states that the drug has no effect. If the researcher rejects the null hypothesis and concludes that the drug is effective when it is actually not, it would be a Type I error. The researcher would have falsely claimed a significant effect.

The significance level chosen for the hypothesis test determines the threshold at which a Type I error is considered acceptable. A lower significance level (e.g., $\alpha = 0.05$) reduces the risk of Type I error but increases the chance of Type II error.

(B) Type II Error:

Type II error, also known as a false negative, occurs when the null hypothesis (H_0) is incorrectly accepted, implying no significant effect or relationship, even when there is one in the population. It represents the failure to detect a true effect or relationship. Type II error is denoted by the symbol β (beta) and is related to the statistical power of the test.

Building upon the previous example, if the researcher fails to reject the null hypothesis and concludes that the drug is not effective, even though it is, it would be a Type II error. The researcher would have missed detecting a real effect.

Type II error is influenced by factors such as the sample size, effect size, variability in the data, and the chosen significance level. To minimize the risk of Type II error, researchers often aim to maximize the statistical power of their study by using larger sample sizes, employing more sensitive measurement techniques, or increasing the significance level.

It's important to note that Type I and Type II errors are inversely related: reducing one type of error increases the likelihood of the other. Researchers need to strike a balance between these two types of errors based on the consequences of each in the specific research context.

(3)Level of Significance in Statistics:

In statistical hypothesis testing, the level of significance, often denoted by the symbol α (alpha), is a predetermined threshold that helps researchers make decisions about the validity of their results. It represents the maximum allowable probability of making a Type I error (rejecting the null hypothesis when it is actually true). The level of significance plays a crucial role in determining the critical region and the acceptance or rejection of the null hypothesis.

The most commonly used level of significance in many fields of research is 0.05 (or 5%). This means that if the calculated probability (p-value) of obtaining the observed data under the null hypothesis is equal to or less than 0.05, the null hypothesis is rejected in favour of the alternative hypothesis. In other words, researchers conclude that there is sufficient evidence to suggest that a relationship, effect, or difference exists in the population being studied. However, the choice of the level of significance is not arbitrary and should be determined based on the specific research question, the consequences of Type I and Type II errors, and the desired level of confidence. Commonly used levels of significance include 0.01 (1%) and 0.10 (10%), depending on the context and the stringency of the decision-making process.

A lower level of significance (e.g., 0.01) reduces the risk of Type I error, providing a more conservative approach to hypothesis testing. It requires stronger evidence to reject the null hypothesis and provides a higher level of confidence in the conclusions drawn from the data. On the other hand, a higher level of significance (e.g., 0.10) increases the risk of Type I error, making it easier to reject the null hypothesis. This approach is less conservative and may be appropriate when the consequences of Type II error are more severe or when exploratory analysis is conducted. It's important to note that the level of significance does not directly indicate the magnitude or practical importance of the observed effect. It solely reflects the strength of evidence against the null hypothesis. Therefore, researchers need to carefully interpret the results in the context of the specific research question and consider the practical implications of their findings.

(4) P-value

In statistical hypothesis testing, the p-value is a measure that helps researchers assess the strength of evidence against the null hypothesis (H_0) and make informed decisions about its rejection or acceptance. The p-value represents the probability of obtaining the observed data, or more extreme data, if the null hypothesis were true. The calculation of the p-value involves comparing the observed test statistic (e.g., t-statistic, z-score, chi-square statistic) with the distribution of the test statistic under the assumption that the null hypothesis is true. The p-value provides a quantitative measure of the likelihood of observing the data under the null hypothesis.

Interpreting the p-value is based on a chosen level of significance (α) that represents the threshold for rejecting the null hypothesis. If the p-value is smaller than the chosen level of significance, typically 0.05 (or 5%), it is considered statistically significant, and the null hypothesis is rejected. This indicates that the observed data is unlikely to occur by random chance alone and provides evidence in favour of the alternative hypothesis (H_1).On the other hand, if the p-value is larger than the chosen level of significance, the null hypothesis is not rejected. This suggests that the observed data is reasonably likely to occur by random chance, and there is insufficient evidence to support the alternative hypothesis. It's important to note that failing to reject the null hypothesis does not prove its truthfulness; it simply suggests that there is not enough evidence to support the alternative hypothesis.

(5) Critical Region

The critical region, also known as the rejection region, is a defined range of values or outcomes of a test statistic that leads to the rejection of the null hypothesis (H_0). The critical region is determined based on the chosen level of significance (α) and the distribution of the test statistic under the assumption that the null hypothesis is true.

The critical region represents the extreme or unlikely values of the test statistic that would cast doubt on the validity of the null hypothesis. If the observed test statistic falls within the critical region, it provides evidence against the null hypothesis and leads to its rejection in favour of the alternative hypothesis (H_1).

To determine the critical region, researchers specify the desired level of significance (α) before conducting the hypothesis test. The level of significance represents the maximum allowable probability of making a Type I error (rejecting the null hypothesis when it is actually true). The critical region is then defined such that the probability of observing a test statistic within that region, assuming the null hypothesis is true, is equal to or less than the chosen level of significance (α).The critical region is determined based on the specific distribution associated with the test statistic being used and the nature of the research question. For example, in a t-test, the critical region is defined by critical values obtained from the t-distribution, while in a z-test, it is determined by the critical values of the standard normal distribution. The critical region is often represented graphically on a probability distribution, showing the area in the tail(s) of the distribution associated with rejection of the null hypothesis. The critical values divide the distribution into the critical region (rejection region) and the non-critical region (non-rejection region).

When the calculated test statistic falls within the critical region, the null hypothesis is rejected, indicating that the observed data is unlikely to occur by random chance alone and supports the alternative hypothesis. Conversely, if the test statistic falls within the non-critical region, the null hypothesis is not rejected, suggesting that the observed data is reasonably likely to occur by random chance, and there is insufficient evidence to support the alternative hypothesis.It's important to note that the size and location of the critical region are influenced by the chosen level of significance. A smaller level of significance (e.g., $\alpha = 0.01$) results in a more stringent critical region, making it more difficult to reject the null hypothesis. On the other hand, a larger level of significance (e.g., $\alpha = 0.10$) widens the critical region, making it easier to reject the null hypothesis.

(6)One-Tailed and Two-Tailed Tests in Statistics:

In statistical hypothesis testing, researchers can choose between one-tailed and two-tailed tests based on the specific research question and the directionality of the effect being investigated. These tests differ in the way they assess the evidence against the null hypothesis (H_0) and the corresponding critical region.

One-Tailed Test

In a one-tailed (or one-sided) test, the alternative hypothesis (H_1) specifies the direction of the effect or difference between variables. It predicts that the observed data will be either significantly greater or significantly less than what would be expected under the null hypothesis. Therefore, the critical region is located entirely in one tail of the distribution of the test statistic.

The one-tailed test is appropriate when there is a clear theoretical or practical basis for predicting the direction of the effect. It allows researchers to focus their analysis on that specific direction and increases the power to detect the effect in that direction. One-tailed tests are often used in situations where previous research or knowledge suggests a particular directionality. For example, in a study investigating whether a new treatment improves test scores, the one-tailed test would focus on determining if the treatment leads to significantly higher test scores, neglecting the possibility of significantly lower scores.

Two-Tailed Test

In a two-tailed (or two-sided) test, the alternative hypothesis does not specify a particular direction of the effect. It predicts that the observed data will be significantly different from what would be expected under the null hypothesis, without specifying whether it will be greater or smaller. Therefore, the critical region is divided into two equal tails, one in each direction of the distribution of the test statistic.

The two-tailed test is appropriate when there is no prior expectation or theoretical basis to predict the direction of the effect. It provides a more conservative approach to hypothesis testing, as it requires stronger evidence to reject the null hypothesis compared to a one-tailed test. For example, in a study investigating whether a new teaching method affects test scores, the two-tailed test would examine if the teaching method leads to significantly different test scores, without specifying whether the scores will be higher or lower. The choice between one-tailed and two-tailed tests should be based on careful consideration of the research question, previous knowledge, and theoretical expectations. While a one-tailed test increases the power to detect an effect in a specific direction, it may miss effects in the opposite direction. A two-tailed test is more conservative but captures effects in both directions.

7. Non-parametric Test

In statistics, non-parametric tests, also known as distribution-free tests, are statistical methods used to make inferences and draw conclusions about populations or samples without assuming a specific probability distribution. Unlike parametric tests, which rely on assumptions about the underlying data distribution, non-parametric tests make fewer assumptions and are more robust to violations of distributional assumptions.

Non-parametric tests are often used when the data does not meet the assumptions required for parametric tests, such as when the data is skewed, have outliers, or when the sample size is small. These tests are also useful when dealing with ordinal or nominal data, as they do not require interval or ratio level measurements. Some common non-parametric tests include:

1. **Mann-Whitney U test:** This test is used to compare the medians of two independent groups. It is a non-parametric alternative to the independent samples t-test.
2. **Wilcoxon signed-rank test:** This test is used to compare the medians of two related or paired samples. It is a non-parametric alternative to the paired samples t-test.
3. **Kruskal-Wallis test:** This test is used to compare the medians of three or more independent groups. It is a non-parametric alternative to the one-way analysis of variance (ANOVA).
4. **Friedman test:** This test is used to compare the medians of three or more related groups. It is a non-parametric alternative to the repeated measures ANOVA.
5. **Spearman's rank correlation coefficient:** This test is used to assess the strength and direction of the monotonic relationship between two variables. It is a non-parametric alternative to Pearson's correlation coefficient.

Non-parametric tests rely on ranks or other orderings of the data rather than the actual numerical values. They use statistical techniques that compare the distributions of the data or evaluate the degree of association between variables without assuming a specific probability distribution. Advantages of non-parametric tests include their robustness to outliers and their ability to handle data that does not meet the assumptions of parametric tests. However, they generally have less statistical power than parametric tests when the data does conform to the assumptions of the parametric tests. Non-parametric tests are widely used in various fields, including psychology, sociology, biology, medicine, and environmental science, where the assumptions of parametric tests may not be met or when dealing with categorical or ranked data.

References

Agrawal, B.L. (2006) Basic Statistics, New Age International, India

Gupta, S.C. and Kapoor, V.K. (2020) Fundamentals of Mathematical Statistics, Sultan Chand and Sons, India

MULTIVARIATE STATISTICAL TECHNIQUES

Prabina Kumar Meher, Atmakuri Ramakrishna Rao
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Prabina.Meher@icar.gov.in

Multivariate data consist of observations on several different variables for a number of individuals or subjects. Data of this type arise in all the branches of science, ranging from psychology to biology, and methods of analyzing multivariate data constitute an increasingly important area of statistics. Indeed, the vast majority of data in forestry is multivariate and proper handling of such data is highly essential. Principal components analysis (PCA) and Factor analysis (FA) are multivariate techniques applied to a single set of variables to discover which sets of variables in the set form coherent subsets that are relatively independent of one another. The details of PCA and FA are discussed as below.

Principal Components Analysis

Most of the times the variables under study are highly correlated and as such they are effectively "saying the same thing". To examine the relationships among a set of *p* correlated variables, it may be useful to transform the original set of variables to a new set of uncorrelated variables called *principal components*. These new variables are linear combinations of original variables and are derived in decreasing order of importance so that, for example, the first principal component accounts for as much as possible of the variation in the original data.

Let $x_1, x_2, x_3, \ldots, x_p$ are variables under study, then first principal component may be defined as

$$z_1 = a_{11}\, x_1 + a_{12}\, x_2 + \ldots\ldots + a_{1p}\, x_p$$

such that variance of z_1 is as large as possible subject to the condition that

$$a_{11}^2 + a_{12}^2 + \ldots\ldots + a_{1p}^2 = 1$$

This constraint is introduced because if this is not done, then Var(z_1) can be increased simply by multiplying any a_{1j}s by a constant factor

The second principal component is defined as

$$z_2 = a_{21}\, x_1 + a_{22}\, x_2 + \ldots\ldots\ldots + a_{2p}\, x_p$$

Such that Var(z_2) is as large as possible next to Var(z_1)subject to the constraint that

$a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$ and $\text{cov}(z_1, z_2) = 0$ and so on.

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial p variables in subsequent analysis, thus, reducing the effective dimensionality of the problem. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result. However, Principal Component Analysis is more of a means to an end rather than an end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique which does not require user to specify the statistical model or assumption about distribution of original varieties. It may also be mentioned that principal components are artificial variables and often it is not possible to assign physical meaning to them. Further, since Principal Component Analysis transforms original set of variables to new set of uncorrelated variables, it is worth stressing that if original variables are uncorrelated, then there is no point in carrying out principal component analysis.

Computation of principal component

Let us consider the following data on average minimum temperature (x_1), average relative humidity at 8 hrs. (x_2), average relative humidity at 14 hrs. (x_3) and total rainfall in cm. (x_4) pertaining to Raipur district from 1970 to 1986 for kharif season from 21st May to 7th Oct.

X_1	x_2	x_3	x_4
25.0	86	66	186.49
24.9	84	66	124.34
25.4	77	55	98.79
24.4	82	62	118.88
22.9	79	53	71.88
7.7	86	60	111.96
25.1	82	58	99.74
24.9	83	63	115.20
24.9	82	63	100.16
24.9	78	56	62.38
24.3	85	67	154.40
24.6	79	61	112.71
24.3	81	58	79.63
24.6	81	61	125.59
24.1	85	64	99.87

	24.5	84	63	143.56
	24.0	81	61	114.97
Mean	23.56	82.06	61.00	112.97
S.D.	4.13	2.75	3.97	30.06

with the variance co-variance matrix.

$$\Sigma = \begin{bmatrix} 17.02 & -4.12 & 1.54 & 5.14 \\ & 7.56 & 8.50 & 54.82 \\ & & 15.75 & 92.95 \\ & & & 903.87 \end{bmatrix}$$

Find the eigen values and eigen vectors of the above matrix. Arrange the eigen values in decreasing order. Let the eigen values in decreasing order and corresponding eigen vectors are

$\lambda_1 = 916.902 \quad a_1 = (0.006, \; 0.061, \; 0.103, \; 0.993)$

$\lambda_2 = 18.375 \quad a_2 = (0.955, \; -0.296, \; 0.011, \; 0.012)$

$\lambda_3 = 7.87 \quad a_3 = (0.141, \; 0.485, \; 0.855, \; -0.119)$

$\lambda_4 = 1.056 \quad a_4 = (0.260, \; 0.820, \; -0.509, \; 0.001)$

The principal components for this data will be

$$z_1 = 0.006 \; x_1 + 0.061 \, x_2 + 0.103 \, x_3 + 0.993 \, x_4$$

$$z_2 = 0.955 \, x_1 - 0.296 \, x_2 + 0.011 \, x_3 + 0.012 \, x_4$$

$$z_3 = 0.141 \, x_1 + 0.485 \, x_2 + 0.855 \, x_3 - 0.119 \, x_4$$

$$z_4 = 0.26 \; x_1 + 0.82 \; x_2 - 0.509 \, x_3 + 0.001 \, x_4$$

The variance of principal components will be eigen values i.e.

$\text{Var}(z_1) = 916.902$, $\text{Var}(z_2) = 18.375$, $\text{Var}(z_3) = 7.87$, $\text{Var}(z_4) = 1.056$

The total variation explained by original variables is

$$= \text{Var}(x_1) + \text{Var}(x_2) + \text{Var}(x_3) + \text{Var}(x_4)$$

$$= 17.02 + 7.56 + 15.75 + 903.87 = 944.20$$

The total variation explained by principal components is

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 916.902 + 18.375 + 7.87 + 1.056 = 944.20$$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.

The proportion of total variation accounted for by the first principal component is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{916.902}{944.203} = .97$$

Continuing, the first two components account for a proportion

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{935.277}{944.203} = .99$$

of the total variance.

Hence, in further analysis, the first or first two principal components z_1 and z_2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of x_i s in equations of z_i s. For above data, the first two principal components for first observation i.e. for year 1970 can be worked out as

$$z_1 = 0.006 \text{ x } 25.0 + 0.061 \text{ x } 86 + 0.103 \text{ x } 66 + 0.993 \text{ x } 186.49 = 197.380$$

$$z_2 = 0.955 \text{ x } 25.0 - 0.296 \text{ x } 86 + 0.011 \text{ x } 66 + 0.012 \text{ x } 186.49 = 1.383$$

Similarly for the year 1971

$$z_1 = 0.006 \text{ x } 24.9 + 0.061 \text{ x } 84 + 0.103 \text{ x } 66 + 0.993 \text{ x } 124.34 = 135.54$$

$$z_2 = 0.955 \text{ x } 24.9 - 0.296 \text{ x } 84 + 0.011 \text{ x } 66 + 0.012 \text{ x } 124.34 = 1.134$$

Thus the whole data with four variables can be converted to a new data set with two principal components.

Note: The principal components depend on the scale of measurement, for example, if in the above example X_1 is measured in ^{0}F instead of ^{0}C and X_4 in mm in place of cm, the data gives different principal components when transformed to original x's. In very specific situations results are same. The conventional way of getting around this problem is to use standardized variables with unit variances, i.e., correlation matrix in place of dispersion matrix. But the principal components obtained from original variables as such and from correlation matrix will not be same and they may not explain the same proportion of variance in the system. Further more, one set of principal components is not simple function of the other. When the variables are standardized, the resulting variables contribute almost equally to the principal components determined from correlation matrix. Variables should probably be standardized if they are measured on scales with widely differing ranges or if measured units are not commensurate. Often population dispersion matrix or correlation matrix are not available. In such situations sample dispersion matrix or correlation matrix can be used.

Applications of principal components:

- The most important use of principal component analysis is reduction of data. It provides the effective dimensionality of the data. If first few components account for most of the variation in the original data, then first few components' scores can be utilized in subsequent analysis in place of original variables.
- Plotting of data becomes difficult with more than three variables. Through principal component analysis, it is often possible to account for most of the variability in the data by first two components, and it is possible to plot the values of first two components scores for each individual. Thus, principal component analysis enables us to plot the data in two dimensions. Particularly detection of outliers or clustering of individuals will be easier through this technique. Often, use of principal component analysis reveals grouping of variables which would not be found by other means.
- Reduction in dimensionality can also help in analysis where no. of variables is more than the number of observations, for example, in discriminant analysis and regression analysis. In such cases, principal component analysis is helpful by reducing the dimensionality of data.
- Multiple regression can be dangerous if independent variables are highly correlated. Principal component analysis is the most practical technique to solve the problem. Regression analysis can be carried out using principal components as regressors in place of original variables. This is known as principal component regression.

Discriminant Analysis

Discriminant analysis and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a separatory procedure, it is often employed on a one - time basis in order to investigate observed differences when causal relationships are not well understood. Classification procedures are less explanatory in the sense that they lead to well- defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination.

Thus, the immediate goals of discrimination and classification, respectively, are as follows.

Goal 1. To describe either graphically (in three or lower dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose numerical values are such that the collections are separated as much as possible.

Goal 2. To sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign a new object to the labeled classes.

We shall follow convention and use the term discrimination to refer to Goal 1. This terminology was introduced by R.A. Fisher in the first modern treatment of separatory problems. A more descriptive term for this goal, however, is separation; we shall refer to the second goal as classification, or allocation.

A function that separates may sometimes serve as an allocation, and conversely, an allocatory rule may suggest a discriminatory procedure. In practice, Goals 1 and 2 frequently overlap and the distinction between separation and allocation becomes blurred.

Here we discuss Fisher's linear discriminant function for two multivariate populations having same dispersion matrix. For more general cases readers are requested to go through the references cited at the end.

Fisher's Discriminant Function

Here Fisher's idea was to transform the multivariate observations **x** to univariate observations y such that the y's derived from populations π_1 and π_2 were separated as much as possible. Fisher's approach assumes that the populations are normal and also assumes the population covariances matrices are equal because a pooled estimate of common covariance matrix is used.

A fixed linear combination of the **x**'s takes the values y_{11}, y_{12}, ..., y_{1n1}, for the observations from the first population and the values y_{21}, y_{22}, ..., y_{2n2}, for the observations from the second population. The separation of these two sets of univariate y's is assessed in terms of the differences between $\bar{y}_1$ and $\bar{y}_2$ expressed in standard deviation units. That is,

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \text{ where } s_y^2 = \frac{\sum_{j=1}^{n1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the variance. The objective is to select the linear combination of the **x** to achieve maximum separation of the sample means $\bar{y}_1$ and $\bar{y}_2$.

Result: The linear combination y = $\hat{\boldsymbol{l}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\mathbf{pooled}}^{-1}\mathbf{x}$ maximizes the ratio

$$\frac{\text{(Squared distance between sample means of y)}}{\text{(Sample variance of y)}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{(\hat{\boldsymbol{l}}'\bar{\mathbf{x}}_1 - \hat{\boldsymbol{l}}'\bar{\mathbf{x}}_2)^2}{\hat{\boldsymbol{l}}'\boldsymbol{S}_{\boldsymbol{pooled}}\hat{\boldsymbol{l}}} = \frac{(\hat{\boldsymbol{l}}'\mathbf{d})^2}{\hat{\boldsymbol{l}}'\boldsymbol{S}_{\boldsymbol{pooled}}\hat{\boldsymbol{l}}}$$

overall possible coefficient vectors $\hat{\boldsymbol{l}}'$ where $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. The maximum of the above ratio is $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{s}_{\mathbf{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, the Mahalanobis distance.

Fisher's solution to the separation problem can also be used to classify new observations. An allocation rule is as follows.

Allocate $\mathbf{x_0}$ to π_1 if

$$y_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{s}_{\mathbf{pooled}}^{-1}\mathbf{x_0} \geq \hat{\mathbf{m}} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{s}_{\mathbf{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

and to π_2 if

$$y_0 < \hat{\boldsymbol{m}}$$

If we assume the populations π_1 and π_2 are multivariate normal with a common covariance matrix, then a test of **H₀: μ₁ = μ₂** versus **H₁: μ₁ ≠ μ₂** are accomplished by referring

$$\frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p}\left(\frac{n_1 n_2}{n_1 + n_2}\right)\mathbf{D^2}$$

to an F-distribution with $\upsilon_1 = p$ and $\upsilon_2 = n_1 + n_2 - p - 1$ d.f. If **H₀** is rejected, we can conclude the separation between the two populations is significant.

Example:

To construct a procedure for detecting potential hemophilia ‘A’ carriers, blood samples were analyzed for two groups of women and measurements on the two variables, $x_1 = \log_{10}$(AHF activity) and $x_2 = \text{loh}_{10}$(AHF-like antigens) recorded. The first group of $n_1 = 30$ women were selected from a population who do not carry hemophilia gene (normal group). The second group of $n_2 = 22$ women were selected from known hemophilia ‘A’ carriers (obligatory group). The mean vectors and sample covariance matrix are given as

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} \text{ and } \mathbf{S}_{\mathbf{pooled}}^{-1} = \begin{bmatrix} 131158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Now the linear discriminant function is

$$y_0 = \hat{\boldsymbol{l}}'\,\mathbf{x}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\mathbf{pooled}}^{-1}\mathbf{x}_0$$

$$= \begin{bmatrix} .2418 & -0.0652 \end{bmatrix} \begin{bmatrix} 131158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 37.61x_1 - 28.92\,x_2$$

Moreover

$$\bar{y}_1 = \hat{\boldsymbol{l}}'\,\bar{\mathbf{x}}_1 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} = 0.88$$

$$\bar{y}_2 = \hat{\boldsymbol{l}}'\,\bar{\mathbf{x}}_2 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.2483 \\ -0.0262 \end{bmatrix} = -10.10$$

and the mid-point between these means is

$$\hat{\mathbf{m}} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{s}_{\mathbf{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = -4.61$$

Now to classify a women who may be a hemophilia ‘A’ carrier with $x_1 = -.210$ and $x_2 = -0.044$, we calculate

$y_0 = \hat{\boldsymbol{l}}'\,\mathbf{x}_0 = 37.61x_1 - 28.92\,x_2 = -6.62$

Since $y_0 < \hat{m}$ we classify the women in π_2 population, i.e., to obligatory carrier group.

Factor Analysis

Some Basics

Factor analysis is a data reduction technique, which often requires large sample size to have a valid interpretation. The basic idea in factor analysis is that a large number of explanatory variables having similar type of responses can be captured with a single latent variable that cannot be measured directly. For example, the latent variable (or factor) socioeconomic status is associated with the observed variables income, education, health status, occupation, on which the peoples' responses are of similar type.

In factor analysis, the number of factors is same as the number of variables, where each factor captures a certain amount of variation of all the variations present in the observed variables. The factors are always arranged in the decreasing order of their variances. In factor analysis, one expects three outputs viz., common factor variances, factor loadings and factor scores. The common factor variance is the measure of the amount variation explained by a factor present in the observed variables. Factor loading measures the underlying relationship that an observed variable have with a factor. The factor scores are the transformed data, commonly the weighted sum/mean of the observed variables (or manifest variables).

The factor scores are not the penultimate output rather than act as an intermediate step (dimensionality reduction) for carrying out further statistical analysis, a much important one. In other words, factor scores enable user to use a single variable, instead of set of variables, as a measure of the factor in the other statistical investigation. For example, in case of linear model or mixed model, the factor scores can be used as variable (fixed factors or random factors), but here it refers to the categorical independent variable. Further, technically the factor scores are continuous and hence can be used as covariates in the model rather than as factors.

Type of Factor Analysis

There are two types of factor analysis, one is Exploratory Factor Analysis (EFA) and other is Confirmatory Factor Analysis (CFA). In CFA, one assumption is that there should be prior information about the number of factors likely to be encountered as well as which variables will be loaded onto which factors. On the other hand, CFA allows the researchers to test the hypothesis that whether the relationship between a variable and the underlying factor exits or not. Initially, the researcher postulates a certain a priori relationship pattern based on existing knowledge i.e., published research (empirical and/or theoretical) and then test the hypothesis statistically. In EFA, the researcher tries to find out the number of underlying constructs (factors) without having any a priori information about the number of factors. In other words, in EFA, the number of factors is determined on the basis of the dataset supplied by the user, and also depends upon user interpretation. Linking these two approaches, one can use EFA first to explore the underlying factors and then perform CFA to validate the structure of factors in a new dataset that has not been used for performing EFA. For example, a factor "depression" can be obtained with underlying variables depressed mood, fatigue, exhaustion and social dysfunction through EFA for a sample of rural women, and then the CFA can be used to validate this factor using a sample of urban women. In EFA, the cut-off of loading are much relaxed than that of CFA. In other words, a variable having loading value <|0.7| is disqualified from its loading onto a certain factor (Thumb rule). Generally, the EFA is most commonly used in day-to-day life than that of CFA. So, in this study material we only focused on EFA.

Exploratory Factor Analysis (EFA)

Before carrying out factor analysis, some important points need to be considered. At first, the reliability of the dataset should be checked for factor analysis. In other words, for factor analysis, the values of the variables should be in interval scale, each variable should be normally distributed, pairs of variables should follow bi-variate normal distribution and the dataset as a whole should follow multivariate normal distribution. Further, the sample size should be large. Field (2000) suggested 10-15 observations per variable. Habing (2003) state that there should be at least 50 observations and the number of observations should be at least 5 times as many variables. Comrey (1973) categorized the sample size for its suitability to factor analysis i.e., 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 or more as excellent. Also, one can conduct Kaiser-Meyer-Olkin (KMO) test to check the sample adequacy. The sample is said to be adequate if KMO value is more than 0.5.

As far as correlation matrix is concerned, the observed variables should be linearly related but not highly correlated that may lead to the matrix as singular and create difficulty in determining the unique contribution of the variables to the factors. To check the correlation among variables, one can use Bartlett's test of spherity to test the null hypothesis that the correlation matrix is a identity matrix and the result should come out as significant. After rejecting the null hypothesis, one can validate the presence of multi-collinearity via the determinant of the correlation matrix ie., if the determinant is greater than 0.00001, then there is no multi-collinearity (Field, 2000).

After getting correlation matrix, it is essential to determine whether factor analysis (FA) or principal component analysis (PCA) is to be performed. The main difference between these two lies on the way the eigen values are used. In PCA, all the diagonal elements of the correlation matrix are 1 and all the variance present in the dataset are accounted by the components. However, in FA, the diagonal of the correlation matrix are squared multiple correlation coefficient, which is further used to get the eigen values and thereby the factor scores. Also, all the variances are not accounted by the factors as there is also an error variance. Further, in PCA the sum of square of the factor loadings of a variable provided the variance accounted for by that variable, which is not same in FA as it is assumed that the variables do not account for 100% of the variance. Theoretically, FA is more correct than PCA (Field, 2000) but practically there is little difference and is further decreased with decrease in the number of variables and increase in the value of factor loadings (Rietveld and Van Hout, 1993).

In conducting FA, one of the most important questions is the number of factors to be retained in the model. In PCA, the number of components is same as the number of positive eigen value. However eigen values are sometime positive and close to zero, and in that situation deciding the number of factor is difficult. In literature certain thumb rules are there to take decision about the number of factors. Guttman-Kaiser rule state that the factor with eigen value >1 should be retained in the model. Hair et al, (1995) stated that in the natural sciences the number factors retained in the model should explain at least 95% of the total variance present in the observed variables. In humanities, the number factors that can explain up to 60-70% variation may be retained in the model (Hair et al, 1995; Pett et al, 2003). Besides, another option is that first draw a scree plot (Cattell, 1966) and retained all those factors appeared before reaching the point of inflection.

After extracting the factors, the next task is to name the factors and interpret them. Since, most variable have higher value of loading on the most important factors and less amount of loadings on the remaining factors, it is always a difficult task to interpret about the factors. However, the factor rotation can help in this respect to a large extent. Factor rotation transforms the original loadings and thereby the interpretation becomes easier. Rotation maximizes the high loading items and minimizes the less loading items. There are two rotation techniques viz., orthogonal/

varimax and oblique/promax that are commonly used in factor analysis. Varimax rotation (Thomson, 2004) is the most common rotational technique used in factor analysis that produces uncorrelated factors. On the other hand, in oblique rotation, the factors are correlated. Often, the oblique rotation provides more accurate results when the data does not meet the prior assumptions. Further, to decide the type of rotation technique is almost difficult and therefore first carryout the analysis with oblique rotaions, and if the oblique rotation demonstrates a negligible correlation between the extracted factors then it is reasonable to use orthogonally rotated factors (Field, 2000). Regardless of the rotation techniques uses, the objective is to provide easier interpretation of the results.

Interpretation of EFA is nothing but to determine which variables are attributed to a factor and labeling of that factor. However, the labeling of a factor is a subjective process (Henson and Roberts, 2006), where the meaningful of the factor is dependent on the researchers definition. Moreover, through and systematic factor analysis is nothing but to find those factors that together explain the majority of the responses.

Mathematical aspects of EFA

Consider a dataset with *n* observations and *p* standardized variables $x_1, x_2, ..., x_p$. Then, in EFA the observed variables are expressed as the linear combination of the common factors and unique factor i.e., $x_i = a_{i1}F_1 + a_{i2}F_2 + a_{i3}F_3 + ... + a_{ik}F_k + e_i$, where i=1,2,…, p, k<p and a_{ik} is the factor loading of i^{th} variable on k^{th} factor which is not same as that of eigen vector. The assumptions of this model are $E(e_i) = 0$, $V(e_i) = \psi_i$, $E(e_i e_j) = 0$, $E(e_i F_j) = 0$ and $E(F_i F_j) = 0$. In matrix notation we can write $\mathbf{X}_{p\times n} = \mathbf{L}_{p\times k}\mathbf{F}_{k\times n} + \mathbf{E}_{p\times n}$, where

$$\mathbf{X}_{p\times n} = \begin{bmatrix} x_{11}\ x_{12}\ x_{13} \dots x_{1n} \\ x_{21}\ x_{22}\ x_{23} \dots x_{2n} \\ \dots\ \dots\ \dots\ \dots\ \dots \\ x_{p1}\ x_{p2}\ x_{p3} \dots x_{pn} \end{bmatrix},\ \mathbf{F}_{k\times n} = \begin{bmatrix} F_{11}\ F_{12}\ F_{13} \dots F_{1n} \\ F_{21}\ F_{22}\ F_{23} \dots F_{2n} \\ \dots\ \dots\ \dots\ \dots\ \dots \\ F_{k1}\ F_{k2}\ F_{k3} \dots F_{kn} \end{bmatrix},\ \mathbf{L}_{p\times k} = \begin{bmatrix} a_{11}\ a_{12}\ a_{13} \dots a_{1k} \\ a_{21}\ a_{22}\ a_{23} \dots a_{2k} \\ \dots\ \dots\ \dots\ \dots\ \dots \\ a_{p1}\ a_{p2}\ a_{p3} \dots a_{pk} \end{bmatrix} \text{ and}$$

$$\mathbf{E}_{p\times n} = \begin{bmatrix} e_{11}\ e_{12}\ e_{13} \dots e_{1n} \\ e_{21}\ e_{22}\ e_{23} \dots e_{2n} \\ \dots\ \dots\ \dots\ \dots\ \dots \\ e_{p1}\ e_{p2}\ e_{p3} \dots e_{pn} \end{bmatrix}.$$

Also, it is assumed that $E(\mathbf{E}) = 0$, $E(\mathbf{F}) = 0$, $\text{cov}(\mathbf{F}, \mathbf{E}) = 0$, $V(\mathbf{E}) = Diag(\psi_1, \psi_2, ..., \psi_p) = \boldsymbol{\psi}(say)$ and $\text{var}(\mathbf{F}) = \mathbf{I}$. The correlation matrix is generally used for performing the factor analysis. Here the diagonal elements are 1 (often described as the variance of the observed variable). In PCA, this matrix is used as such but factor analysis involves the replacing of diagonal element with communality estimate. The communality estimate is the estimated proportion of variance of the variable that is free of error variance and is shared with other variables in the matrix. These estimates reflect the variance of a variable in common with all others together. The initial estimate of the communality is taken as the squared multiple correlation coefficients and then the communalities of the variables are estimated as the sum of the square of the loadings onto different factors. Once the correlation matrix of the observed variables are obtained, the factor analysis can be written as $\mathbf{\Sigma} = \mathbf{LL}' + \boldsymbol{\psi}$, which nothing but $\mathrm{var}(\mathbf{X}_{p\times n}) = \mathrm{var}(\mathbf{L}_{p\times k}\mathbf{F}_{k\times n} + \mathbf{E}_{p\times n})$. So, for the i^{th} variable, one can write $1 = (a_{i1}^2 + a_{i2}^2 + ... + a_{ip}^2) + \psi_i$ or $1 = h_i^2 + \psi_i$ or Total variance=Variance explained by the common factors + Error variance. Here h_i^2 is the communality and 1-h_i^2 is the variance accounted for by the i^{th} unique factor. In this model, there is a need to estimate the common factor loadings (***L***) as well as the factor scores (***F***). For estimating ***L***, there are two methods available one is Principal Axis Factor (PAF) method and other is Maximum Likelihood (ML) method. PAF makes no assumption about the error and minimizes the sum of squares of the residual matrix i.e., $\frac{1}{2}tr\left[(S-\Sigma)^2\right] = \sum_i \sum_j (s_{ij} - \sigma_{ij})^2$, where s_{ij} and σ_{ij} are the observed correlation matrix and implied correlation matrix, respectively (Jöreskog, 2007). The maximum likelihood (ML) estimation is derived from the theory of normal distribution. The ML value is obtained by minimizing $\ln|\Sigma| - \ln|S| + tr[S\Sigma^{-1}] - p$, which similar to minimizing the discrepancy function $\sum_i \sum_j \left[\frac{(s_{ij} - \sigma_{ij})^2}{\psi_i^2 \psi_j^2}\right]$ (MacCallum et al, 2007).

For estimation of factor scores, generally three types of methods are used viz., ordinary least squares, weighted least squares and regression method. Let ***x****ᵢ* be the i$^{\text{th}}$ observation vector and ***f****ᵢ* is the corresponding vector of factor scores, then we can write $\mathbf{x}_i = \mathbf{Lf}_i + \mathbf{e}_i$, where i=1,2,.., *n*, and the estimates of factor scores for this model by different methods are provided as follows:

(I) ***Ordinary Least Square***

The estimate of $\mathbf{f}_i$ can be obtained by minimizing the error sum of squares i.e., $\sum_{j=1}^{p} e_{ij}^2 = \sum_{j=1}^{p} (x_{ij} - a_{i1} f_1 - a_{i2} f_2 - ... - a_{ik} f)^2 = (\mathbf{x}_i - \mathbf{L}\mathbf{f}_i)'(\mathbf{x}_i - \mathbf{L}\mathbf{f}_i)$. This is like a least squares regression, except in this case we already have estimates of the parameters (the factor loadings). In matrix notations, it can be written as $\hat{\mathbf{f}}_i = (\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}'\mathbf{x}_i$. Using the principal component method with the unrotated factor loadings, the results can be obtained as

$$\hat{\mathbf{f}}_i = \begin{pmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}} \hat{\zeta}_1 \mathbf{x}_i \\ \frac{1}{\sqrt{\hat{\lambda}_2}} \hat{\zeta}_2 \mathbf{x}_i \\ ... \\ \frac{1}{\sqrt{\hat{\lambda}_k}} \hat{\zeta}_k \mathbf{x}_i \end{pmatrix},$$

where $\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_k$ are the eigen vectors and $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k$ are the estimate of eigen values.

(II) ***Weighted Least Squares***

In this method, larger weights are given to the variables having low specific variances. Variables with low specific variances are those for which the model fits the data best. In other words, the variable with the low specific variance provides more information regarding the true values for the specific factors. For the above considered model, we wish to minimize

$$\sum_{j=1}^{p} \frac{e_{ij}^2}{\psi_j} = \sum_{j=1}^{p} \frac{(x_{ij} - a_{i1} f_1 - a_{i2} f_2 - ... - a_{ik} f)^2}{\psi_j} = (\mathbf{x}_i - \mathbf{L}\mathbf{f}_i)'\boldsymbol{\psi}^{-1}(\mathbf{x}_i - \mathbf{L}\mathbf{f}_i) \quad ,$$

that resulted in the estimate as $\hat{\mathbf{f}}_i = (\mathbf{L}'\boldsymbol{\psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\psi}^{-1}\mathbf{x}_i$. Both OLS and WLS methods are used for estimating the factor scores, while PAF method is used to estimate the factor loadings.

(III) ***Regression method***

This method is used when maximum likelihood is used for estimating the factor loadings. Now, for standardized variables the joint distribution of $\mathbf{x}_i$

and $\mathbf{f}_i$ can be writes as $\begin{pmatrix}\mathbf{x}_i\\ \mathbf{f}_i\end{pmatrix} \sim N\left[\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\mathbf{LL}'+\boldsymbol{\psi} & \mathbf{L}\\ \mathbf{L}' & \mathbf{I}\end{pmatrix}\right]$. Then, we can calculate the conditional expectation of the factor score $\mathbf{f}_i$ given the observed data $\mathbf{x}_i$ as $E(\mathbf{f}_i|\mathbf{x}_i)=\mathbf{L}'(\mathbf{LL}'+\boldsymbol{\psi})^{-1}\mathbf{x}_i$, which is nothing but the estimate of $\mathbf{f}_i$.

Step by step procedure for performing exploratory factor analysis using R

Step 1: Set the working directory. Let my directory is “meher” present in “D” drive. Then, set the directory as

```
setwd(‟C:/Documents and Settings/Prabin/Desktop/meher”)
```

Step 2: Read the data from the specified directory. Let my data file is *fact.txt* present in the directory. Then data file can be imported to R as

```
x <- read.table (file= ‟fact.txt”)
```

Step 3: Check the normality assumption of each variable using Shapiro-Wilk’s test.

`shapiro.test (x[,i])` # This is for ith variable. If P-value is >level of significance, the variable is normally distributed.

Step 4: Check the adequacy of the each variable and sample as a whole for factor analysis using KSA and KMO and test. The desired value of KMO is > 0.5. Variables with MSA being below 0.5 indicate that item does not belong to a group and may be removed from the factor analysis.

```
kmo <- function(x)
{
x <- subset(x, complete.cases(x)) # Omit missing values
r <- cor(x)                                    # Correlation matrix
r2 <- r^2                  # Squared correlation coefficients
i <- solve(r)                  # Inverse matrix of correlation matrix
d <- diag(i)                 # Diagonal elements of inverse matrix
p2 <- (-i/sqrt(outer(d, d)))^2          # Squared partial correlation
coefficients
diag(r2) <- diag(p2) <- 0         # Delete diagonal elements
KMO <- sum(r2)/(sum(r2)+sum(p2))
MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
return(list(KMO=KMO, MSA=MSA))
}
kmo (x)
```

Step 5: Check that the correlation matrix is not an identity matrix using Bartlett’s sphericity test. The test should come out significant.

```
bst <- function(x)
{
method <- "Bartlett's test of sphericity"
data.name <- deparse(substitute(x))
x <- subset(x, complete.cases(x))  # Omit missing values
n <- nrow(x)
p <- ncol(x)
chisq <- (1-n+(2*p+5)/6)*log(det(cor(x)))
df <- p*(p-1)/2
p.value <- pchisq(chisq, df, lower.tail=FALSE)
names(chisq) <- "X-squared"
names(df) <- "df"
return(structure(list(statistic=chisq,       parameter=df,
p.value=p.value,
method=method, data.name=data.name), class="htest"))
}
bst (x)
```

Step 6: Test that there is no presence of high degree of multicollinearity. The determinant of the matrix should come out > 0.0001 to pass the test.

```
det(cor(x))
```

Step 7: Carryout factor analysis to extract the factor loadings (by ML estimate method), common variances and specific variances.

```
factanal (x=swiss, factors=2, rotation= “varimax or
promax”)
or
factanal (~., factors=2, data=swiss, rotation= “varimax
or promax”)
```

# In the result one cannot see the complete factor loadings but it is possible with the following commands.

```
factanal (~., factors=2, rotation= “varimax or
promax”)$loadings[,i] # for complete ith factor loading.
```

Step 8: Estimate the factor scores either by Bartlett’s WLS method or Johnson’s regression method.

```
factanal (~., factors=2, rotation= “varimax or promax”,
scores=”Bartlett or regression”)$scores
```

Step 9: The factor loadings, common variances, specific variances can also be computed by supplying the covariance matrix and number of observations. However, the scores can only be obtained when full data set is available.

```
factanal (factors=2, covmat=cor(swiss),rotation= “varimax
or promax”, n.obs=47)
```

Step 10: Interpretation of the result and conclusion

Note: One can use the “psych” package of R-software for KMO test and Barlett’s test of sphericity using single line code as provided below.

`KMO(r)` # r is the correlation matrix. This will provide the values of both KMO and KSA

`cortest.bartlett(r, n)` # r is the correlation matrix and n is the number of observation in the dataset.

References

Cattell, RB (1966) The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

Chatfield, C. and Collins, A.J. (1990). Introduction to multivariate analysis. *Chapman and Hall publications*.

Comrey, AL (1973) *A First Course in Factor Analysis*. New York: Academic Press, Inc.

Field, A (2000) *Discovering Statistics using SPSS for Windows*. London – Thousand Oaks – New Delhi: Sage publications.

Habing, B (2003) *Exploratory Factor Analysis*. Website: http://www.stat.sc.edu/~habing/courses/530EFA.pdf

Hair, J., Anderson, RE., Tatham, RL., Black, WC (1995) *Multivariate data analysis*. 4th edn. New Jersey: Prentice-Hall Inc.

Henson, RK., Roberts, JK (2006) Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, 66(3).

Johnson, R.A. and Wichern, D.W. (1996). Applied multivariate statistical analysis. *Prentice-Hall of India Private Limited*.

Jöreskog, G (2007) *Factor analysis and its extensions*, in *Factor analysis at 100: Historical Developments and Future Directions*, R. Cudeck and R.C. MacCallum, eds., Lawrence Erlbaum, Mahwah, NJ, pp. 47–77.

MacCallum, RC., Browne, MW., Cai, L (2007) *Factor analysis models as approximations*, in *Factor Analysis at 100: Historical Developments and Future Directions*, R. Cudeck and R.C. MacCallum eds., Lawrence Erlbaum, Mahwah, NJ, pp. 153–175.

Pett, MA., Lackey, NR., Sullivan, JJ (2003) *Making Sense of Factor Analysis: The use of factor analysis for instrument development in health care research*. California: Sage Publications Inc.

Rietveld, T., Van Hout, R (1993) *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin – New York: Mouton de Gruyter.

Thompson, B (2004) *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington, DC: American Psychological Association.

CORRELATION AND REGRESSION ANALYSIS

Dr.Kanchan Sinha
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
kanchan.sinha@icar.gov.in

1. Introduction

Correlation is a powerful statistical concept that enables us to explore the relationships between variables and uncover hidden patterns in complex data. By measuring the extent to which two variables move together, correlation helps us gain insights into the interconnectedness of phenomena. In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). Regression analysis is primarily used for two distinct purposes. First, it is widely used for prediction and forecasting, which overlaps with the field of machine learning. Second, it is also used to infer causal relationships between independent and dependent variables. This methodology is widely used in business, social and behavioral sciences, biological sciences including agriculture. For example, yield of a crop can be predicted by utilizing the relationship between yield and other factors like water temperature, rainfall, quantity of fertilizer, quantity of seeds, irrigation level and relative humidity, etc.

A functional relationship between two variables can be expressed by a mathematical formula. If x denotes the independent variable and y the dependent variable, then y can be related x through a functional relation of the form $y = f(x)$. Given a particular value of x, the function f indicates the corresponding value of y. In regression analysis, the variable x is known as input variable, explanatory variable or predictor variable. This is an exact mathematical relationship. In statistical relation, may not be perfect owing to sampling. The above functional form is made a statistical model by adding an error term as $y = f(x) + \varepsilon$, where ε denotes the error term.

Depending on the nature of the relationships between x and y, regression approach may be classified into two broad categories *viz.*, linear regression models and nonlinear regression models. The response variable is generally related to other causal variables through some parameters. The models that are linear in these parameters are known as linear models; whereas in nonlinear models parameters appear nonlinearly.

2. The Concept of Correlation

2.1 ***Defining Correlation*****:** Correlation refers to the statistical association between two or more variables, indicating the degree to which they tend to change together. It measures the direction (positive or negative) and strength (weak or strong) of the relationship.

2.2Significance of Correlation

Identifying Associations: Correlation helps us identify relationships between variables, providing a foundation for further analysis.

Prediction: Correlated variables can be used to make predictions about one variable based on the other(s).

Variable Selection: Correlation assists in selecting relevant variables for analysis, weeding out redundant or irrelevant ones.

2.3 Measuring Correlation

2.3.1 Pearson's Correlation Coefficient

The Pearson correlation coefficient (r) quantifies the linear relationship between two continuous variables and can be expressed as:

$$r = \frac{\Sigma \quad (x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\Sigma \quad (x_i - \underline{x})^2 \Sigma \quad (y_i - \underline{y})^2}}$$

Where, r is the correlation coefficient

x_i are the values of the x-variable.

y_i are the values of the y-variable.

$\underline{x}$ is the mean of the values of x-variable.

$\underline{y}$ is the mean of the values of y-variable.

Range and Interpretation: r ranges from -1 to 1, where -1 denotes a perfect negative correlation, 1 signifies a perfect positive correlation, and 0 indicates no linear relationship.

Strength of Correlation: Various criteria, such as effect size or correlation coefficient magnitude, determine the strength of the relationship.

2.3.2 Spearman's Rank Correlation Coefficient

Spearman's rho (ρ) measures the monotonic relationship (increasing or decreasing) between variables, especially when the relationship is not strictly linear and can be expressed as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} \quad d_i^2}{n(n^2 - 1)}$$

where, ρ is the Spearman's Rank Correlation Coefficient.

d_i is the difference between the two ranks of each observation.

n is the number of observations.

Advantages: It is robust to outliers and can handle ordinal or non-normal data.

Interpretation: Similar to Pearson's r, ρ ranges from -1 to 1, with the same interpretations.

2.4 Types of Correlation

2.4.1 Positive Correlation

Definition: Positive correlation exists when an increase in one variable corresponds to an increase in the other, and vice versa.

Examples: Height and weight, income and education level.

2.4.2 Negative Correlation

Definition: Negative correlation occurs when an increase in one variable corresponds to a decrease in the other, and vice versa.

Examples: Temperature and heating costs, exercise duration and body weight.

2.4.3 Zero Correlation

Definition: Zero correlation indicates no discernible relationship between variables.

Examples: Shoe size and IQ, number of siblings and favourite colour.

2.4.4 Interpreting Correlation

2.4.4.1 Causation vs. Correlation

Correlation does not imply causation; a strong relationship between two variables does not necessarily mean one variable causes the other.

Spurious Correlation: Be cautious of coincidental associations without a meaningful underlying connection.

2.4.4.2 Scatterplots

Visualizing Correlation: Scatter plots are graphical representations that help us assess the relationship between variables.

Patterns: Scatterplots can exhibit various patterns, such as linear, nonlinear, or clusters, aiding in understanding the correlation visually.

2.4.4.3 Applications of Correlation

Finance and Economics

Analyzing stock market trends and investment portfolios.

Examining relationships between economic indicators, such as GDP and unemployment rates.

Social Sciences

Investigating relationships between variables like crime rates and income levels.

Studying the impact of education on health outcomes.

Medicine and Health

Exploring the correlation between risk factors and disease prevalence.

Assessing the effectiveness of treatments or interventions.

Agriculture

Crop Yield and Environmental Factors

Pest and Disease Management

Crop Nutrient Requirements

Crop-Livestock Interactions

Climate Change Impact Assessment

Water Management

Market Analysis and Price Forecasting, etc.

3. Simple Linear Regression (SLR) Model

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

The simple linear regression model is usually written as

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$(3)

where the ε_i's are normal random variables with mean 0 and variance σ^2. The model implies (i) The average y-value at a given $x-$value is linearly related to x.

(ii) The variation in responses y at a given x value is constant.

(iii) The population of responses y at a given x is normally distributed.

(iv) The observed data are a random sample.

Regression model (3) is said to be simple and linear regression model. It is “simple” in the sense that there is only one predictor variable and “linear” in the sense that all parameters appeared linearly with the predictor variables. The parameters $\beta_0$and β_1 in regression model (3) are called regression coefficients, β_1 is the slope of the regression line. It indicates the change in the mean of the probability distribution of y per unit increase in x. The parameter β_0 is the y intercept of the regression line.

3.1 Estimation of Parameters in a Simple Linear Regression Model

In the above models the variables y and x are known, these are observed. The only unknown quantities are the parameters β’s. In regression analysis, our main concern is how precisely we can estimate these parameters. Once these parameters are estimated, our model becomes known and we can use it for further analysis. The method of least squares is generally used to estimate these parameters. For each observations (x_i, y_i), the method of least squares considers the error of each observation, i.e, for a simple model $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$. The method of least squares requires the sum of the n squared errors. This criterion is denoted by S:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

According to the method of least squares, the estimators of β_0 and β_1 are those values of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, that minimize the criterion S for the given observations. To minimize S, we differentiate S with respect to each parameter and equate to zero. We get as many equations as the number of parameters. Solving these equations simultaneously, we get the estimates of parameters. For example, for the regression model (3) the values of $\hat{\beta}_0$ and $\hat{\beta}_1$that minimizes S for any particular set of sample data are given by the following simultaneous equations:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \qquad (6)$$

These two equations are called normal equations and can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$ as follows

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \ (x_i-\underline{x})(y_i-\underline{y})}{\sum_{i=1}^{n} \ (x_i-\underline{x})^2} \qquad (7)$$

$$\hat{\beta}_0 = \frac{1}{n}(\sum_{i=1}^{n} \quad y_i - \beta_1 \sum_{i=1}^{n} \quad x_i) = \underline{y} - \beta_1\underline{x} \ (8)$$

where,$\underline{y}$ and $\underline{x}$ are the means of the y_i and x_i observations, respectively.

3. Multiple Linear Regression Model (MLR) Model

A regression model that involves more than one regressor variable is called a multiple regression model i.e., the multiple linear regression model is used to study the relationship between a dependent variable and one or more independent variables. The generic form of the linear regression model is

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad (9)$$

where,$y$is the dependent or explained variable and $x_1, x_2, \ldots, x_p$are the independent or explanatory variables. The regression model in the equation describes above is linear in the sense, it is a linear function of the unknown parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. In general, any regression model that is linear in the parameters (β's) is a linear regression model, regardless of the shape of the surface that it generates. We have also assumed that the expected value of the error term ε is zero. The parameter β_0 is the intercept of the regression model. If the range of the data includes $x_1 = x_2 = \cdots = x_p = 0$, then β_0 is the mean of y when $x_1 = x_2 = \cdots = x_p = 0$. Otherwise β_0 has no physical interpretation. The parameter β_1 indicates the expected change in response (y) per unit change in x_1 when $x_2, \ldots, x_p$ are held constant. Similarly $\beta_2$measures the expected change in response (y) per unit change in x_2 when $x_1, \ldots, x_p$ are held constant. For this reason the parameters $\beta_i, \forall\ i = 1,2, \ldots, p$ are often called as partial regression coefficients.

A. Assumptions of the Multiple Linear Regression Model

1. Linearity

The model defined by the following equation

$y = f(x_1, x_2, \ldots, x_p) + \varepsilon = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$ specifies a linear relationship between y and $x$and our primary interest is in estimation and inference about the parameter vector $\boldsymbol{\beta}$. For the regression to be linear in the sense described here, it must be of the form in the original variables or after some suitable transformation.

i. Full rank

There are no exact linear relationships among the variables in the model. x is an $n \times p$ matrix with rank p. Hence x has full column rank; the columns of x are linearly independent and there are at least p observations $(n \geq p)$.

i. Exogeneity of the independent variables:

The disturbance is assumed to have conditional expected value zero at every observation, which we can write as $E[x] = 0$.

In this equation, the left hand side states, in principle, that the mean of each ε_i conditioned on all observations $x$is zero. This strict exogeneity assumption states, in words, that no observations on x convey information about the expected value of the disturbance.

i. Homoscedasticity:

The fourth assumption concerns the variances and covariance of the disturbances:

$$Var(x) = \sigma^2, \forall\ i = 1, \dots, n$$

$$Cov(x) = 0\ \forall\ i \neq j \qquad (10)$$

Constant variance is labelled **homoscedasticity**. Consider a model that describes the profits of firms in an industry as a function of, say, size. Even accounting for size, measured in dollar terms, the profits of large firms will exhibit greater variation than those of smaller firms. The homoscedasticity assumption would be inappropriate here. Survey data on household expenditure patterns often display marked **heteroscedasticity**, even after accounting for income and household size. The two assumptions imply that

$$E[x] = [\sigma^2\ 0\ \cdots\ 0\ 0\ \sigma^2\ \cdots\ 0\ \vdots\ \vdots\ \ddots\ \vdots\ 0\ 0\ \cdots\ \sigma^2\ ] = \sigma^2 I\ (11)$$

i. Data generating process for the regressors

It is common to assume that $x_i$is nonstochastic, as it would be in an experimental situation. Here the analyst chooses the values of the regressors and then observes y_i. This process might apply, for example, in an agricultural experiment in which y_i is yield and $x_i$is fertilizer concentration and water applied.

i. Normality

It is convenient to assume that the disturbances are normally distributed with zero mean and constant variance. This is a convenience that we will dispense with after some analysis of its implications. The normality assumption is useful for defining the computations behind statistical inference about the regression, such as confidence intervals and hypothesis tests. For practical purposes, it will be useful then to extend those results and in the process develop a more flexible approach that does not rely on this specific assumption.

$\varepsilon|x \sim N(0, \sigma^2 I)(12)$

The validity of these assumptions is needed for the results to be meaningful. If these assumptions are violated, the result can be incorrect and may have serious consequences. If these departures are small, the final result may not be changed significantly. But if the deviations are large, the model obtained may become unstable in the sense that a different sample could lead to an entirely different model with opposite conclusions. So such underlying assumptions have to be verified before attempting to regression modeling. One crucial point to keep in mind is that these assumptions are for the population, and we work only with a sample. So the main issue is to make a decision about the population on the basis of a sample of data. Several diagnostic methods to check the violation of regression assumption are based on the study of model residuals and also with the help of various types of graphics.

4.1Estimation of Parameters in a Multiple Linear Regression (MLR) Model

The method of least squares can be used to estimate the regression coefficients in Eq. (9). Suppose that $n > p$ observations are available, and let y_i denote the ith observed response and x_{ij} denote ith observation or level of regressor x_j. The data will appear in the following table 1. We also assume that the error term ε in the model has $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, and the errors are uncorrelated.

Table 1: Data for Multiple Linear Regression

		Regressors		
Observation, i	Response, y	x_1	x_2	x_p
1	y_1	x_{11}	x_{12}	x_{1p}
2	y_2	x_{21}	x_{22}	x_{2p}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	y_n	x_{n1}	x_{n2}	x_{np}

We may write the sample regression model corresponding to (9) as

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon$$

$= \beta_0 + \sum_{j=1}^{p} \quad \beta_j x_{ij} + \varepsilon_i, \forall\, i = 1,2,\ldots,n$

The least - squares function is then used to estimate the model parameters, which are obtained by minimizing the error sum of squares with respect to the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

It is more convenient to deal with multiple regression models if they are expressed in matrix notation. This allows a very compact display of the model, data, and results. In matrix notation, we can express the multiple regression model as

$y = X\beta + \varepsilon$(14)

Where

$$y = [y_1\ y_2\ .\ ..y_n\ ]X = \left[1\ x_{11}\ \cdots\ x_{1p}\ 1\ x_{21}\ \cdots\ x_{2p}\ \vdots \vdots \ddots \vdots\ 1\ x_{n1}\ \cdots\ x_{np}\ \right]\beta$$

$$= \left[\beta_0\ \beta_1\ .\ ..\beta_p\ \right]\varepsilon = [\varepsilon_1\ \varepsilon_2\ .\ ..\varepsilon_n\ ]$$

$y$is a $n \times 1$ vector of responses

$X$is a $n \times p$ matrix of the regressor variables

$\beta$is a $n \times 1$ vector of unknown constants, and

$\varepsilon$is a $n \times 1$ vector of random errors with $\varepsilon_i \sim NID(0, \sigma^2)$

We wish to find the vector of least-squares estimators, $\hat{\beta}$that minimizes

$S(\beta) = \sum_{i=1}^{n} \quad \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$

Note that $S(\beta)$ may be expressed as

$$S(\beta) = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta$$

$= y'y - 2\beta'X'y + \beta'X'X\beta$(16)

Since$\beta'X'y$is a1×1matrix, or a scalar, and its transpose $(\beta'X'y)' = y'X\beta$is the same scalar. The least square estimators must satisfy

$$\frac{\partial S}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0$$

Which simplifies

$X'X\hat{\beta} = X'y$ (17)

To solve the normal equations, multiply both sides of (iv) by the inverse of $X'X$. Thus the least squares estimator of

$\hat{\beta} = (X'X)^{-1}X'y$(18)

So, the vector of fitted values $\hat{y}_i$ corresponding to the observed value y_i is

$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$(19)

The difference between the observed value y_i and the corresponding fitted values $\hat{y}_i$ is the residual i.e., $e_i = y_i - \hat{y}_i$. The $n$residuals may be conveniently written in matrix notation as

$e = y - \hat{y}$(20)

3. Estimation of Error Term Variance (σ^2)

The variance σ^2 of the error terms ε_i in regression model needs to be estimated to know the variability of the probability distribution of y. In addition, a variety of inferences concerning the regression function and the prediction of y require an estimate of σ^2. Denote by $SSE = \sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \quad r_i^2$, is the residual sum of squares. Then an estimate of σ^2 is given by,

$\hat{\sigma}^2 = \frac{SSE}{n-p}$ (21)

where p is the total number of parameters involved in the model including the intercept term, if the model contains it. We also denote this quantity by MSE.

3. Inferences in Linear Regression Models

In multiple linear regression model, all variables may not be contributing significantly to the model. In other word, each of the parameters may not be significant. Therefore, these parameters must be tested whether they are significantly different from zero or not. That is, we test the null hypothesis (H_0) against the alternative hypothesis $(H_1)$for a parameter β_i (say) as follows:

$H_0: \beta_i = 0$

$H_1: \neq 0$

when $H_0: \beta_i = 0$is accepted we infer that there is no linear association between y and x_i. For normal error regression model, the condition β_i implies even more than no linear association between y and x_i. $\beta_i = 0$ for the normal error regression model implies not only that there is no linear association between y and x_i but also that there is no relation of any kind between y and x_i, since the probability distribution of $y$are then identical at all levels of x_i. The test is based on t test

$$t = \frac{\beta_i}{s(\beta_i)} \qquad (23)$$

where $s(\beta_i)$ is the standard error of $\beta_i$and calculated as $s(\beta_i) = \sqrt{\frac{MSE}{\sum_{i=1}^{n} \quad (x_i - \underline{x})^2}}$

The decision rule with this test statistic when controlling level of significance at α is

if $|t| \leq t\left(1 - \frac{\alpha}{2}; n - p\right)$ conclude H_0,

if $|t| > t\left(1 - \frac{\alpha}{2}; n - p\right)$ conclude H_1.

Similarly testing for other parameters can be carried out.

3. Measures of Fitting (R^2)

The overall fitting of a regression line can be judged by the F-statistic by carrying out an analysis of variance. If the F-statistic is significant, we say that our model is fitted well. However, there are times when the degree of linear association is of interest. A frequently used statistic is R^2. We describe this descriptive measure to describe the degree of linear association between y and x.

Denote by $TSS = \sum_i^n \quad \left(y_i - \underline{y}\right)^2$, total sum of squares which measures the variation in the observation y_i, or the uncertainty in predicting y, when no account of the predictor variable x is taken. Thus TSS is a measure of uncertainty in predicting y when x is not considered. Similarly, SSE measures the variation in the y_i when a regression model utilizing the predictor variable x is employed. A natural measure of the effect of x in reducing the variation in y, i.e., in reducing the uncertaintity in predicting y, is to express the reduction in variation ($TSS - SSE = SSR$ as a proportion of the total variation and it is denoted by

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad (24)$$

The measure R^2 is called coefficient of determination and $0 \leq R^2 \leq 1$. In practice R^2 is not likely to be 0 or 1 but somewhere between these limits. The closer it is to 1, the greater is said to be the degree of linear association between x and y. Remember that R^2 statistic should be used only when in the model an intercept term is involved. For the model with no intercept, $R^2$is not a good statistic. In case of "no intercept" model, sum of all residuals may not be equal to 0, making R^2 inflated.

3. An Illustration of a MLR model

Consider the following data:

Table 2: y as a response variable and x's as explanatory variables

Case No.	x_1	x_2	x_3	y	*Case No.*	x_1	x_2	x_3	Y
1	12.98	0.317	9.99	57.70	14	14.23	10.40	1.04	41.89
2	14.29	2.028	6.77	59.29	15	15.22	1.220	6.14	63.26
3	15.53	5.305	2.94	56.16	16	15.74	10.61	-1.91	45.79
4	15.13	4.738	4.20	55.76	17	14.95	4.815	4.11	58.69
5	15.3	7.038	2.05	51.72	18	14.12	3.153	8.45	50.08
6	17.14	5.982	-0.0	60.44	19	16.39	9.698	-1.7	48.89
7	15.46	2.737	4.65	60.71	20	16.45	3.912	2.14	62.21
8	12.80	10.66	3.04	37.44	21	13.53	7.625	3.85	45.62
9	17.03	5.132	0.25	60.97	22	14.19	4.474	5.11	53.92
10	13.17	2.039	8.73	55.27	23	15.83	5.753	2.08	55.79
11	16.12	2.271	2.10	59.28	24	16.56	8.546	8.97	56.74
12	14.34	4.077	5.54	54.02	25	13.32	8.589	4.01	43.14
13	12.92	2.643	9.33	53.19	26	15.94	8.290	-0.2	50.70

In the present example, we have 3 three predictor variables x_1, x_2 and x_3 and there are 26 observations. The response variable denoted by y. Applying least square method we obtain the parameter estimates as follows:

Table 3: ANOVA of a MLR model

Source	**Degrees of freedom**	**Sum of Square**	**Mean Square**	**F-value**	**Prob. > F**
Model	3	1062.34	354.11	109.69	<0.0001
Error	22	71.02	3.22		
Corrected Total	25	1133.37			

Table 4: Parameter Estimates of a MLR model

Variable	Degrees of freedom	Parameter Estimates	Standard Error	t-value	Prob. > |t|
Intercept	1	8.19	6.29	1.30	0.2060
x_1	1	3.56	0.36	9.86	<.0001
x_2	1	-1.64	0.15	-10.28	<.0001
x_3	1	0.33	0.17	1.88	0.0741

The value of $R^2$of this model is 0.93. From Table 3, we see that F-statistic is highly significant, indicating that overall model fitting is good. $R^2$is also very high. The fitted regression line is $\hat{y} = 8.19 + 3.56x_1 - 1.64x_2 + 0.33x_3$. The corresponding standard errors are given in the 4th column of Table 3. However, while testing the significance of the parameter estimates, we find that the intercept and the parameter for the variable x_3, i.e., are not significant at 5% level of significance (probability values for these parameters are greater than 0.05).

3. Practical Applications of regression analysis

Economics and Finance

Predicting stock market returns based on various economic indicators.

Analyzing the impact of interest rates on housing prices.

Marketing and Consumer Behavior

Understanding the factors influencing consumer purchasing decisions.

Predicting sales based on advertising expenditure and market demographics.

Healthcare and Medicine

Assessing the relationship between risk factors and disease outcomes.

Predicting patient outcomes based on treatment protocols and patient characteristics.

Agriculture

Crop Yield Prediction

Soil Fertility Assessment

Pest and Disease Management

Livestock Production

Economic Analysis and Market Forecasting, etc.

3. Conclusion

Correlation serves as a fundamental tool for analyzing relationships and unveiling hidden associations in data. By understanding the concept, measuring techniques, types, and interpretation of correlation, we can gain valuable insights and make informed decisions across a wide range of fields. Embracing correlation empowers us to unlock the intricate connections underlying the phenomena we observe, fostering a deeper understanding of the complex world around us.

Regression analysis serves as a versatile tool for understanding and predicting the relationship between variables. By comprehending the principles, assumptions, and types of regression analysis, we can harness its power to uncover patterns, make predictions, and inform decision-making across diverse fields. Embracing regression analysis empowers us to unravel the dynamics of complex systems, enabling us to navigate the intricacies of the world we inhabit with greater clarity and confidence.

3. References

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example*, New York: John Wiley & Sons.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, New York: Wiley Eastern Ltd.

Montgomery, D. C., Peck, E. and Vining, G. (2003). *Introduction to Linear Regression Analysis*, 3rd Edition, New York: John Wiley and Sons.

OVERVIEW OF SURVEY SAMPLING

Ankur Biswas
ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012
Ankur.Biswas@icar.gov.in

1. Introduction

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conservation and controversy require numerical data for their resolution. The data collected and analyzed in an objective manner and presented suitably serve as a basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include government, industry, business, research institutions, public organizations and international agencies and organizations. To discharge its various responsibilities, the government needs variety of information regarding different sectors of economy, trade, industrial production, health and mortality, population, livestock, agriculture, forestry, environment and available resources. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. For instance, the information on cost of living for different categories of people, living in various parts of the country is of importance in shaping its policies in respect of wages and price levels. Data on agricultural production are of immense use to the state for planning to feed the nation. In case of industry and business, the information is to be collected on labour, cost and quality of production, stock and demand and supply positions for proper planning of production levels and sales campaigns.

1.1 Complete enumeration

One way of obtaining the required information at regional and country level is to collect the data for each and every unit (person, household, field, factory, shop etc. as the case may be) belonging to the population which is the aggregate of all units of a given type under consideration and this procedure of obtaining information is termed as complete enumeration. The effort, money and time required for the carrying out complete enumeration to obtain the different types of data will, generally, be extremely large. However, if the information is required for each and every unit in the domain of study, a complete enumeration is clearly necessary. Examples of such situations are preparation of “voter list” for election purposes and recruitment of personnel in an establishment, etc. But there are many situations, where only summary figures are required for the domain of study as a whole or for group of units.

1.2 Need for sampling

An effective alternative to a complete enumeration can be sample survey where only some of the units selected in a suitable manner from the population are surveyed and an inference is drawn about the population on the basis of observations made on the selected units. It can be easily seen that compared to sample survey, a complete enumeration is time-consuming, expensive, has less scope in the sense of restricted subject coverage and is subject to greater coverage, observational and tabulation errors. In certain investigations, it may be essential to use specialized equipment or highly trained field staff for data collection making it almost impossible to carry out such investigations. It is of interest to note that if a sample survey is carried out according to certain specified statistical principles, it is possible not only to estimate the value of the characteristic of the population as a whole on the basis of the sample data, but also to get a valid estimate of the sampling error of the estimate. There are various steps involved in the planning and execution of the sample survey. One of the principal steps in a sample survey relates to methods of data collection.

1.3. Various concepts and definitions

i. Element:

An element is a unit about which we require information. For example, a field growing a particular crop is an element for collecting information on the yield of a crop.

ii. Population

The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

iii. Sampling unit

Elementary units or group of such units which besides being clearly defined, identifiable and observable, are convenient for the purpose of sampling are called sampling units. For instance, in a family budget enquiry, usually a family is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit.

iv. Sampling frame

A list of all the sampling units belonging to the population to be studied with their identification particulars or a map showing the boundaries of the sampling units is known as sampling frame. Examples of a frame are a list of farms and a list of suitable area segments like villages in India or counties in the United States. The frame should be up to date and free from errors of omission and duplication of sampling units.

v. Random sample

One or more sampling units selected from a population according to some specified procedures are said to constitute a sample. The sample will be considered as random or probability sample, if its selection is governed by ascertainable laws of chance. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection. For example, if a population consists of the N sampling units$U_1,U_2,\ldots,U_i,\ldots,U_N$ then, we may select a sample of n units by selecting them unit by unit with equal probability for every unit at each draw with or without replacing the sampling units selected in the previous draws.

vi. Non-random sample

A sample selected by a non-random process is termed as non-random sample. A non-random sample, which is drawn using certain amount of judgment with a view to get a representative sample, is termed as judgment or purposive sample. In purposive sampling units are selected by considering the available auxiliary information more or less subjectively with a view to ensuring a reflection of the population in the sample. This type of sampling is seldom used in large-scale surveys mainly because it is not generally possible to get strictly valid estimates of the population parameters under consideration and of their sampling errors due to the risk of bias in subjective selection and the lack of information on the probabilities of selection of the units.

vii. Population parameters

Suppose a finite population consists of the N units $U_1,U_2,\ldots,U_N$ and let Y_i be the value of the variable y, the characteristic under study, for the ith unit U_i, (i=1,2,…,N). For instance, the unit may be a farm and the characteristic under study may be the area under a particular crop. Any function of the values of all the population units is known as a population parameter or simply a parameter. Some of the important parameters usually required to be estimated in surveys are population total and population mean.

viii. Statistic, estimator and estimate

Suppose, a sample of n units is selected from a population of N units, according to some probability scheme and let, the sample observations be denoted by $y_1,y_2,\ldots,y_n$. Any function of these values which is free from unknown population parameters is called a statistic.An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample and the samples are selected with specified probabilities. The particular value, which the estimator takes for a given sample, is known as an estimate.

ix. Sampling and non-sampling error

The error arises due to drawing inferences about the population on the basis of observations on a part (sample) of it, is termed sampling error. The sampling error is non-existent in a complete enumeration survey since the whole population is surveyed. On the contrary, the errors other than sampling errors such as those arising through non-response, in- completeness and inaccuracy of response are termed non-sampling errors and are likely to be more wide-spread and important in a complete enumeration survey than in a sample survey. Non-sampling errors arise due to various causes right from the beginning stage when the survey is planned and designed to the final stage when the data are processed and analyzed.

The sampling error usually decreases with increase in sample size (number of units selected in the sample) while the non-sampling error is likely to increase with increase in sample size.

As regards the non-sampling error, it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former.

2. Simple Random Sampling

Simple random sampling (SRS) can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. Simple random sampling is a method of selecting n units out of the N such that every one of the $\binom{N}{n}$ distinct samples has an equal chance of being drawn. In practice a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. Since a number that has been drawn is removed from the population for all subsequent draws, this method is also called random sampling without replacement. In case of a random sampling with replacement, at any draw all N members of the population are given an equal chance of being drawn, no matter how often they have already been drawn. The with-replacement assumption simplifies the estimation under complex sampling designs and is often adopted, although in practice sampling is usually carried out under a without replacement type scheme. Obviously, the difference between with replacement and without replacement sampling becomes less important when the population size is large and the sample size is noticeably smaller than it.

2.1 Procedure of selecting a random sample

Since probability sampling theory is based on the assumption of random sampling, the technique of random sampling is of basic significance. Some of the procedures used for selecting a random sample are as follows:

i) Lottery method

ii) Use of random number tables

i) Lottery Method:

Each unit in the population may be associated with a chit/ticket such that each sampling unit has its identification mark from 1 to N. All the chits are placed in a container, drum or metallic spherical device, in which a thorough mixing is possible before each draw. Chits may be drawn one by one and may be continued until a sample of the required size is obtained. When the size of population is large, this procedure of numbering units on chits and selecting one after reshuffling becomes cumbersome. In practice, it may be too difficult to achieve a thorough shuffling. Human bias and prejudice may also creep in this method.

ii) Use of Random Number Tables:

A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern where each position is filled with one of these digits. A Table of random numbers is so constructed that all numbers 0, 1, 2,…,9 appear independent of each other. Some random number tables in common use are:

- Tippett's random number Tables
- Fisher and Yates Tables
- Kendall and Smith Tables
- A million random digits Table

A practical method of selecting a random sample is to choose units one-by-one with the help of a Table of random numbers. By considering two-digit numbers, we can obtain numbers from 00 to 99, all having the same frequency. Similarly, three or more digit numbers may be obtained by combining three or more rows or columns of these Tables. The simplest way of selecting a sample of the required size is to select a random number from 1 to N and then taking the unit bearing that number. This procedure involves a number of rejections since all numbers greater than N appearing in the Table are not considered for selection. The procedure of selection of sample through the use of random numbers is, therefore, modified and one of these modified procedures is:

- **Remainder Approach:**

Let N be an r-digit number and let its r-digit highest multiple be N'. A random number k is chosen from 1 to N' and the unit with serial number equal to the remainder obtained on dividing k by N is selected, *i.e.* the selected number is reduced mod (N). If the remainder is zero, the last unit is selected. As an illustration, let N = 123, then highest three-digit multiple of 123 is 984. For selecting a unit, one random number from 001 to 984 has to be selected. Let the random number selected be 287. Dividing 287 by 123 gives the remainder as 41. Hence, the unit with serial number 41 is selected in the sample. Suppose that another random number selected is 245. Dividing 245 by 123 leaves 122 as remainder. So the unit bearing the serial number 122 is selected. Similarly, if the random number selected is 369, then dividing 369 by 123 leaves remainder as 0. So the unit bearing serial number 123 is selected in the sample.

2.2 Estimation of Population Total

Let Y be the character of interest and $Y_1, Y_2, \cdots, Y_i, \cdots, Y_N$ be the values of the character from*N* units of the population. Further, let $y_1, y_2, \cdots, y_i, \cdots, y_n$ be the sample of size n selected by simple random sampling without replacement. For the total $Y = \sum_{i=1}^{N} Y_i$ we have an estimator

$$\hat{Y} = N \sum_{i=1}^{n} y_i / n = N\overline{y}_n$$

i.e., the sample mean $\overline{y}_n$ multiplied by the population size N.

The estimator can be expressed as

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i = \left(N/n\right) \sum_{i=1}^{n} y_i \text{ , where } w_i = N/n.$$

The constant N/n is the sampling weight and is the inverse of the sampling fraction $n/N.$

The estimator has the statistical property of unbiasedness in relation to the sampling design. Variance of the estimator $\hat{Y}$ of the population total is given by

$$V_{SRS}\ (\hat{Y}) = \frac{N^2}{n} \left(1\text{-}\frac{n}{N}\right) \sum_{i=1}^{N} \ (Y_i - \overline{Y})^2 \big/ (N-1)$$

where $\overline{Y} = \sum_{i=1}^{N} Y_i / N$ is the population mean and $S^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 / (N-1)$ is the population mean square.

An unbiased estimator of variance of the estimator $\hat{Y}$ of the total, $V_{SRS}(\hat{Y})$ is given by

$$\begin{aligned} \hat{V}_{SRS}(\hat{Y}) &= N^2 \left(1\text{-}\frac{n}{N}\right) \sum_{i=1}^{n} (y_i - \overline{y}_n)^2 / n(n-1) \\ &= N^2 \left(1\text{-}\frac{n}{N}\right) s^2 / n \end{aligned}$$

where $\overline{y}_n = \sum_{i=1}^{n} y_i / n$ is the sample mean and s^2 is an unbiased estimator of the population mean square S^2.

3. Use of Auxiliary Information

In sampling theory if the auxiliary information, related to the character under study, is available on all the population units, then it may be advantageous to make use of this additional information in survey sampling. One way of using this additional information is in the sample selection with unequal probabilities of selection of units. The knowledge of auxiliary information may also be exploited at the estimation stage. The estimator can be developed in such a way that it makes use of this additional information. Ratio estimator, difference estimator, regression estimator, generalized difference estimators are the examples of such estimators. Obviously, it is assumed that the auxiliary information is available on all the sampling units. In case the auxiliary information is not available then it can be obtained easily without much burden on the cost.

Another way the auxiliary information can be used is at the stage of planning of survey. An example of this is the stratification of the population units by making use of the auxiliary information.

4. Sampling with Varying Probability

Under certain circumstances, selection of units with unequal probabilities provides more efficient estimators than equal probability sampling, and this type of sampling is known as unequal or varying probability sampling. In the most commonly used varying probability sampling scheme, the units are selected with probability proportional to a given measure of size (PPS) where the size measure is the value of an auxiliary variable x related to the characteristic y under study and this sampling scheme is termed as probability proportional to size sampling. For instance, the number of persons in some previous period may be taken as a measure of the size in sampling area units for a survey of socio-economic characters, which are likely to be related to population. Similarly, in estimating crop characteristics the geographical area or cultivated area for a previous period, if available, may be considered as a measure of size, or in an industrial survey, the number of workers may be taken as the size of an industrial establishment.

Since a large unit, that is, a unit with a large value for the study variable y, contributes more to the population total than smaller units, it is natural to expect that a scheme of selection which gives more chance of inclusion in a sample to larger units than to smaller units would provide estimators more efficient than equal probability sampling. Such a scheme is provided by pps sampling, size being the value of an auxiliary variable x directly related to y. It may appear that such a selection procedure would give biased estimators as the larger units are over-represented and the smaller units are under-represented in the sample. This would be so, if the sample means is used as an estimator of population mean. Instead, if the sample observations are suitably weighted at the estimation stage taking into consideration their probabilities of selection, it is possible to obtain unbiased estimators. Mahalanobis (1938) has referred to this procedure in the context of sampling plots for a crop survey and this procedure has been discussed in detail by Hansen and Hurwitz (1943).

5. Stratified Random Sampling

The basic idea in stratified random sampling is to divide a heterogeneous population into sub-populations, usually known as strata, each of which is internally homogeneous in which case a precise estimate of any stratum mean can be obtained based on a small sample from that stratum and by combining such estimates, a precise estimate for the whole population can be obtained. Stratified sampling provides a better cross section of the population than the procedure of simple random sampling. It may also simplify the organization of the field work. Geographical proximity is sometimes taken as the basis of stratification. The assumption here is that geographically contiguous areas are often more alike than areas that are far apart. Administrative convenience may also dictate the basis on which the stratification is made. For example, the staff already available in each range of a forest division may have to supervise the survey in the area under their jurisdiction. Thus, compact geographical regions may form the strata. If the characteristic under study is known to be correlated with a supplementary variable for which actual data or at least good estimates are available for the units in the population, the stratification may be done using the information on the supplementary variable. For instance, the volume estimates obtained at a previous inventory of the forest area may be used for stratification of the population.

In stratified sampling, the variance of the estimator consists of only the ‘within strata’ variation. Thus the larger the number of strata into which a population is divided, the higher, in general, the precision, since it is likely that, in this case, the units within a stratum will be more homogeneous. For estimating the variance within stratum, there should be a minimum of 2 units in each stratum. The larger the number of strata the higher will, in general, be the cost of enumeration. So, depending on administrative convenience, cost of the survey and variability of the characteristic under study in the area, a decision on the number of strata will have to be arrived at.

6. Cluster Sampling

A sampling procedure presupposes division of the population into a finite number of distinct and identifiable units called the sampling units. The smallest units into which the population can be divided are called the elements of the population, and group of elements the clusters. A cluster may be a class of students or cultivators’ fields in a village. When the sampling unit is a cluster, the procedure of sampling is called cluster sampling.

For many types of population a list of elements is not available and the use of an element as the sampling unit is, therefore, not feasible. The method of cluster or area sampling is available in such cases. Thus, in a city a list of all the houses may be available, but that of persons is rarely so. Again, list of farms are not available, but those of villages or enumeration districts prepared for the census are. Cluster sampling is, therefore, widely practiced in sample surveys.

For a given number of sampling units cluster sampling is more convenient and less costly than simple random sampling due to the saving time in journeys, identification and contacts etc., but cluster sampling is generally less efficient than simple random sampling due to the tendency of the units in a cluster to be similar. In most practical situations, the loss in efficiency may be balanced by the reduction in the cost and the efficiency per unit cost may be more in cluster sampling as compares to simple random sampling.

7. Multistage Sampling

Cluster sampling is a sampling procedure in which clusters are considered as sampling units and all the elements of the selected clusters are enumerated. One of the main considerations of adopting cluster sampling is the reduction of travel cost because of the nearness of elements in the clusters. However, this method restricts the spread of the sample over population which results generally in increasing the variance of the estimator. In order to increase the efficiency of the estimator with the given cost it is natural to think of further sampling the clusters and selecting more number of clusters so as to increase the spread of the sample over population. This type of sampling which consists of first selecting clusters and then selecting a specified number of elements from each selected cluster is known as sub-sampling or two stage sampling, since the units are selected in two stages. In such sampling designs, clusters are generally termed as first stage units (fsu’s) or primary stage units (psu’s) and the elements within clusters or ultimate observational units are termed as second stage units (ssu’s) or ultimate stage units (usu’s). It may be noted that this procedure can be easily generalized to give rise to multistage sampling, where the sampling units at each stage are clusters of units of the next stage and the ultimate observational units are selected in stages, sampling at each stage being done from each of the sampling units or clusters selected in the previous stage. This procedure, being a compromise between uni-stage or direct sampling of units and cluster sampling, can be expected to be (i) more efficient than uni-stage sampling and less efficient than cluster sampling from considerations of operational convenience and cost, and (ii) less efficient than uni-stage sampling and more efficient than cluster sampling from the view point of sampling variability, when the sample size in terms of number of ultimate units is fixed.

It may be mentioned that multistage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is prohibitive or where the cost of locating and physically identifying the usu’s is considerable. For instance, for conducting a socio-economic survey in a region, where generally household is taken as the usu, a complete and up-to-date list of all the households in the region may not be available, whereas a list of villages and urban blocks which are group of households may be readily available. In such a case, a sample of villages or urban blocks may be selected first and then a sample of households may be drawn from each selected village and urban block after making a complete list of households. It may happen that even a list of villages is not available, but only a list of all tehsils (group of villages) is available. In this case a sample of households may be selected in three stages by selecting first a sample of tehsils, then a sample of villages from each selected tehsil after making a list of all the villages in the tehsil and finally a sample of households from each selected village after listing all the households in it. Since the selection is done in three stages, this procedure is termed as three stage sampling. Here, tehsils are taken as first stage units (fsu’s), villages as second stage units (ssu’s) and households as third or ultimate stage units (tsu’s).

8. Systematic Sampling

In all other sampling methods, the successive units (whether elements or clusters) are selected with the help of random numbers. But a method of sampling in which only the first unit is selected with the help of random number while the rest of the units are selected according to a pre-determined pattern, is known as systematic sampling. The systematic sampling has been found very useful in forest surveys for estimating the volume of timber, in fisheries surveys for estimating the total catch of fish, in milk yield surveys for estimating the lactation yield etc.

9. Conclusion

Simple random sampling and probability proportional size designs are most important uni-stage design. In most of the practical situations, complex sampling designs are utilized on the basis of these uni-stage sampling designs. Stratified random sampling, multistage sampling, multiphase sampling, etc. are efficient complex designs widely used in agricultural and socio-economic surveys.

References

Cochran, W.G. (1977). *Sampling techniques*. Wiley Eastern Ltd.

Des Raj, (1968). *Sampling theory*. Tata-Mcgraw-Hill Publishing Company Ltd.

Hansen, M.H. and Hurwitz, W.H. (1943). On the theory of sampling from finite populations.*Ann. Math. Statist.*, **14**, 333-362.

Hansen, M.H., Hurwitz, W.H. and Madow, W.G. (1993). *Sample survey methods and theory*. Vol. 1 and Vol. 2, John Wiley & Sons, Inc.

Murthy, M.N. (1977). *Sampling theory and methods*. Statistical Publishing Society.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984). *Sampling theory of surveys with applications*. Indian Society of Agricultural Statistics.

INTRODUCTION AND OVERVIEW OF THE NONLINEAR GROWTH MODEL

Mrinmoy Ray
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
mrinmoy.ray@icar.gov.in

1. Introduction

Growth is defined as "an irreversible increase in size and volume that occurs as a result of differentiation and distribution in the plant/animal." A model is a schematic representation of a system's conception, an act of mimicry, or a set of equations that represents a system's behaviour. A model is also defined as "a representation of an object, system, or idea in a form other than that of the entity itself." Its purpose is typically to aid in the explanation, comprehension, or improvement of a system's performance.

TYPES OF MODELS

Models are classified into different groups or types based on the purpose for which they are designed. Among them are a few:

a. Statistical models: These models describe the relationship between. Relationships are measured in a system using statistical techniques in these models. Example: regression model, Time series model, etc.

b. Mechanistic models: These models explain not only the relationship between variables, but also how these models work (explains the relationship of influencing dependent variables). Physical selection is the basis for these models.

c. Deterministic models: The exact value of the dependent variable is estimated using these models. These models have defined coefficients as well.

d. Stochastic models: Each output has a probability element attached to it. Different outputs, along with probabilities, are provided for each set of inputs. At a given rate, these models define the state of the dependent variable.

e. Dynamic models: Time is accounted for as a variable. Both dependent and independent variables have values that remain constant over a given time period.

f. Static: Time is not considered a variable. Dependent and independent variables with values that remain constant over time.

g. Simulation models: In general, computer models are mathematical representations of real-world systems. Crop simulation models' primary goal is to estimate agricultural production as a function of weather and soil conditions, as well as crop management. These models employ one or more sets of differential equations to compute rate and state variables over time, typically from planting to harvest maturity or final harvest.

Statistical Modelling

A fundamental problem in statistics is developing models based on a sample of observations and making inferences based on the model. Huge amounts of data pertaining to crop production/productivity, import-export of various agricultural commodities, and so on are being collected sequentially over time in almost all branches of agriculture, including animal sciences and fisheries. One feature of such data is that successive observations are dependent on one another. Each observation of the observed data series, Y_t,may be considered as a realization of a stochastic process $\{Y_t\}$, which is a family of random variables $\{Y_t,\ t \in T\}$, where $T = \{\ 0, \pm 1, \pm 2, \ldots\}$, and apply standard time-series approach to develop an ideal model which will adequately represent the set of realizations and also their statistical relationships in a satisfactory manner. Forecasting of time-series data is critical for planners and policymakers. Over the last few decades, a new field known as "Nonlinear time-series modelling" has emerged. There are essentially two approaches available here: parametric or nonparametric. Obviously, we should use the former if we are certain about the functional form in a given situation; otherwise, the latter may be used.

Parametric and Nonparametric Approaches

Regression analysis has grown in popularity as a tool for statistical modelling and data analysis over the last several decades. This information describes the relationship between a response variable and one or more predictor variables. The primary goal is to express the mean of the response as a function of the predictor variables. The general regression model takes the following form:

$$Y \quad = \quad m\,(X) + \ \varepsilon$$

Where Y is the response variable, $m(X) = E\ (Y|\ X)$ is the mean response or regression function and ε is the error. The regression function $m(X)$ is usually unknown and the objective is to obtain a suitable estimator of $m(X)$ using a sample of observations.

In the linear regression, it is assumed that the mean of the response variable Y is a linear function of predictor variable(s) X of the form

$$E\ (Y|\ X) \ = \ X\beta$$

i.e. m (X) is linear in parameters. The parameter vector β is usually estimated by the Method of least squares. In nonlinear regression, it is assumed that the mean of the response variable is a nonlinear function of the predictor variable (s) X of the form

$E(Y|X)=m(X,\beta)$

i.e.$m(X)$ is nonlinear in parameters. Generally, there will be no closed form expression for the estimates of β and iterative procedures are required for estimation of parameters.

A parametric regression model (linear or nonlinear) assumes that the form of m is known with the exception of some unknown parameters, and that the shape of the regression function is entirely dependent on the parameters. It is frequently difficult to guess the most appropriate functional form simply by looking at the data. There may be times when no suitable parametric form exists to express the regression function. In such cases, the nonparametric regression approach is very useful because it does not require strong assumptions about the shape of the regression function. A nonparametric regression model only assumes that m is part of an infinitely large collection of functions. One limitation of the preceding approach is that it generally relies on certain assumptions about the smoothness of the function being estimated, which may or may not be true in practice. As a result, the data under consideration may be over smoothed.

LINEAR MODEL

A mathematical model is an equation or set of equations that represents a system's behaviour. It can be 'linear' or 'nonlinear.' A linear model is one in which all of the parameters appear linearly.

NONLINEAR MODELS

Any type of statistical investigation in which principles from a body of knowledge are seriously considered in the analysis is likely to result in a 'Nonlinear model.' Such models are critical in understanding the complex interrelationships between variables. A ‘nonlinear model’ is one in which at least one of the parameters appears nonlinearly. More formally, in a ‘nonlinear model’, at least one derivative with respect to a parameter should involve that parameter.

- Examples of a nonlinear model are:

$$Y(t) = \exp(at+bt^2) \qquad (1a)$$

$$Y(t) = at + \exp(-bt) \qquad (1b)$$

Note. Some authors use the term ‘intrinsically nonlinear’ to indicate a nonlinear model which can be transformed to a linear model by means of some transformation. For example, the model given by Eq. (1a) is ‘intrinsically nonlinear’ in view of the transformation $X(t) = \log_e Y(t)$.

a. MALTHUS MODEL:

Thomas R. Malthus, an Englishman, proposed a mathematical model of population growth in 1798. Despite its simplicity, the model has become the foundation for most future modelling of biological populations. His essay, "An Essay on the Principle of Population," contains an excellent discussion of the limitations of mathematical modelling and should be required reading for all serious students of the subject. Malthus observed that, if not restrained by environmental or social constraints, human populations appeared to double every twenty-five years, regardless of initial population size. In other words, he proposed that populations increased by a fixed proportion over a given period of time and that, in the absence of constraints, this proportion was unaffected by population size. According to Malthus, if a population of 100 people increased to a population of 135 people over the course of, say, five years, then a population of 1000 people would increase to 1350 people over the same period of time. Malthus' model is an example of a one-variable, one-parameter model. The quantity we are interested in observing is referred to as a variable. They typically evolve over time. Parameters are quantities known to the modeller before the model is built. They are frequently constants, though a parameter can change over time. The variable in the Malthusian model is population, and the parameter is population growth rate.

If N(t) denotes the population size or biomass at time t and r is the intrinsic growth rate, then the rate of growth of population size is given by

$dN/dt = rN$

Therefore, $N(t) = N_o \exp(rt)$

Note : Malthus model can be used for describing growth of simplistic organisms, which begin to grow by binary splitting of cells.

Drawback:$N(t)\rightarrow\infty$ as $t\rightarrow\infty$, which cannot happen in reality.

Malthus predicted that unchecked population growth would quickly outstrip carrying capacity, resulting in overpopulation and social problems.

a. MONOMOLECULAR MODEL:

Because the monomolecular model assumes a carrying capacity of one, which means that the maximum level of disease is one, disease severity or incidence is measured as a proportion. Plant tissue that is diseased may only have a value between zero (healthy) and one (complete disease).It also assumes the absolute rate of change is proportional to the healthy tissue i.e., (1-*y*).

It describes growth progress in which it is assumed that the rate of growth at any point in time is proportional to the resources yet to be obtained, i.e.

$dN/dt = r(K–N)$,

where K is the carrying capacity.

or $N(t) = K– (K–N_o) \exp (–rt)$

Drawback: No point of inflexion.

a. LOGISTIC MODEL:

Logistic model was developed by Belgian mathematician Pierre Verhulst (1838) who suggested that the rate of population increase may be limited, i.e., it may depend on population density. Population growth rate declines with population numbers, N, and reaches 0 when N = K. Parameter K is the upper limit of population growth and it is called carrying capacity. It is commonly interpreted as the amount of resources expressed in the number of organisms that these resources can support. If the population exceeds K, the population growth rate becomes negative and the population decreases.

The differential equation represents this model:

$dN/dt = rN (1–N/K)$ (1)

Therefore, $N(t) = K/[1+(K/N_o–1) \exp(–rt)]$. The graph of N(t) versus t is elongated S-shaped and the curve is symmetrical about its point of inflexion.

a. GOMPERTZ MODEL

This is another model with sigmoid behaviour that has been found to be quite useful in biological work. Benjamin Gompertz developed the Gompertz curve to estimate human mortality (Gompertz, B. "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies." Phil. Trans. Roy. Soc. London 123, 513-585, 1832). An early description of the use of this equation to describe growth processes is given by

CharlesWinsor (1932). However, unlike the logistic model, this does not have a symmetric point of inflexion.

This model's differential equation is

$dN/dt = rN \log_e (K/N)$ (2)

or $N(t) = K \exp[\log_e (N_o / K) \exp(-rt)]$

a. RICHARDS MODEL:

The Richards curve, also known as generalised logistic, is a popular growth model that can fit a wide range of S-shaped growth curves. Both 4 and 5 parameter versions are commonly used. The logistic curve is symmetrical about its point of inflection. Richards (1959) introduced an additional parameter to deal with asymmetrical growth curves.

This model is given by

$$N(t) = K N_o / [N_o + (K^m - N_o^m) \exp(-rt)]^{1/m} \quad . \qquad (4)$$

However, unlike the earlier models, this model has four parameters.

Drawback. Number of parameters is more.

a. MIXED-INFLUENCE MODEL:

This is a mixture of ‘Monomolecular’ and ‘Logistic’ Models. It is given by

$dN/dt = r (K-N) + s N (1-N/K)$,

FITTING OF NONLINEAR MODELS

The models presented above have been posed deterministically. This is obviously unrealistic, so we replace these deterministic models with statistical models by including an error term on the right hand side and making appropriate assumptions about them. This produces a 'Nonlinear statistical model.' The 'Method of least squares' can be used to estimate parameters in non-linear regression, just as it can in linear regression. However, minimising the residual sum of squares produces normal equations with nonlinear parameters. Because exact solutions to nonlinear equations are not possible, iterative procedures are used to obtain approximate analytic solutions.

- Four main methods of this kind are:

  i) Linearization (or Taylor Series) method

  ii) Steepest Descent method

  iii) Levenberg-Marquardt’s method

  iv) Do not use Derivatives method

Draper and Smith discuss the specifics of these methods, as well as their benefits and drawbacks (1998). Neither the Linearization nor the Steepest descent methods are perfect. The Levenberg-Marquardt method is the most widely used method for computing nonlinear least squares estimates. This method is a compromise between the other two methods, successfully combining the best features of both while avoiding their significant disadvantages. It's good because it almost always converges and doesn't' slow down' at the end of the iterative process.

CHOICE OF INITIAL VALUES

All nonlinear estimation procedures require initial parameter values, and selecting good initial values is critical. There is, however, no standard procedure for obtaining preliminary estimates. The use of prior information is the most obvious method for making initial guesses. Estimates based on previous experiments, known values for similar systems, and values derived from theoretical considerations all combine to form ideal first guesses.

Some other methods are:

(i) Linearization:

After ignoring the error term, check the form of the model to see if it could be transformed into a linear form by means of some transformation. In such cases, linear regression can be used to obtain initial values.

(ii) Solving a system of equations:

If there are p parameters, substitute for p sets of observations into the model ignoring the error. Solve these equations for the parameters, if possible. Widely separated x_i often work best.

R code

Monomolecular growth model

```
z=read.csv(file.choose(), header=TRUE)
head(z)
kk=data.frame(z)
grz1=nls(y~k-(k-y0)*exp(-r*t),data=kk,  start=list(k=1 ,y0=0.03,r=0.1))
summary(grz1)
 fitted=kk$y-resid(grz1)
kkk=data.frame(fitted)
MSE.nn<- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors<- c("blue","red")
```

```
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,1),axes=FALSE, ann=FALSE)
axis(1, at=1:20, lab=c(0:19))
axis(2, las=1, at=0.2*0:5)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22,
lty=1:2);
zz=resid(grz1)
predicted= 0.99651-(0.99651-0.08844)*exp(-0.26727*20)
```

Gompertz model

```
z=read.csv(file.choose(), header=TRUE)
 head(z)
 kk=data.frame(z)
gr1=nls(y~k*exp(log(y0/k)* exp(-r*t)),data=kk,  start=list(k=50,y0=11.72,r=0.1))
summary(gr1)
fitted=kk$y-resid(gr1)
kkk=data.frame(fitted)
MSE.nn<- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors<- c("blue","red")
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,35),axes=FALSE, ann=FALSE)
axis(1, at=1:38, lab=c(0:37))
axis(2, las=1, at=5*0:8)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22,
lty=1:2);
```

logistic model

```
z=read.csv(file.choose(), header=TRUE)
```

```
 head(z)
 kk=data.frame(z)
gr2=nls(y~k/(1+(k/y0-1)* exp(-r*t)), data=kk,  start=list(k=50,y0=11.72,r=0.1))
summary(gr2)
fitted=kk$y-resid(gr2)
kkk=data.frame(fitted)
MSE.nn<- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors<- c("blue","red")
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,35),axes=FALSE, ann=FALSE)
axis(1, at=1:38, lab=c(0:37))
axis(2, las=1, at=5*0:8)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22,
lty=1:2);
```

LOGIT AND PROBIT ANALYSIS

Himadri Shekhar Roy
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
himadri.roy@icar.gov.in

1. Introduction

Regression analysis is a technique used to examine the relationships between variables. These relationships are expressed through equations or models that connect a response or dependent variable with one or more explanatory or predictor variables. Typically, the variables involved in regression analysis are quantitative in nature. The estimation of parameters in this type of analysis relies on four key assumptions. The first assumption is that the response variable is linearly related to the explanatory variables. In other words, there is a linear relationship between the dependent variable and the predictors. The second assumption is that the errors in the model are independently and identically distributed, following a normal distribution with a mean of zero and a common variance. This assumption ensures that the errors are random and have a consistent distribution. The third assumption assumes that the explanatory variables are measured without any errors. This means that the predictor variables are accurate and reliable. The last assumption relates to the equal reliability of observations. It assumes that each observation used in the analysis is equally reliable and contributes equally to the analysis. In cases where the response variable in the model is qualitative, instead of directly modeling the response variable itself, probabilities of belonging to different categories can be modelled using the same regression framework. However, this approach comes with additional constraints and assumptions for multiple regression models. The first constraint is that probabilities range between 0 and 1, while the right-hand side function in multiple regression models is unbounded. This means that adjustments need to be made to ensure that the predicted probabilities remain within the valid range. The second constraint is related to the error term of the model. In this case, the error term can only take limited values, and the variance of the errors is not constant but depends on the probability of the response variable falling into a particular category. There are several notable references available that provide a comprehensive overview of logistic regression, such as the works of Fox (1984) and Klienbaum (1994). For Probit analysis, a useful resource is Finney (1971).

2. Assumptions of Linear Regression Model if Response is Qualitative

To illustrate the limitations of using linear regression when the response variable is qualitative, let's examine a simple linear regression model that involves a single predictor variable and a binary response variable.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{ , } i = 1, 2, \dots, n$$

where, the outcome Y_i is binary (taking values 0,1), $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and are independent and n is the number of observations.

Let π_i denote the probability that Y_i =1 when X_i = x, i.e.

$$\pi_i = P(Y_i = 1 | X_i = x) = P(Y_i = 1)$$

thus $\quad P(Y_i = 0) = 1 - \pi_i$.

Under the assumption $E(\varepsilon_i) = 0$, the expected value of the response variable is

$$E(Y_i) = 1.(\pi_i) + 0.(1 - \pi_i) = \pi_i$$

If the response is binary, then the error terms can take on two values, namely,

$$\varepsilon_i = 1 - \pi_i \qquad \text{when } Y_i = 1$$

$$\varepsilon_i = -\pi_i \qquad \text{when } Y_i = 0$$

Because the error is dichotomous (discrete), normality assumption is violated. Moreover, the error variance is given by:

$$\begin{aligned} V(\varepsilon_i) &= \pi_i (1 - \pi_i)^2 + (1 - \pi_i)(-\pi_i)^2 \\ &= \pi_i (1 - \pi_i) \end{aligned}$$

It can be seen that variance is a function of π_i's and it is not constant. Therefore, the assumption of homoscedasticity (equal variance) does not hold.

3. Logistic regression

3.1 Binary Logistic regression

Logistic regression is often recommended when the multivariate normality assumption is not met by the independent variables and the response variable is qualitative. This situation, where the response variable is qualitative and the independent variables include a mix of categorical and continuous variables, is commonly encountered in statistical applications such as agriculture and medical science. The binary logistic regression model, developed by researcher Cox in the late 1950s, is the preferred statistical model for analysing binary (dichotomous) responses. Agricultural data often exhibit sigmoidal or elongated S-shaped curves, making logistic regression models more appropriate. These models can capture non-linear relationships between the response variable and the qualitative and quantitative factors that influence it. Logistic regression addresses similar questions as discriminant function analysis and multiple regression, but it does not rely on distributional assumptions for the predictors. In other words, the predictors do not need to follow a normal distribution, the relationship between the response and predictors can be non-linear, and the observations do not need to have equal variance in each group. For a comprehensive understanding of logistic regression, informative resources can be found in the works of Fox (1984) and Kleinbaum (1994).

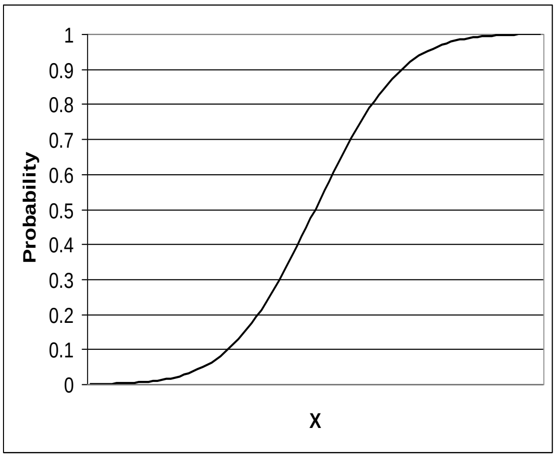
The issue of non-normality and heteroscedasticity, as discussed in section 2, renders least square estimation unsuitable for the linear probability model. When attempting to use weighted least square estimation as an alternative, the resulting fitted values may not be constrained within the interval (0, 1), making them inappropriate for interpretation as probabilities. Furthermore, there is a possibility of negative error variances arising. To address this problem, one solution is to constrain the values of π (the response variable) to the unit interval while still maintaining the linear relationship between π and the regressor X within that interval. By doing so, we can ensure that the predicted values of π remain within the valid range of probabilities.

$$\pi = \begin{cases} 0 & ,\beta_0 + \beta_1 X < 0 \\ \beta_0 + \beta_1 X & ,0 \le \beta_0 + \beta_1 X \le 1 \\ 1 & ,\beta_0 + \beta_1 X > 1 \end{cases}$$

However, this constrained linear probability model has certain unattractive features such as abrupt changes in slope at the extremes 0 and 1 making it hard for fitting the same on data. A smoother relation between π and X is generally more sensible. To correct this problem, a positive monotone (i.e. non-decreasing) function is required to transform ($\beta_0 + \beta_1 x_i$) to unit interval. Any cumulative probability distribution function (CDF) P, meets this requirement. That is, respecify the model as $\pi i = P(\beta_0 + \beta_1 x_i)$. Moreover, it is advantageous if P is strictly increasing, for then, the transformation is one-to-one, so that model can be rewritten as $P^{-1}(\pi i) = (\beta 0 + \beta 1xi)$, where P^{-1} is the inverse of the CDF P. Thus the non-linear model for itself will become both smooth and symmetric, approaching $\pi = 0$ and $\pi = 1$ as asymptotes. Thereafter maximum likelihood method of estimation can be employed for model fitting.

3.2 Properties of Logistic Regression Model

The logistic response function exhibits a characteristic S-shaped curve, which can be visualized in the accompanying figure. As X increases, the probability π initially experiences a gradual increase, followed by a rapid acceleration. Eventually, the increase in probability tapers off and stabilizes, but it never exceeds the value of 1.



The shape of the S-curve can be reproduced if the probabilities can be modeled with only one predictor variable as follows:

$\pi = P(Y=1|X=x) = 1/(1+e^{-z})$

where $z = \beta_0 + \beta_1 x$, and e is the base of the natural logarithm. Thus for more than one (say r) explanatory variables, the probability π is modeled as

$\pi = P(Y=1|X_1 = x_1 ... X_r = x_r)$

$=1/(1+e^{-z})$

where $z = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r$.

This equation is called the logistic regression equation. It is nonlinear in the parameters β_0, β_1… β_r. Modeling the response probabilities by the logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression. The method of estimation generally used is the maximum likelihood estimation method.

To explain the popularity of logistic regression, let us consider the mathematical form on which the logistic model is based. This function, called f (z), is given by

$$f(z) = 1/(1+e^{-z}), -\infty < z < \infty$$

Now when $z = -\infty$, $f(z) = 0$ and when $z = \infty$, $f(z) = 1$. Thus the range of f (z) is 0 to1. So the logistic model is popular because the logistic function, on which the model is based, provides. Estimates that lie in the range between zero and one.

An appealing S-shaped description of the combined effect of several explanatory variables on the probability of an event.

3.6 Multinomial logistic regression modeling

Let **X** is a vector of explanatory variables and π denotes the probability of binary response variable then logistic model is given by

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \mathbf{X}\beta = g(\pi)$$

where, ‘alpha’ is the intercept parameter and ‘beta’ is a vector of slope parameters. In case response variable has ordinal categories say 1,2,3,--------, I, I+1 then generally logistic model is fitted with common slope based on cumulative probabilities of response categories instead of individual probabilities. This provides parallel lines of regression model with following form

g [Prob ($\mathbf{y} \le \mathbf{i}(\mathbf{x})$)] $\bar{\bar{\alpha}}_i + \chi\beta$, $1 \le i \le I$

where, $\alpha_1, \alpha_2, ------- \alpha_{\mathbf{k}},$ are k intercept parameters and β is the vector of slope parameters.

Multinomial logistic regression (taking qualitative response variable with three categories, for simplicity) is given by

logit[Pr(Y ≤ j – 1 / **X**)] = α_j + $\boldsymbol{\beta}^T$ **X** , j = 1,2

where α_j are two intercept parameters ($\alpha_1 < \alpha_2$), $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots\ldots, \beta_k)$ is the slope parameter vector not including the intercept terms, $\mathbf{X}^T = (X_1, X_2, \ldots, X_k)$ is vector of explanatory variables. This model fits a common slope cumulative model i.e. ‘parallel lines’ regression model based on the cumulative probabilities of the response categories.

logit(π_1) = $$\log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k,$$

logit($\pi_1 + \pi_2$) = $$\log\left(\frac{\pi_1+\pi_2}{1-\pi_1-\pi_2}\right) = \alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k$$

π_j (X) denotes classification probabilities Pr(Y=j-1 / X) of response variable Y, j = 1,2,3, at X^T.

These models can be fitted through maximum likelihood procedure.

4. Probit analysis

4.1 Introduction

Probit analysis is widely utilized in various fields when the response variable is qualitative. One of its main applications is observed in toxicological studies, where it transforms the sigmoid dose-response curve into a linear relationship that can be analyzed using regression techniques like least squares or maximum likelihood. In essence, probit analysis is a methodology that converts the complex relationship between the percentage affected and the dose response into a linear relationship between probit and the dose response. The probit values can then be translated back into percentages. This approach is appropriate because of the typical shape exhibited by dose-response curves. While the method is approximate, it enables the quantification of consequences resulting from exposure. The term "probit" originates from the phrase "probability unit" and was coined by Bliss. It was the first model developed and studied for analyzing data such as the percentage of pests killed by a pesticide.

4.2 Probit Model

In the realm of probability theory and statistics, the probit function represents the inverse of the cumulative distribution function (CDF) linked to the standard normal distribution. Alternatively, one can consider the logistic distribution, which results in the logit or logistic model. Both the logistic and probit curves are highly similar, producing almost indistinguishable outcomes. In practice, they provide estimated probabilities that exhibit very little variation (Aldrich and Nelson, 1984). The selection between the logistic and probit approaches is primarily based on practical preferences and prior experience.

For the standard normal distribution N (0, 1), the CDF is commonly denoted by Φ (*z*) (continuous, monotone increasing sigmoid function) given by,

$$\Phi(z) = \mathrm{P}(Z \le z) = \int_{-\infty}^{\mathrm{z}} \varphi(u)\mathrm{du} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathrm{z}} e^{-\frac{\mathrm{u}^2}{2}} \mathrm{du}$$

As an example, considering the familiar fact that the N (0, 1) distribution places 95% of probability between -1.96 and 1.96, and is symmetric about zero, it follows that

$$\Phi(-1.96) = 0.025 = 1 - \Phi(1.96)$$

The probit function gives the 'inverse' computation, generating a value of an N (0, 1) random variable, associated with specified cumulative probability. Formally, the probit function is the inverse of Φ (*z*), denoted by Φ^{-1}(p). Continuing the example,

$$\Phi^{-1}(0.025) = -1.96 = -\Phi^{-1}(0.975)$$

In general,

$$\Phi(\text{probit(p)}) = \text{p} \ \text{ and } \ \text{probit}(\Phi(\text{z})) = \text{z}$$

In statistics, a probit model is a popular specification of a generalized linear model. If Y be a binary response variable, and let X be the single predictor variable, then the probit model assumes that,

$$P(Y_i = 1|X_i = x) = \Phi(\alpha + \beta x_i)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta x_i} e^{-\frac{1}{2}z^2} dz$$

where Φ is the CDF of the standard normal distribution. The parameters β are estimated by maximum likelihood.

In any dose-response scenario, there are two key components: the stimulus (such as a vitamin, drug, mental test, or physical force) and the subject (which could be an animal, plant, human volunteer, etc.). The stimulus is administered to the subject at a specific dose or intensity, measured in units such as concentration, weight, time, or other appropriate metrics, within a controlled environmental setting. Consequently, the subject exhibits a response. The response in this context is quantal, meaning it can either occur or not occur depending on the intensity of the stimulus. Under controlled conditions, a response is observed when the stimulus intensity surpasses a certain threshold or limen. However, the term "tolerance" is now more commonly used to refer to this value. The tolerance value varies among individuals within the population being studied. For quantal response data it is therefore necessary to consider distribution of tolerance over the population studied. If the dose or intensity of stimulus is measured by z, the distribution of tolerance may be expressed by $dP = f(z)dz$.

5. Classificatory ability of the models

There are several different classification accuracy measures that are commonly used to assess the performance of a classification model. Here are a few examples:

Accuracy: This is the most basic measure and represents the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances.

Precision: Precision focuses on the proportion of correctly classified positive instances (true positives) out of all instances predicted as positive (true positives plus false positives). It measures the model's ability to avoid false positives.

Recall (Sensitivity or True Positive Rate): Recall calculates the proportion of correctly classified positive instances (true positives) out of all actual positive instances (true positives plus false negatives). It quantifies the model's ability to capture true positives and avoid false negatives.

Specificity (True Negative Rate): Specificity evaluates the proportion of correctly classified negative instances (true negatives) out of all actual negative instances (true negatives plus false positives). It measures the model's ability to identify true negatives and avoid false positives.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that combines both precision and recall into a single value, useful when there is an imbalance between positive and negative instances.

Area Under the Receiver Operating Characteristic curve (AUC-ROC): The AUC-ROC measure quantifies the overall performance of a binary classifier by considering the trade-off between true positive rate (sensitivity) and false positive rate across different classification thresholds. It provides a single value that represents the model's ability to distinguish between positive and negative instances.

References:

Finney, D.J. (1971). *Probit Analysis* (3rd edition). Cambridge University Press, Cambridge, England.

Fox, J. (1984). *Linear statistical models and related methods with application to social research*, Wiley, New York.

Kleinbaum, D.G. (1994). *Logistic regression*: A self learning text, New York: Springer.

OVERVIEW OF TIME SERIES ANALYSIS

Mrinmoy Ray
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
mrinmoy.ray@icar.gov.in

Introduction:

A data set containing sequence of observations on a single phenomenon observed over time is called time-series data. In time series, past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship.

Time Series Components:

Trend: A trend exists when there is a long-term increase or decrease in the data. It does not have to be lin-ear. Some-times we will refer to a trend "changing direction" when it might go from an increas-ing trend to a decreasing trend.

Seasonal: A seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a fixed and known period.

Cyclic: A cyclic pattern exists when data exhibit rises and falls that are *not of fixed period*. The duration of these fluctuations is usually of at least 2 years.

Irregular component: Unobserved component exhibit in a time series

Exponential Smoothing Methods:

This method is suit-able for forecasting data with no trend or seasonal pattern. For exam-ple, the data in fig-ure do not dis-play any clear trend-ing behav-iour or any sea-son-al-ity, although the mean of the data may be chang-ing slowly over time. Simple moving average method assigns equal weights (1/k) to all k data points. Arguably, recent observations provide more information than do observations in the past. Exponential smoothing methods give larger weights to more recent observations, and the weights decrease exponentially as the observations become more distant. These methods are most effective when the parameters describing the time series are changing slowly over time

Types

- Simple exponential smoothing
- Holt's trend corrected exponential smoothing
- Holt-Winters method

Simple Exponential Smoothing (SES):

The SES method is used forecasting a time series when there is no trend or seasonal pattern, but the mean (or level) of the time series y_t is slowly changing over time

No trend model: $y_t = \beta_0 + \varepsilon_t$

Steps for SES method:

1. Compute the initial estimate of the mean (or level) of the series at time period $t = 0$

$$l_0 = \bar{y} = {\textstyle\sum_{t=1}^{n} y_t}/n$$

2. Compute the updated estimate by using the smoothing equation

$$\ell_T = \alpha y_T + (1-\alpha)\ell_{T-1}$$

where α is a smoothing constant between 0 and 1

Note that,

$$\ell_T = \alpha y_T + (1-\alpha)\ell_{T-1}$$

$$= \alpha y_T + (1-\alpha)[\alpha y_{T-1} + (1-\alpha)\ell_{T-2}]$$

$$= \alpha y_T + (1-\alpha)\alpha y_{T-1} + (1-\alpha)^2 \ell_{T-2}$$

$$= \alpha y_T + (1-\alpha)\alpha y_{T-1} + (1-\alpha)^2 \alpha y_{T-2} + \ldots + (1-\alpha)^{T-1}\alpha y_1 + (1-\alpha)^T \ell_0$$

<u>Holt's Trend Corrected Exponential Smoothing</u>

- A smoothing approach for forecasting such a time series that employs two smoothing constants, denoted by αandγ.
- There are two estimates $\ell_{T\text{-}1}$ and $b_{T\text{-}1}$
- $\ell_{T\text{-}1}$ is the estimate of the level of the time series constructed in time period T–1 (This is usually called the <u>permanent component</u>).
- $b_{T\text{-}1}$ is the estimate of the growth rate of the time series constructed in time period T–1 (This is usually called the <u>trend component</u>).
- Level estimate

$$\ell_T = \alpha y_T + (1-\alpha)(\ell_{T-1} + b_{T-1})$$

- Trend estimate

$$b_T = \gamma(\ell_T - \ell_{T-1}) + (1-\gamma)b_{T-1}$$

where α = smoothing constant for the level (0 ≤ α ≤ 1)

γ = smoothing constant for the trend (0 ≤ γ ≤ 1)

Holt-Winters Method

- Estimate of the level

$$\ell_T = \alpha(y_T / sn_{T-L}) + (1-\alpha)(\ell_{T-1} + b_{T-1})$$

- Estimate of the growth rate or trend

$$b_T = \gamma(\ell_T - \ell_{T-1}) + (1-\gamma)b_{T-1}$$

- Estimate of the seasonal factors

$$sn_T = \delta(y_T / \ell_T) + (1-\delta)sn_{T-L}$$

where α, γ, and δ are smoothing constants between 0 and 1, *L* = number of seasons in a year (*L* = 12 for monthly data, and *L* = 4 for quarterly data)

ARIMA Model:

Auto Regressive Integrated Moving Average (ARIMA) is a prediction model for time series analysis and forecasting

- **Here the terms indicate:**

Auto Regressive: lags of variables itself

Integrated: Differencing steps required to make time series stationary

Moving Average: lags of previous information shocks

- **ARIMA model is denoted as ARIMA(*p,d,q*)**

where

p=number of autoregressive terms

d=number of non-seasonal differences needed to make time series stationary

q=number of lagged forecast errors in the prediction equation

For ARIMA model building process there is a minimum of 30 data points required

In an autoregressive integrated moving average model, the future value of a variable is assumed to be a linear function of several past observations and random errors. The underlying process that generate the time series has the form

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}$$

where, and are the actual and random error at time period t, respectively; (i= 1, 2, …, p) and (*j= 1, 2, …, q*) are model parameters *p* and *q* are integers and often referred to as orders of the model

Random errors are assumed to be independently and identically distributed with a mean zero and a constant variance of σ^2

If *q*= 0, then the above equation becomes an AR model of order *p*. When *p*= 0, the model reduces to an MA model of order *q*.One central task of the ARIMA (*p, d, q*) model building is to determine the appropriate model order (*p, q*) where *d* is the order of differencing.

ANN approach to time series forecasting:

In the domain of time series analysis, the inputs are typically the past observations series and the output is the future value. The ANN performs the following nonlinear function mapping between the input and output

$$y_t = f(y_{t-1} + y_{t-2}, ..., y_{t-p}, w) + \varepsilon_t$$

where, w is a vector of all parameters and f is a function of network structure and connection weights. Therefore, the neural network resembles a nonlinear autoregressive model.

Single hidden layer multilayer feed forward network is the most popular for time series modeling and forecasting. This model is characterized by a network of three layers of simple processing units. The first layer is input layer, the middle layer is the hidden layer and the last layer is output layer.

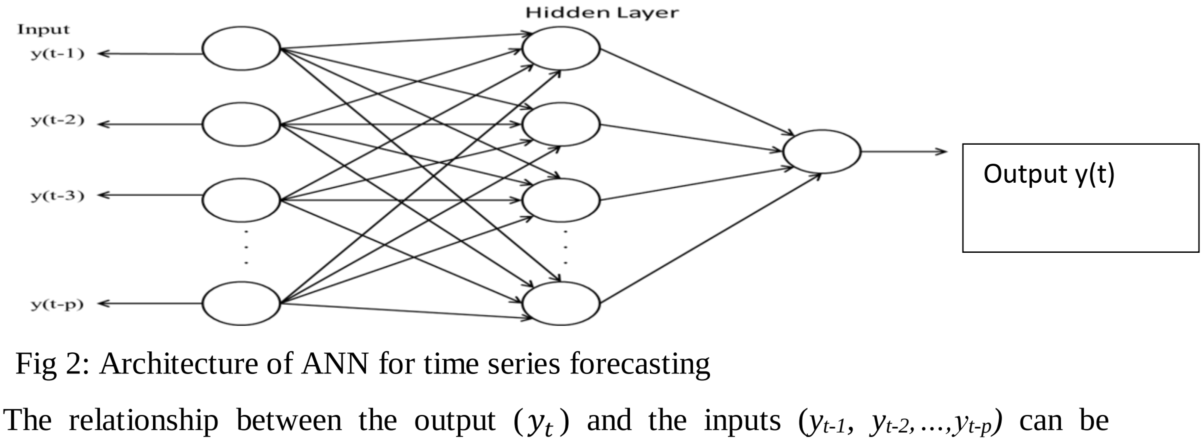


Fig 2: Architecture of ANN for time series forecasting

The relationship between the output (y_t) and the inputs (*y_{t-1}, y_{t-2},…,y_{t-p}*) can be mathematically represented as follows:

$$y_t = f\left(\sum_{j=0}^{q} \omega_j g\left(\sum_{i=0}^{p} \omega_{ij} y_{t-i}\right)\right)$$

where, $\omega_j (j = 0,1,2, \dots\dots, q)$ and $\omega_{ij} (i = 0,1,2, \dots\dots, p,\ j = 0,1,2, \dots\dots, q)$ are the model parameters often called the connection weights, p is the number of input nodes and q is the number of hidden nodes, g and f denote the activation function at hidden and output layer respectively. Activation function defines the relationship between inputs and outputs of a network in terms of degree of the non-linearity. Most commonly used activation functions are as follows-

Activation function	**Equation**
Identity	x
Sigmoid	$\frac{1}{1+e^{-x}}$
TanH	$\tanh(x) = \frac{2}{1+e^{-2x}} - 1$
ArcTan	$\tan^{-1}(x)$
Sinusoid	$\sin(x)$
Gaussian	e^{-x^2}

For time series forecasting sigmoid activation function is employed in hidden layer and identity activation function is employed in the output layer.

The selection of appropriate number of hidden nodes as well as optimum number of lagged observation p for input vector is important in ANN modeling for determination of the autocorrelation structure present in a time series. Though there are no established theories available for the selection of p and q, hence experiments are often conducted for the determination of the optimal values of p and q. The connection weights of ANNs are determined by learning method. There are three common learning algorithms for ANN –

1) Supervised Learning

The supervised learning strategy consists of having available the desired outputs for a given set of input signals; in other words, each training sample is composed of the input signals and their corresponding outputs. Henceforth, it requires a table with input/output data, also called attribute/value table, which represents the process and its behavior.

2) Unsupervised Learning

Different from supervised learning, the application of algorithm based on unsupervised learning does not require any knowledge of the respective desired outputs. Thus, the network needs to organize itself when there are existing particularities between the elements that compose the entire sample set, identifying subsets (or clusters) presenting similarities. The learning algorithm adjusts the synaptic weights and thresholds of the network in order to reflect these clusters within the network .itself.

3) Reinforcement Learning

It is the hybrid of supervised and unsupervised learning.

For time series forecasting supervised learning approach is utilized. Gradient decent back propagation algorithm is one of the popular approach of supervised learning.

Gradient decent back propagation algorithm

The objective of training is to minimize the error function that measures the misfit between the predicted value and the actual value. The error function which is widely used is mean squared error which can be written as:

$$E = \frac{1}{N}\sum_{n=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{n=1}^{N}\left\{ y_t - f\left(\sum_{j=0}^{q}\omega_j g\left(\sum_{i=0}^{p}\omega_{ij} y_{t-i}\right)\right)\right\}^2$$

Where N is the total number of error terms. The parameters of the neural network are ω_j and ω_{ij} estimated by iteration. Initial connection weights are taken randomly from uniform distribution. In each iteration the connection weights changed by an amount $\Delta\omega_j$

$$\Delta\omega_j(t) = -\eta\frac{\partial E}{\partial\omega_j} + \delta\Delta\omega_j(t-1)$$

where, η is the learning rate and $\frac{\partial E}{\partial\omega_j}$ is the partial derivative of the function E with respect to the weight ω_j. δ is the momentum rate. The $\frac{\partial E}{\partial\omega_j}$ can be represented as follows-

$$\frac{\partial E}{\partial w_j} = -e_j(n)\times f'(x)\times y_j(n)$$

where $e_j(n)$ is the residual at nth iteration

$f'(x)$= derivative of the activation function in the output layer. As in time series forecasting the activation function in the output layer is identity function hence $f'(x)$ =1. $y_j(n)$ is the desired output. Now connection weights in from input to hidden nodes changed by an amount $\Delta\omega_{ij}$

$$\Delta\omega_{ij}(t) = -\eta\frac{\partial E}{\partial\omega_{ij}} + \delta\Delta\omega_{ij}(t-1)$$

where

$$\frac{\partial E}{\partial w_{ij}} = g'(x)\times\sum_{j=0}^{q} e_j(n) * w_j(n)$$

where $g'(x)$ is the activation function in the hidden layer. For sigmoid activation function

$$g'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

Learning rate is user defined parameter known as tuning parameter of neural network which determine how slow or fast the optimal weight is obtained. The learning rate must be set small enough to avoid divergence. The momentum term prevents the learning process from setting in a local minimum. Though there are no established theories available for the selection of learning rate and momentum, hence experiments are often conducted for the determination of the learning rate and momentum.

Step by Step Modeling Procedure:

1. Testing of Nonlinearity:

As ANNs is suitable for nonlinear time series forecasting. Hence, prior to application of ANN the nonlinearity should be check. There are several tests for checking nonlinearity. BDS (Brock-Dechert-Scheinkman) test is of the popular approach for checking nonlinearity. This test utilizes the concept of spatial correlation from chaos theory. The computational procedure is given as follows

i) Let the considered time series is

$$\{x_i\} = [x_1, x_2, x_3, ..., x_N]$$

ii) The next step is to specify a value of m (embedding dimension), embed the time series into m dimensional vectors, by taking each m successive points in the series. This transforms the series of scalars into a series of vectors with overlapping entries

$$x_1^m = (x_1, x_2, ..., x_m)$$
$$x_2^m = (x_2, x_3, ..., x_{m+1})$$
$$.$$
$$.$$
$$.$$
$$x_{N-m}^m = (x_{N-m}, x_{N-m+1}, ..., x_N)$$

iii) In the third step correlation integral is computed, which measures the spatial correlation among the points, by adding the number of pairs of points (*i, j*), where $1 \leq i \leq N$ and $1 \leq j \leq N$, in the m-dimensional space which are "close" in the sense that the points are within a radius or tolerance ε of each other.

$$C_{\varepsilon,m} = \frac{1}{N_m(N_m - 1)} \sum_{i \neq j} I_{i,j;\varepsilon}$$

where $I_{i,j;\varepsilon}$= 1 if $\left\| x_i^m - x_j^m \right\| \leq \varepsilon$

= 0 otherwise

iv) If the time series is i.i.d. then $C_{\varepsilon,m} \approx [C_{\varepsilon,1}]^m$

v) The BDS test statistics is as follows

$$BDS_{\varepsilon,m} = \frac{\sqrt{N}[C_{\varepsilon,m} - (C_{\varepsilon,1})^m]}{\sqrt{V_{\varepsilon,m}}}$$

where $V_{\varepsilon,m} = 4[K^m + 2\sum_{j=1}^{m-1} K^{m-j} C_\varepsilon^{2j} + (m-1)^2 C_\varepsilon^{2m} - m^2 K C_\varepsilon^{2m-2}]$

$$K = K_\varepsilon = \frac{6}{N_m(N_m - 1)(N_m - 2)} \sum_{i<j<N} h_{i,j,N;\varepsilon}$$

$$h_{i,j,N;\varepsilon} = \frac{[I_{i,j;\varepsilon} I_{j,N;\varepsilon} + I_{i,N;\varepsilon} I_{N,j;\varepsilon} + I_{j,i;\varepsilon} I_{i,N;\varepsilon}]}{3}$$

The choice of m and ε depends on number of data. The null hypothesis is data are independently and identically distributed (i.i.d) against the alternative hypothesis the data are not i.i.d.; this implies that the time series is non-linearly dependent. BDS test is a two-tailed test; the null hypothesis should be rejected if the BDS test statistic is greater than or less than the critical values.

2. Division of the data:

Data is divided into training and test sets. The training sample is used for ANN for model development and the test sample is utilized to evaluate the forecasting performance. Sometimes a third one called the validation sample is also utilized to avoid the over fitting problem or to determine the stopping point of the training process. It is common to use one test set for both validation and testing purposes particularly for small data sets. The literature suggests little guidance in selecting the training and testing sets. Most commonly used rule are 90% vs. 10%, 80% vs. 20% or 70% vs. 30%, etc.

3. Data Normalization:

Nonlinear activation functions such as the sigmoid function typically have the squashing role in restricting the possible output from a node to, typically, (0, 1). Hence, data normalization is done prior to training process begins.

Normalization procedure

Linear transformation to [0,1]: $X_n=(X_0-X_{min})/ (X_{max}-X_{min})$

Statistical normalization: $X_n=(X_0\text{-mean}(X))/\text{var}(X)$

simple normalization: $X_n=X_0/ X_{max}$

4. Selection of appropriate number of hidden nodes as well as optimum number of lagged:

There are no established theories available for the selection of *p* and *q*, hence experiments are often conducted for the determination of the optimal values of *p* and *q*.

5. Estimation of connection weights:

Estimation of connection weights are determined by learning algorithm. For time series forecasting most commonly used learning approach is gradient decent back propagation algorithm.

6. Evaluating forecasting Performance

Forecasting performance can be computed by several approaches. Some of the approaches are given below-

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| / y_t \times 100$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}$$

where n is the total number of forecast values. y_t is the actual value at period t and $\hat{y}_t$ is the corresponding forecast value. The model with less MAPE/MSE/RMSE is preferred for forecasting purposes.

Limitations of ANN for time series forecasting:

i) ANNs are nonlinear time series model hence, for linear time series data the approach may not be better than linear statistical model.

ii) ANNs are black-box methods. There is no exact form to describe and analyze the relationship between inputs and outputs. This causes troublesome for interpretation of results. In addition, no formal statistical test is available.

iii) ANNs are subjected to have over fitting problems owing to its large number of parameters.

iv) There are no established theories available for the selection of p and q, hence experiments are often conducted for the determination of the optimal values of p and q which is tedious.

v) ANNs usually require more data for time series forecasting.

Support Vector Machine (SVM) in time series:

Application of SVM in time series is generally utilized when the series shows non stationarity and non-linearity process. A tremendous advantage of SVM is that it is not model dependent as well as independent of stationarity and linearity. However, it may be computationally expensive during the training. The training of the data driven prediction process SVM is done by a function which is estimated utilizing the observed data. Let, a time series $y(t)$ which takes the data at time $t\{t = 0,1,2,3,\ldots,N\}$.

Now, the prediction function for linear regression is defined as:

$$f(y) = (w.y) + c$$

Whereas, for non linear regression, it will be:

$$f(y) = \big(w.\emptyset(y)\big) + c$$

Where, w dentoes the weights, c represents threshold value and $\emptyset(y)$ is known as kernel function.If the observed data is linear, then equation (1) will be used. But, for non-lineadata,the mapping of $y(t)$ is done to the higher dimension feature space through some function which is denoted as $\emptyset(y)$ and eventually it is transformed into the linear process. Afer that, a linear regression will carry out in that feature space. The first and foremost objective is to find out the value of w and c which will be optimal. In SVM, there are two things viz., flatness of weights and error after the estimation which are to be minimized. The flatness of the weights is denoted by $||w||^2$ which is the eucledian norm. Firstly, one has to concentrate on minimization the $||w||^2$. Second important thing is the minimization of the error. This is also called as empirical risk. However, the overall aim is to minimize the regularized risk which is sum of empirical risk and the half of the product of the flatness of weight and a constant term which is known as regularized constant. The regularized risk can be written as-

$$R_{reg}(f) = R_{emp}(f) + \frac{\tau}{2}||w||^2$$

Where, $R_{reg}(f)$is the regularized risk, $R_{emp}(f)$ denotes the empirical risk, τ is as constant which is called as regularized constant/capacity control term and $||w||^2$ is the flatness of weights.

The regularization constant has a significant impact on a better fitting of the data and it can also be useful for the minimization of bad generalization effects. In the other words, this constant deals with the problem of over-fitting. The overfitting of the data can be redued by the proper selection of this constant value. The empirical risk can be defined as:-

$$R_{emp}(f) = \frac{1}{N}\sum_{i=0}^{N-1} L\big(y(i), \alpha(i), f(y(i), w)\big)$$

Where, $\alpha(i)$ denotes the truth data of predicted value, $L(.)$ is known as loss function and i represents the index to the time series.There are various types of loss function in literature. But, two functions viz., vapnik loss function and quadratic loss function are most popular and they are generally used. The quadratic programming problem has been made to minimize the regularised risk which is-

$$\text{Minimize, } \frac{1}{2}||w||^2 + D\sum_{i=1}^{n} L\big(\alpha(i), f(y(i), w)\big)$$

Where,

$$L\big(\alpha(i), f(y(i), w)\big) = |\alpha(i) - f(y(i), w)| - \in \text{ if } |\alpha(i) - f(y(i), w)| \geq \in$$

$$= 0; \text{ otherwise.}$$

Where, D is a constant which equals to the summation normalization factor and $\in$ represents the size of the tube.

The computation of $\in$ and D is done empirically because they are user defined. On has to choose proper value of D and $\in$. Now, dual optimization problem is formed using the lagrange multiplier which can be written as:

Maximize, $-\frac{1}{2}\sum_{i,j=1}^{N}(\beta_i - \beta_i{}^{*})\left(\beta_j - \beta_j{}^{*}\right)\langle y(i), y(j)\rangle - \in \sum_{i=1}^{N}(\beta_i - \beta_i{}^{*}) + \sum_{i=1}^{N} \alpha(i)(\beta_i - \beta_i{}^{*})$

Subject to, $\sum_{i-1}^{N}(\beta_i - \beta_i{}^{*}) = 0$; $\beta_i, \beta_i{}^{*} \in [0, D]$

The function $f(x)$ is defined as;

$$f(x) = \sum_{i=1}^{N}(\beta_i - \beta_i{}^{*})\,\langle y, y(i)\rangle + C$$

KKT conditions are used to get the solution of the weights.

The significance of kernel function in non-linear support vector machine (NLSVR) is very much imporatnt for mapping the data $y(i)$into higher dimension feature space $\emptyset\big(\,y(i)\big)$in which the data becomes linear. Generally notation for kernel function is given as;

$$k(y, y') = \langle \emptyset(y), \emptyset(y')\rangle;$$

There are many methods in literature to solve the quadartic programming. However, the most used method is sequential minimization optimization (SMO) algorithm.

References:

Anjoy, P., Paul, R. K., Sinha, K., Paul, A. K. and Ray, M. (2017). A hybrid wavelet based neural networks model for predicting monthly WPI of pulses in India. *Indian Journal of Agricultural Sciences*. **87 (6)**, 834-839.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2009), *Time Series Analysis: Forecasting and Control (3rd ed.), San Francisco: Holden-Day.*

Broock, W., Scheinkman, J. A., Dechert, W. D. and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Review*, **15,** 197–235.

Jha, G. K., and Sinha, K.2014.Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India. *Neural Computing and Applications*24 (3): 563-571.

Makridakis, S., Wheelwright, S.C. and Hyndman, R. J. (1998).*Forecasting: Methods and Applications (3rd ed.),* Chichester: Wiley.

Mukherjee, A., Rakshit, S., Nag, A., Ray, M**.,**Kharbikar, H. L., Kumari, S., Sarkar, S., Paul, S., Roy, S., Maity, A., Meena, V. S. and Burman, R. R. (2016). Climate Change Risk Perception, Adaptation and Mitigation Strategy: An Extension Outlook in Mountain Himalaya. In: Jaideep Kumar Bisht, Vijay Singh Meena, Pankaj Kumar Mishra and Arunava Pattanayak Edition. Conservation Agriculture (pp. 257-292). Singapore. Springer Singapore.

Ray, M., Rai, A., Ramasubramanian, V. and Singh, K. N. (2016). ARIMA-WNN hybrid model for forecasting wheat yield time series data. *Journal of the Indian Society of Agricultural Statistics,* **70(1)**, 63-70.

Ray, M., Rai, A., Singh, K. N., Ramasubramanian, V. and Kumar, A. (2017). Technology forecasting using time series intervention based trend impact analysis for wheat yield scenario in India. *Technological Forecasting & Social Change*, **118**, 128–133.

Remus, W. and O'Connor, M.(2001). *Neural Networks for Time-Series Forecasting, New york, Springer.*

Zhang, G., Patuwo, B. E. and Hu, M. Y. 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14: 35-62.

Planning of Experiments and Basic Experimental Designs

Seema Jaggi and Anindita Datta
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
seema.jaggi@icar.gov.in, anindita.datta@icar.gov.in

An experiment is usually associated with a scientific method for testing certain phenomena. An experiment facilitates the study of such phenomena under controlled conditions and thus creating controlled condition is an essential component. Scientists in the biological fields who are involved in research constantly face problems associated with planning, designing and conducting experiments. Basic familiarity and understanding of statistical methods that deal with issues of concern would be helpful in many ways. Researchers who collect data and then look for a statistical technique that would provide valid results will find that there may not be solutions to the problem and that the problem could have been avoided first by a properly designed experiment. Obviously it is important to keep in mind that we cannot draw valid conclusions from poorly planned experiments. Second, the time and cost involved in many experiments are enormous and a poorly designed experiment increases such costs in time and resources. For example, an agronomist who carries out fertilizer experiment knows the time limitation of the experiment. He knows that when seeds are to be planted and harvested. The experimenter plot must include all components of a complete design. Otherwise what is omitted from the experiment will have to be carried out in subsequent trials in the next cropping season or next year. The additional time and expenditure could be minimized by a properly planned experiment that will produce valid results as efficiently as possible. Good experimental designs are products of the technical knowledge of one's field, an understanding of statistical techniques and skill in designing experiments.

Any research endeavor may entail the phases of Conception, Design, Data collection, Analysis and Dissemination. Statistical methodologies can be used to conduct better scientific experiments if they are incorporated into entire scientific process, i.e., From inception of the problem to experimental design, data analysis and interpretation. When planning experiments we must keep in mind that large uncontrolled variations are common occurrences. Experiments are generally undertaken by researchers to compare effects of several conditions on some phenomena or in discovering an unknown effect of particular process. An experiment facilitates the study of such phenomena under controlled conditions. Therefore the creation of controlled condition is the most essential characteristic of experimentation. How we formulate our questions and hypotheses are critical to the experimental procedure that will follow. For example, a crop scientist who plants the same variety of a crop in a field may find variations in yield that are due to periodic variations across a field or to some other factors that the experimenter has no control over. The methodologies used in designing experiments will separate with confidence and accuracy a varietal difference of crops from the uncontrolled variations.

The different concepts in planning of experiment can be well explained through chapati tasting experiment.

Consider an experiment to detect the taste difference in chapati made of wheat flour of c306 and pv 18 varieties. The null hypothesis we can assume here is that there is no taste difference in chapatis made of c306 or pv18 wheat flours. After the null hypothesis is set, we have to fix the level of significance at which we can operate. The pv18 is a much higher yielding variety than c306. Hence a false rejection may not help the country to grow more pv18 and the wheat production may decrease while a false acceptance may give more production of pv18 wheat and the consumption may be less or practically nil. Thus the false acceptance or false rejection are of practically equal consequence and we agree to choose the level of significance at $\alpha = 0.05$. Now to execute the experiment, a subject is to be found with extrasensory powers who can detect the taste differences. The colours of c306 and pv18 are different and anyone, even without tasting the chapatis, can distinguish the chapatis of either kind by a mere glance. Thus the taster of the chapatis has to be blindfolded before the chapatis are given for tasting. Afterwards, the method is to be decided in which the experiment will be conducted. The experiment can be conducted in many ways and of them three methods are discussed here:

- Give the taster equal number of chapatis of either kind informing the taster about it.
- Give the taster pairs of chapatis of each kind informing the taster about it.
- Give the taster chapatis of either kind without providing him with any information. Let us use 6 chapatis in each of these methods.

Under first method of experimentation, if the null hypothesis is true, then the experimenter cannot distinguish the two kinds of chapaties and he will randomly select 3 chapatiS out of 6 chapaties given to him, as made of pvl8 wheat. In that case, all correct guesses are made if selection exactly coincides with the exactly used wheat variety and the probability for such an occurrence is:

$$1 \Big/ \binom{6}{3} = 1 \Big/ 20 = 0.05$$

Under second method,the pv18 wheat variety chapaties are selected from each pair given if the null hypothesis is true. Furthermore, independent choices are made of pv18 variety chapaties from each pair. Thus the probability of making all correct guesses is

$$1/(2)^3 = 1/8 = 0.125.$$

In third method the experimenter has to make the choice for each chapati and the situation is analogous at calling heads or tails in a coin tossing experiment. The probability of making all correct guesses would then be:

$$1/2^6 = 1/64 = .016.$$

If the experimenter makes all correct guesses in third method as its probability is smaller than the selected $\alpha = 0.05$, we can reject the null hypothesis and conclude that the two wheat varieties give different tastes at chapaties. In other methods the probability of making all correct guesses does not exceed $\alpha = 0.05$ and hence with either method, we cannot reject the null hypothesis even if all correct guesses are made.

However, if 8 chapaties are used by first method and if the taster guesses all of them, we can reject the null hypothesis, at 0.05 level of significance, as the probability of making all correct guesses would then be $1 \Big/ \binom{8}{3} = 1 \Big/ 56$ which is smaller than 0.05. 8 chapaties will not enable us to reject the null hypothesis even if all correct guesses are made by second method as the probability of making all correct guesses is $\left(\frac{1}{4}\right)^4 = \frac{1}{16} = 0.06$ it is easy to see that if 10 chapaties are given by second method and if all correct guesses are made, then we can reject the null hypothesis at 0.05 level of significance. Not to unduly influence the taster in making guesses, we should also present the chapaties in a random order rather than systematically presenting them for tasting.

The above discussed chapati tasting experiment brings home the following salient features of experimentation:

- All the extraneous variations in the data should be eliminated or controlled excepting the variations due to the treatments under study. One should not artificially provide circumstances for one treatment to show better results than others.
- Far a given size of the experiment, though the experiment can be done in many ways, even the best results may not turn out to be significant with some designs, while some other design can detect the treatment differences. Thus there is an imperative need the choose the right type of design, before the commencement of the experiment, lest the results may be useless.
- If for some specific reasons related to the nature .of the experiment, a particular method has to be used in experimentation, then adequate number of replications of each treatment have to be provided in order to get valid inferences.
- The treatments have to be randomly allocated to the experimental units.

The terminologies often used in planning and designing of experiments are listed below.

Treatment

Treatment refers to controllable quantitative or qualitative factors imposed at a certain level by the experimenter. For an agronomist several fertilizer concentrations applied to a particular crop or a variety of crop is a treatment. Similarly, an animal scientist looks upon several concentrations of a drug given to animal species as a treatment. In agribusiness we may look upon impact of advertising strategy on sales a treatment. To an agricultural engineer, different levels of irrigation may constitute a treatment.

Experimental Unit

An experimental unit is an entity that receives a treatment e.g., for an agronomist or horticulturist it may be a plot of a land or batch of seed, for an animal scientist it may be a group of pigs or sheep, for a scientist engaged in forestry research it may be different tree species occurring in an area, and for an agricultural engineer it may be manufactured item. Thus, an experimental unit maybe looked upon as a small subdivision of the experimental material, which receives the treatment.

Experimental Error

Differences in yields arising out of experimental units treated alike are called Experimental Error.

Controllable conditions in an experiment or experimental variable are terms as a factor. For example, a fertilizer, a new feed ration, and a fungicide are all considered as factors. Factors may be qualitative or quantitative and may take a finite number of values or type. Quantitative factors are those described by numerical values on some scale. The rates of application of fertilizer, the quantity of seed sown are examples of quantitative factors. Qualitative factors are those factors that can be distinguished from each other, but not on numerical scale e.g., type of protein in a diet, sex of an animal, genetic make up of plant etc. While choosing factors for any experiment researcher should ask the following questions, like What treatments in the experiment should be related directly to the objectives of the study? Does the experimental technique adopted require the use of additional factors? Can the experimental unit be divided naturally into groups such that the main treatment effects are different for the different groups? What additional factors should one include in the experiment to interact with the main factors and shed light on the factors of direct interest? How desirable is it to deliberately choose experimental units of different types?

Basic Principles of Design of Experiments

Given a set of treatments which can provide information regarding the objective of an experiment, a design for the experiment, defines the size and number of experimental units, the manner in which the treatments are allotted to the units and also appropriate type and grouping of the experimental units. These requirements of a design ensure validity, interpretability and accuracy of the results obtainable from an analysis of the observations.

These purposes are served by the principles of:

- Randomization
- Replication
- Local (Error) control

Randomization

After the treatments and the experimental units are decided the treatments are allotted to the experimental units at random to avoid any type of personal or subjective bias, which may be conscious or unconscious. This ensures validity of the results. It helps to have an objective comparison among the treatments. It also ensures independence of the observations, which is necessary for drawing valid inference from the observations by applying appropriate statistical techniques.

Depending on the nature of the experiment and the experimental units, there are various experimental designs and each design has its own way of randomization. Various speakers while discussing specific designs in the lectures to follow shall discuss the procedure of random allocation separately.

Replication

If a treatment is allotted to r experimental units in an experiment, it is said to be replicated r times. If in a design each of the treatments is replicated r times, the design is said to have r replications. Replication is necessary to

- Provide an estimate of the error variance which is a function of the differences among observations from experimental units under identical treatments.
- Increase the accuracy of estimates of the treatment effects.

Though, more the number of replications the better it is, so far as precision of estimates is concerned, it cannot be increased infinitely as it increases the cost of experimentation. Moreover, due to limited availability of experimental resources too many replications cannot be taken.

The number of replications is, therefore, decided keeping in view the permissible expenditure and the required degree of precision. Sensitivity of statistical methods for drawing inference also depends on the number of replications. Sometimes this criterion is used to decide the number of replications in specific experiments.

Error variance provides a measure of precision of an experiment, the less the error variance the more precision. Once a measure of error variance is available for a set of experimental units, the number of replications needed for a desired level of sensitivity can be obtained as below.

Given a set of treatments an experimenter may not be interested to know if two treatment differ in their effects by less than a certain quantity, say, d. In other words, he wants an experiment that should be able to differentiate two treatments when they differ by d or more.

The significance of the difference between two treatments is tested by t-test where

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{2s^2 / r}},$$

Here, $\bar{y}_i$, and $\bar{y}_j$ are the arithmetic means of two treatment effects each based on r replications, s^2 is measure of error variation.

Given a difference d, between two treatment effects such that any difference greater than d should be brought out as significant by using a design with r replications, the following equation provides a solution of r.

$$t = \frac{|d\ |}{\sqrt{2s^2/r}},$$

$$r = \frac{t_0^2}{d^2} x 2s^2$$

where t_0 is the critical value of the t-distribution at the desired level of significance, that is, the value of t at 5 or 1 per cent level of significance read from the t-table. If s^2 is known or based on a very large number of observations, made available from some pilot pre-experiment investigation, then t is taken as the normal variate. If s^2 is estimated with n degree of freedom (d.f.) then t_0 corresponds to n d.f.

When the number of replication is r or more as obtained above, then all differences greater than d are expected to be brought out as significant by an experiment when it is conducted on a set of experimental units which has variability of the order of s^2. For example, in an experiment on wheat crop conducted in a seed farm in Bhopal, to study the effect of application of nitrogen and phosphorous on yield a randomized block design with three replications was adopted. There were 11 treatments two of which were (i) 60 Kg/ha of nitrogen (ii) 120 Kg/ha of nitrogen. The average yield figures for these two application of the fertilizer were 1438 and 1592 Kg/ha respectively and it is required that differences of the order of 150 Kg/ha should be brought out significant. The error mean square (s^2) was 12134.88. Assuming that the experimental error will be of the same order in future experiments and t_0 is of the order of 2.00, which is likely as the error d.f. is likely to be more than 30 as there are 11 treatments; Substituting in (1), we get:

$$r = \frac{2t_0^2 s^2}{d^2} = \frac{2x2^2 x2134.88}{150^2} = 4\,(\text{approx.})$$

Thus, an experiment with 4 replications is likely to bring out differences of the order of 150 Kg/ha as significant.

Another criterion for determining r is to take a number of replications which ensures at least 10 d.f. for the estimate of error variance in the analysis of variance of the design concerned since the sensitivity of the experiment will be very much low as the F test (which is used to draw inference in such experiments) is very much unstable below 10 d.f.

Local Control

The consideration in regard to the choice of number of replications ensure reduction of standard error of the estimates of the treatment effect because the standard error of the estimate of a treatment effect is $\sqrt{s^2 / r}$, but it cannot reduce the error variance itself. It is, however, possible to devise methods for reducing the error variance. Such measures are called *error control* or local control. One such measure is to make the experimental units homogenous. Another method is to form the units into several homogenous groups, usually called blocks, allowing variation between the groups.

A considerable amount of research work has been done to divide the treatments into suitable groups of experimental units so that the treatment effect can be estimated more precisely Extensive use of combinatorial mathematics has been made for formation of such group treatments. This grouping of experiment units into different groups has led to the development of various designs useful to the experimenter. We now briefly describe the various term used in designing of an experiment

Blocking

It refers to methodologies that form the units into homogeneous or pre-experimental subject-similarity groups. It is a method to reduce the effect of variation in the experimental material on the Error of Treatment of Comparisons. For example, animal scientist may decide to group animals on age, sex, breed or some other factors that he may believe has an influence on characteristic being measured. Effective blocking removes considerable measure of variation nom the experimental error. The selection of source of variability to be used as basis of blocking, block size, block shape and orientation are crucial for blocking. The blocking factor is introduced in the experiment to increase the power of design to detect treatment effects.

The importance of good designing is inseparable from good research (results). The following examples point out the necessity for a good design that will yield good research. First, a nutrition specialist in developing country is interested in determining whether mother's milk is better than powdered milk for children under age one. The nutritionist has compared the growth of children in village A, who are all on mother's milk against the children in village B, who use powdered milk. Obviously, such a comparison ignores the health of the mothers, the sanitary-conditions of the villages, and other factors that may have contributed to the differences observed without any connection to the advantages of mother's milk or the powdered milk on the children.

A proper design would require that both mother's milk and the powdered milk be alternatively used in both villages, or some other methodology to make certain that the differences observed are attributable to the type of milk consumed and not to some uncontrollable factor. Second, a crop scientist who is comparing 2 varieties of maize, for instance, would not assign one variety to a location where such factors as sun, shade, unidirectional fertility gradient, and uneven distribution of water would either favor or handicap it over the other. If such a design were to be adopted, the researcher would have difficulty in determining whether the apparent difference in yield was due to variety differences or resulted from such factors as sun, shade, soil fertility of the field, or the distribution of water. These two examples illustrate the type of poorly designed experiments that are to be avoided.

Analysis of Variance

Analysis of Variance (ANOVA) is a technique of partitioning the overall variation in the responses into different assignable sources of variation, some of which are specifiable and others unknown. Total variance in the sample data is partitioned and is expressed as the sum of its non-negative components is a measure of the variation due to some specific independent source or factor or cause. ANOVA consists in estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to ascribable factors (causes) with the estimate due to chance factor the latter being known as experimental error or simply the error.

Total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components associated with the nature of classification of the data. The systematic procedure for achieving this is called *Analysis of Variance*. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance.

Thus, ANOVA is a statistical technique that can be used to evaluate whether there are differences between the average value, or mean, across several population groups. With this model, the *response variable is continuous* in nature, whereas the *predictor variables are categorical*. For example, in a clinical trial of hypertensive patients, ANOVA methods could be used to compare the effectiveness of three different drugs in lowering blood pressure. Alternatively, ANOVA could be used to determine whether infant birth weight is significantly different among mothers who smoked during pregnancy relative to those who did not. In a particular case, where two population means are being compared, ANOVA is equivalent to the independent two-sample *t*-test.

The fixed-effects model of ANOVA applies to situations in which the experimenter applies several treatments to the subjects of the experiment to see if the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole. In it factors are fixed and are attributable to a finite set of levels of factor eg. Sex, year, variety, fertilizer etc.

Consider for example a clinical trial where three drugs are administered on a group of men and women some of whom are married and some are unmarried. The three classifications of sex, drug and marital status that identify the source of each datum are known as factors. The individual classification of each factor is known as levels of the factors. Thus, in this example there are 3 levels of factor drug, 2 levels of factor sex and 2 levels of marital status. Here all the effects are fixed. Random effects models are used when the treatments are not fixed. This occurs when the various treatments (also known as factor levels) are sampled from a larger population. When factors are random, these are generally attributable to infinite set of levels of a factor of which a random sample are deemed to occur *eg.* research stations, clinics in Delhi, sire, etc. Suppose new inject-able insulin is to be tested using 15 different clinics of Delhi state. It is reasonable to assume that these clinics are random sample from a population of clinics from Delhi. It describe the situations where both fixed and random effects are present.

In any ANOVA model, general mean is always taken as fixed effect and error is always taken as random effect. Thus class of model can be classified on the basis of factors, other than these two factors. ANOVA can be viewed as a generalization of *t*-tests: a comparison of differences of means across more than two groups.

The ANOVA is valid under certain assumptions. These assumptions are:

- Samples have been drawn from the populations that are normally distributed.
- Observations are independent and are distributed normally with mean zero and variance σ^2.
- Effects are additive in nature.

The ANOVA is performed as one-way, two-way, three-way, etc. ANOVA when the number of factors is one, two or three respectively. In general if the number of factors is more, it is termed as multi-way ANOVA.

Completely Randomized Design

Designs are usually characterized by the nature of grouping of experimental units and the procedure of random allocation of treatments to the experimental units. In a completely randomized design the units are taken in a single group. As far as possible the units forming the group are homogeneous. This is a design in which only randomization and replication are used. There is no use of local control here.

Let there be *v* treatments in an experiment and *n* homogeneous experimental units. Let the i^{th} treatment be replicated r_i times *(i = 1,2,..., v)* such that $\sum_{i=1}^{v} r_i = n$. The treatments are allotted at random to the units.

Normally the number of replications for different treatments should be equal as it ensures equal precision of estimates of the treatment effects. The actual number of replications is, however, determined by the availability of experimental resources and the requirement of precision and sensitivity of comparisons. If the experimental material for some treatments is available in limited quantities, the numbers of their replication are reduced. If the estimates of certain treatment effects are required with more precision, the numbers of their replication are increased.

Randomization

There are several methods of random allocation of treatments to the experimental units. The *v* treatments are first numbered in any order from *1* to *v*. The *n* experimental units are also numbered suitably. One of the methods uses the random number tables. Any page of a random number table is taken. If *v* is a one-digit number, then the table is consulted digit by digit. If *v* is a two-digit number, then two-digit random numbers are consulted. All numbers greater than *v* including zero are ignored.

Let the first number chosen be n_1; then the treatment numbered $n_1$is allotted to the first unit. If the second number is n_2 which may or may not be equal to n_1 then the treatment numbered n_2 is allotted to the second unit. This procedure is continued. When the i^{th} treatment number has occurred r_i times, $(i = 1,2,...,v)$ this treatment is ignored subsequently. This process terminates when all the units are exhausted.

One drawback of the above procedure is that sometimes a very large number of random numbers may have to be ignored because they are greater than *v*. It may even happen that the random number table is exhausted before the allocation is complete. To avoid this difficulty the following procedure is adopted. We have described the procedure by taking *v* to be a two-digit number.

Let *P* be the highest two-digit number divisible by *v*. Then all numbers greater than *P* and zero are ignored. If a selected random number is less than *v*, then it is used as such. If it is greater than or equal to *v*, then it is divided by *v* and the remainder is taken to the random number. When a number is completely divisible by *v*, then the random number is *v*. If *v* is an *n*-digit number, then *P* is taken to be the highest *n*-digit number divisible by *v*. The rest of the procedure is the same as above.

Analysis

This design provides a one-way classified data according to levels of a single factor. For its analysis the following model is taken:

$$y_{ij} = \mu + t_i + e_{ij} \ , \qquad i = 1,\cdots,v; j = 1,\cdots r_i ,$$

where y_{ij} is the random variable corresponding to the observation y_{ij} obtained from the j^{th} replicate of the i^{th} treatment, μ is the general mean, t_i is the fixed effect of the i^{th} treatment and e_{ij} is the error component which is a random variable assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Let $\sum y_{ij} = T_i \quad (i = 1,2,...,v)$ be the total of observations from i^{th} treatment. Let further $\sum_i T_i = G$. Correction factor (*C.F.*) = G^2/n.

Sum of squares due to treatments $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$

Total sum of squares = $\sum_{i=1}^{v} \sum_{j=1}^{r_i} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Treatments	$v - 1$	*SST* $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$	*MST = SST / (v - 1)*	*MST/MSE*
Error	$n - v$	*SSE = by subtraction*	*MSE = SSE / (n - v)*	
Total	$n - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test where F is the ratio *MST / MSE* with *(v - 1)* and *(n - v)* degrees of freedom.

3. Randomized Complete Block Design

It has been seen that when the experimental units are homogeneous then a CRD should be adopted. In any experiment, however, besides treatments the experimental material is a major source of variability in the data. When experiments require a large number of experimental units, the experimental units may not be homogeneous, and in such situations CRD can not be recommended. When the experimental units are heterogeneous, a part of the variability can be accounted for by grouping the experimental units in such a way that experimental units within each group are as homogeneous as possible. The treatments are then allotted randomly to the experimental units within each group (or blocks). The principle of first forming homogeneous groups of the experimental units and then allotting at random each treatment once in each group is known as local control. This results in an increase in precision of estimates of the treatment contrasts, due to the fact that error variance that is a function of comparisons within blocks, is smaller because of homogeneous blocks. This type of allocation makes it possible to eliminate from error variance a portion of variation attributable to block differences. If, however, variation between the blocks is not significantly large, this type of grouping of the units does not lead to any advantage; rather some degrees of freedom of the error variance is lost without any consequent decrease in the error variance. In such situations it is not desirable to adopt randomized complete block designs in preference to completely randomized designs.

If the number of experimental units within each group is same as the number of treatments and if every treatment appears precisely once in each group then such an arrangement is called a ***randomized complete block design.***

Suppose the experimenter wants to study *v* treatments. Each of the treatments is replicated *r* times (the number of blocks) in the design. The total number of experimental units is, therefore, *vr*. These units are arranged into *r* groups of size *v* each. The error control measure in this design consists of making the units in each of these groups homogeneous.

The number of blocks in the design is the same as the number of replications. The *v* treatments are allotted at random to the *v* plots in each block. This type of homogeneous grouping of the experimental units and the random allocation of the treatments separately in each block are the two main characteristic features of randomized block designs. The availability of resources and considerations of cost and precision determine actual number of replications in the design.

Analysis

The data collected from experiments with randomized block designs form a two-way classification, that is, classified according to the levels of two factors, *viz.,* blocks and treatments. There are *vr* cells in the two-way table with one observation in each cell. The data are orthogonal and therefore the design is called an *orthogonal design.* We take the following model:

$$y_{ij} = \mu + t_i + b_j + e_{ij} \ , \qquad \begin{pmatrix} i = 1,2,...,v; \\ j = 1,2,...,r \end{pmatrix},$$

where y_{ij} denotes the observation from *i*th treatment in *j*th block. The fixed effects μ, t_i, b_j denote respectively the general mean, effect of the *i*th treatment and effect of the *j*th block. The random variable e_{ij} is the error component associated with y_{ij}. These are assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Following the method of analysis of variance for finding sums of squares due to blocks, treatments and error for the two-way classification, the different sums of squares are obtained as follows: Let $\sum_j y_{ij} = T_i \ \ (i = 1,2,...,v)$ = total of observations from *i*th treatment and $\sum_j y_{ij} = B_j \quad j = 1,\cdots,r$ = total of observations from *j*th block. These are the marginal totals of the two-way data table. Let further, $\sum_i T_i = \sum_j B_j = G.$

Correction factor (*C.F.*) = G^2/rv, Sum of squares due to treatments $= \sum_i \frac{T_i^2}{r} - C.F.$,

Sum of squares due to blocks $= \sum_j \frac{B_j^2}{v} - C.F.$, Total sum of squares $= \sum_{ij} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	**Degrees of freedom (D.F.)**	**Sum of squares (S.S.)**	**Mean squares (M.S.)**	**F**
Blocks	$r - 1$	$SSB = \sum_j \frac{B_j^2}{v} - C.F.$	$MSB = SSB / (r - 1)$	MSB/MSE
Treatments	$v - 1$	$SST = \sum_i \frac{T_i^2}{r} - C.F.$	$MST = SST / (v - 1)$	MST/MSE
Error	$(r - 1)(v - 1)$	$SSE =$ *by subtraction*	$MSE = SSE / (v - 1)(r - 1)$	
Total	$vr - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test, where F is the ratio *MST / MSE* with *(v - 1)* and *(v - 1)(r - 1)* degrees of freedom. We may then be interested to either compare the treatments in pairs or evaluate special contrasts depending upon the objectives of the experiment. This is done as follows:

The critical difference for testing the significance of the difference of two treatment effects, say $t_i - t_j$ is $C.D. = t_{(v-1)(r-1),\alpha/2}\sqrt{2MSE / r}$, where $t_{(v-1)(r-1),\alpha/2}$ is the value of Student's *t* at the level of significance α and degree of freedom *(v - 1)(r - 1)*. If the difference of any two-treatment means is greater than the C.D. value, the corresponding treatment effects are significantly different.

4. Latin Square Design

Latin square designs are normally used in experiments where it is required to remove the heterogeneity of experimental material in two directions. These designs require that the number of replications equal the number of ***treatments*** or ***varieties***.

Definition 1. A Latin square arrangement is an arrangement of *v* symbols in v^2 cells arranged in *v* rows and *v* columns, such that every symbol occurs precisely once in each row and precisely once in each column. The term *v* is known as the **order** of the Latin square.

If the symbols are taken as *A, B, C, D,* a Latin square arrangement of order 4 is as follows:

$$\begin{array}{cccc} A & B & C & D \\ B & C & D & A \\ C & D & A & B \\ D & A & B & C \end{array}$$

A Latin square is said to be in the ***standard form*** if the symbols in the first row and first column are in natural order, and it is said to be in the ***semi-standard form*** if the symbols of the first row are in natural order. Some authors denote both of these concepts by the term ***standard form***. However, there is a need to distinguish between these two concepts. The standard form is used for randomizing the Latin-square designs, and the semi-standard form is needed for studying the properties of the orthogonal Latin squares.

Definition 2. If in two Latin squares of the same order, when superimposed on one another, every ordered pair of symbols occurs exactly once, the two Latin squares are said to be **orthogonal**. If the symbols of one Latin square are denoted by Latin letters and the symbols of the other are denoted by Greek letters, the pair of orthogonal Latin squares is also called a **graeco-latin square**.

Definition 3. If in a set of Latin squares every pair is orthogonal, the set is called a set of **mutually orthogonal latin squares (MOLS)**. It is also called a **hypergraeco latin square.**

The following is an example of graeco latin square:

$$\begin{array}{cccc} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{array} \qquad \begin{array}{cccc} \alpha & \gamma & \delta & \beta \\ \beta & \delta & \gamma & \alpha \\ \gamma & \alpha & \beta & \delta \\ \delta & \beta & \alpha & \gamma \end{array}$$

$$\begin{array}{cccc} A\alpha & B\gamma & C\delta & D\beta \\ B\beta & A\delta & D\gamma & C\alpha \\ C\gamma & D\alpha & A\beta & B\delta \\ D\delta & C\beta & B\alpha & A\gamma \end{array}$$

We can verify that in the above arrangement every pair of ordered Latin and Greek symbols occurs exactly once, and hence the two latin squares under consideration constitute a graecolatin square.

It is well known that the maximum number of MOLS possible of order v is $v - 1$. A set of $v - 1$ MOLS is known as a complete set of MOLS. Complete sets of MOLS of order v exist when v is a **prime or prime power.**

Randomization

According to the definition of a Latin square design, treatments can be allocated to the v^2 experimental units (may be animal or plots) in a number of ways. There are, therefore, a number of Latin squares of a given order. The purpose of randomization is to select one of these squares at random. The following is one of the methods of random selection of Latin squares.

Let a $v \times v$ Latin square arrangement be first written by denoting treatments by Latin letters *A, B, C, etc.* or by numbers *1, 2, 3, etc.* Such arrangements are readily available in the ***Tables for Statisticians and Biometricians*** (Fisher and Yates, 1974). One of these squares of any order can be written systematically as shown below for a 5×5 Latin square:

$$\begin{array}{ccccc}
A & B & C & D & E \\
B & C & D & E & A \\
C & D & E & A & B \\
D & E & A & B & C \\
E & A & B & C & D
\end{array}$$

For the purpose of randomization rows and columns of the Latin square are rearranged randomly. There is no randomization possible within the rows and/or columns. For example, the following is a row randomized square of the above 5×5 Latin square;

$$\begin{array}{ccccc}
A & B & C & D & E \\
B & C & D & E & A \\
E & A & B & C & D \\
D & E & A & B & C \\
C & D & E & A & B
\end{array}$$

Next, the columns of the above row randomized square have been rearranged randomly to give the following random square:

$$\begin{array}{ccccc}
E & B & C & A & D \\
A & C & D & B & E \\
D & A & B & E & C \\
C & E & A & D & B \\
B & D & E & C & A
\end{array}$$

As a result of row and column randomization, but not the randomization of the individual units, the whole arrangement remains a Latin square.

Analysis of Latin Square Designs

In Latin square designs there are three factors. These are the factors *P, Q,* and treatments. The data collected from this design are, therefore, analyzed as a three-way classified data. Actually, there should have been v^3 observations as there are three factors each at v levels. But because of the particular allocation of treatments to the cells, there is only one observation per cell instead of *v* in the usual three way classified orthogonal data. As a result we can obtain only the sums of squares due to each of the three factors and error sum of squares. None of the interaction sums of squares of the factors can be obtained. Accordingly, we take the model

$$Y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$$

where y_{ijs} denotes the observation in the i^{th} row, j^{th} column and under the s^{th} treatment; $\mu, r_i, c_j, t_s (i, j, s = 1,2,...,v)$ are fixed effects denoting in order the general mean, the row, the column and the treatment effects. The e_{ijs} is the error component, assumed to be independently and normally distributed with zero mean and a constant variance, σ^2.

The analysis is conducted by following a similar procedure as described for the analysis of two-way classified data. The different sums of squares are obtained as below: Let the data be arranged first in a *row × column* table such that y_{ij} denotes the observation of (*i*, *j)*th cell of table.

Let $R_i = \sum_j y_{ij} = i^{th}$ *row total* $(i = 1,2,...,v)$,

$C_j = \sum_i y_{ij} = j^{th}$ *column total* $(j = 1,2,...,v)$, T_s = sum of those observations which come from s^{th} treatment *(s= 1,2,...,v),* $G = \sum_i R_i =$ *grand total.* Correction factor, $C.F. = \dfrac{G^2}{v^2}$. Treatment sum of squares = $\sum_s \dfrac{T_s^2}{v} - C.F.$, Row sum of squares = $\sum_i \dfrac{R_i^2}{v} - C.F.$, Column sum of squares = $\sum_j \dfrac{C_j^2}{v} - C.F.$

Analysis of Variance of *v × v* Latin Square Design

Sources of Variation	D.F.	S.S.	M.S.	F
Rows	$v - 1$	$\sum_i \frac{R_i^2}{v} - C.F.$		
Columns	$v - 1$	$\sum_j \frac{C_j^2}{v} - C.F.$		
Treatments	$v - 1$	$\sum_s \frac{T_s^2}{v} - C.F.$	s_t^2	s_t^2 / s_e^2
Error	$(v - 1)(v - 2)$	By subtraction	s_e^2	
Total	$v^2 - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis of equal treatment effects is tested by *F*-test, where *F* is the ratio of treatment mean squares to error mean squares. If *F* is not significant, treatment effects do not differ significantly among themselves. If *F* is significant, further studies to test the significance of any treatment contrast can be made in exactly the same way as discussed for randomized block designs.

ANCOVA

Anindita Datta
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
anindita.datta@icar.gov.in

Introduction

The meaning of ANVOVA is Analysis of Covariance. It is a general linear model with one continuous outcome variable (quantitative) and one or more factor variables (qualitative). ANCOVA is a merger of ANOVA and regression for continuous variables. ANCOVA tests whether certain factors have an effect on the outcome variable after removing the variance for which quantitative predictors (covariates) account. The inclusion of covariates can increase statistical power because it accounts for some of the variability.

It is well known that in designed experiments the ability to detect existing differences among treatments increases as the size of the experimental error decreases, a good experiment attempts to incorporate all possible means of minimizing the experimental error. Besides proper experimentation, a proper data analysis also helps in controlling experimental error. In situations where blocking alone may not be able to achieve adequate control of experimental error, proper choice of data analysis may help a great deal. By measuring one or more *covariates* - the characters whose functional relationships to the character of primary interest are known - the Analysis of Covariance (ANCOVA) can reduce the variability among experimental units by adjusting their values to a common value of the covariates. For example, in an animal feeding trial, the initial body weight of the animals usually differs. Using this initial body weight as a covariate, the final weights recorded after the animals have been subjected to various physiological feeds (treatments) can be adjusted to the values that would have been obtained had there been no variation in the initial body weights of the animals at the start of the experiment. An another example, in a field experiment where rodents have (partially) damaged some of the plots, covariance analysis with rodent damage as a covariate could be useful in adjusting plot yields to the levels that they should have been had there been no rodent damage in any plot.

ANCOVA requires measurement of the character of primary interest plus the measurement of one or more variables known as *covariates*. It also requires that the functional relationship of the covariates with the character of primary interest is known beforehand. Generally a linear relationship is assumed, though other type of relationships could also be assumed.

Consider the case of a variety trial in which weed incidence is used as a covariate. With a known functional relationship between weed incidence and grain yield, the character of primary interest, the covariance analysis can adjust grain yield in each plot to a common level of weed incidence. With this adjustment, the variation in yield due to weed incidence is quantified and effectively separated from that due to varietal difference.

ANCOVA can be applied to any number of covariates and to any type of functional relationship between variables *viz.* quadratic, inverse polynomial, etc. Here we illustrate the use of covariance analysis with the help of a single covariate that is linearly related with the character of primary interest. It is expected that this simplification shall not unduly reduce the applicability of the technique, as a single covariate that is linearly related with the primary variable is adequate for most of the experimental situations in agricultural research.

Uses of Covariance Analysis in Agricultural Research

There are several important uses of covariance analysis in agricultural research. Some of the most important ones are:

1. To control experimental error and to adjust treatment means.
2. To aid in the interpretation of experimental results.
3. To estimate missing data.

Error Control and Adjustment of Treatment Means

It is now well realized that the size of experimental error is closely related to the variability between experimental units. It is also known that proper blocking can reduce experimental error by maximizing the differences between the blocks and thus minimizing differences within blocks. Blocking, however, can not cope with certain types of variability such as spotty soil heterogeneity and unpredictable insect incidence. In both instances, heterogeneity between experimental plots does not follow a definite pattern, which causes difficulty in getting maximum differences between blocks. Indeed, blocking is ineffective in the case of nonuniform insect incidences because blocking must be done before the incidence occurs. Furthermore, even though it is true that a researcher may have some information on the probable path or direction of insect movement, unless the direction of insect movement coincides with the soil fertility gradient, the choice of whether soil heterogeneity or insect incidence should be the criterion for blocking is difficult. The choice is especially difficult if both sources of variation have about the same importance.

Use of covariance analysis should be considered in experiments in which blocking couldn't adequately reduce the experimental error. By measuring an additional variable (*e.g.,* covariate X) that is known to be linearly related to the primary variable Y, the source of variation associated with the covariate can be deducted from experimental error. This adjusts the primary variable Y linearly upward or downward, depending on the relative size of its respective covariate. The adjustment accomplishes two important improvements:

1. The treatment mean is adjusted to a value that it would have had; had there been no differences in the values of the covariate.
2. The experimental error is reduced and the precision for comparing treatment means is increased.

Although blocking and covariance techniques are both used to reduce experimental error, the differences between the two techniques are such that they are usually not interchangeable. The ANCOVA can be used only when the covariate representing the heterogeneity among the experimental units can be measured quantitatively. However, that is not a necessary condition for blocking. In addition, because blocking is done before the start of the experiment, it can be used only to cope with sources of variation that are known or predictable. ANCOVA, on the other hand, can take care of unexpected sources of variation that occur during the experiment. Thus, ANCOVA is useful, as a supplementary procedure to take care of sources of variation that cannot be accounted for by blocking.

When covariance analysis is used for error control and adjustment of treatment means, the covariate must not be affected by the treatments being tested. Otherwise, the adjustment removes both the variation due to experimental error and that due to treatment effects. A good example of covariates that are free of treatment effects are those that are measured before the treatments are applied, such as soil analysis and residual effects of treatments applied in the past experiments. In other cases, care must be exercised to ensure that the covariates defined are not affected by the treatments being tested. This technique can be illustrated through the following example:

Example 1: A trial was designed to evaluate 15 rice varieties grown in soil with a toxic level of iron. The experiment was in a RCB design with three replications. Guard rows of a susceptible check variety were planted on two sides of each experimental plot. Scores for tolerance for iron toxicity were collected from each experimental plot as well as from guard rows. For each experimental plot, the score of susceptible check (averaged over two guard rows) constitutes the value of the covariate for that plot. Data on the tolerance score of each variety (Y variable) and on the score of the corresponding susceptible check (X variable) are shown below:

Scores of tolerance for iron toxicity (Y) of 15 rice varieties and those of the corresponding guard rows of a susceptible check variety (X) in a RCB trial

Variety Number	Replication-I		Replication-II		Replication-III	
	X	Y	X	Y	X	Y
1.	15	22	16	13	16	14
2.	16	14	15	23	15	23
3.	15	24	15	24	15	23
4.	16	13	15	23	15	23
5.	17	17	17	16	16	16
6.	16	14	15	23	15	23
7.	16	13	15	23	16	13
8.	16	16	17	17	16	16
9.	17	14	15	23	15	24
10.	17	17	17	17	15	26
11.	16	15	15	24	15	25
12.	16	15	15	23	15	23
13.	15	24	15	24	16	15
14.	15	25	15	24	15	23
15.	15	24	15	25	16	16

The usual analysis of variance without using the covariate (X variable) is as follows:

Source	**DF**	**SS**	**Mean Square**	**F Value**	**Pr > F**
Replication	2	104.0444	52.0222	2.85	0.0745
Treatment	14	265.9111	18.9937	1.04	0.4448
Error	28	510.6222	18.2365		
Total	**44**	**880.5778**			

R-Square	**C.V.**	**Root MSE**	**Y - Mean**
0.4201	21.5436	4.2704	19.82222

Using the covariate, the analysis is the following:

Source	**DF**	**S.S.**	**M.S.**	**F-Value**	**Pr > F**
Replication	2	22.4802	11.2402	2.71	0.0844
Treatment	14	152.5606	10.8972	2.63	0.0151
Covariate X	1	398.7516	398.7516	96.24	0.0001
Error	27	111.8707	4.1434		

R-Square	**C.V.**	**Root MSE**	**Y Mean**
0.8730	10.2689	2.0355	19.8222

It is interesting to note that the use of a covariate has resulted into a considerable reduction in the error mean square and hence the CV has also reduced drastically. This has helped in catching the small differences among the treatment effects as significant. This was not possible when the covariate was not used. The covariance analysis will thus result into a more precise comparison of treatment effects.

The probability of significance of pairwise comparisons among the least square estimates of the treatment effects are given below:

```
                    Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

i/j    1       2       3       4       5       6       7       8       9

1      .       0.3370  0.0666  0.4431  0.0019  0.3370  1.0000  0.0252  0.0232

2   0.3370       .     0.3370  0.8425  0.0237  1.0000  0.3370  0.1834  0.1697

3   0.0666  0.3370       .     0.2497  0.1620  0.3370  0.0666  0.6757  0.6751

4   0.4431  0.8425  0.2497       .     0.0157  0.8425  0.4431  0.1320  0.1191

5   0.0019  0.0237  0.1620  0.0157       .     0.0237  0.0019  0.2361  0.2493

6   0.3370  1.0000  0.3370  0.8425  0.0237       .     0.3370  0.1834  0.1697

7   1.0000  0.3370  0.0666  0.4431  0.0019  0.3370       .     0.0252  0.0232

8   0.0252  0.1834  0.6757  0.1320  0.2361  0.1834  0.0252       .     0.9727

9   0.0232  0.1697  0.6751  0.1191  0.2493  0.1697  0.0232  0.9727     .

10  0.0001  0.0019  0.0237  0.0012  0.3370  0.0019  0.0001  0.0361  0.0385

11  0.0874  0.4294  0.8575  0.3249  0.1046  0.4294  0.0874  0.5445  0.5439

12  0.2497  0.8425  0.4431  0.6915  0.0351  0.8425  0.2497  0.2493  0.2361

13  0.1270  0.5524  0.7066  0.4294  0.0739  0.5524  0.1270  0.4298  0.4229

14  0.0446  0.2497  0.8425  0.1803  0.2158  0.2497  0.0446  0.8096  0.8204

15  0.0589  0.3249  0.9860  0.2393  0.1452  0.3249  0.0589  0.6736  0.6809

 Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

  i/j     10      11     12        13        14        15

  1     0.0001  0.0874  0.2497    0.1270    0.0446    0.0589

  2     0.0019  0.4294  0.8425    0.5524    0.2497    0.3249

  3     0.0237  0.8575  0.4431    0.7066    0.8425    0.9860

  4     0.0012  0.3249  0.6915    0.4294    0.1803    0.2393

  5     0.3370  0.1046  0.0351    0.0739    0.2158    0.1452

  6     0.0019  0.4294  0.8425    0.5524    0.2497    0.3249

  7     0.0001  0.0874  0.2497    0.1270    0.0446    0.0589
```

8	0.0361	0.5445	0.2493	0.4298	0.8096	0.6736
9	0.0385	0.5439	0.2361	0.4229	0.8204	0.6809
10	.	0.0124	0.0031	0.0079	0.0351	0.0191
11	0.0124	.	0.5524	0.8425	0.7066	0.8425
12	0.0031	0.5524	.	0.6915	0.3370	0.4294
13	0.0079	0.8425	0.6915	.	0.5671	0.6915
14	0.0351	0.7066	0.3370	0.5671	.	0.8575
15	0.0191	0.8425	0.4294	0.6915	0.8575	.

References

Cochran, W. G., and Cox, G. M. (1957). *Experimental Design*, 2nd edition. New York: Wiley.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York.

FACTORIAL EXPERIMENTS

Seema Jaggi and Anindita Datta
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
seema.jaggi@icar.gov.in, anindita.datta@icar.gov.in

1. Introduction

Factorial Experiments are experiments that investigate the effects of two or more factors or input parameters on the output response of a process. Factorial experiment design, or simply factorial design, is a systematic method for formulating the steps needed to successfully implement a factorial experiment. Estimating the effects of various factors on the output of a process with a minimal number of observations is crucial to being able to optimize the output of the process.

In a factorial experiment, the effects of varying the levels of the various factors affecting the process output are investigated. Each complete trial or replication of the experiment takes into account all the possible combinations of the varying levels of these factors. Effective factorial design ensures that the least number of experiment runs are conducted to generate the maximum amount of information about how input variables affect the output of a process.

For example, an experiment on rooting of cuttings involving two factors, each at two levels, such as two hormones at two doses, is referred to as a 2 x 2 or a 2^2 factorial experiment. Its treatments consist of the following four possible combinations of the two levels in each of the two factors.

Treatment number	**Treatment Combination**	
	Hormone	**Dose (ppm)**
1	NAA	10
2	NAA	20
3	IBA	10
4	IBA	20

The total number of treatments in a factorial experiment is the product of the number of levels of each factor; in the 2^2 factorial example, the number of treatments is 2 x 2 = 4, in the 2^3 factorial, the number of treatments is 2 x 2 x 2 = 8. The number of treatments increases rapidly with an increase in the number of factors or an increase in the levels in each factor. For a factorial experiment involving 5 clones, 4 espacements, and 3 weed-control methods, the total number of treatments would be 5 x 4 x 3 = 60. Thus, indiscriminate use of factorial experiments has to be avoided because of their large size, complexity, and cost. Furthermore, it is not wise to commit oneself to a large experiment at the beginning of the investigation when several small preliminary experiments may offer promising results. For example, a tree breeder has collected 30 new clones from a neighbouring country and wants to assess their reaction to the local environment. Because the environment is expected to vary in terms of soil fertility, moisture levels, and so on, the ideal experiment would be one that tests the 30 clones in a factorial experiment involving such other variable factors as fertilizer, moisture level, and population density. Such an experiment, however, becomes extremely large as factors other than clones are added. Even if only one factor, say nitrogen or fertilizer with three levels were included, the number of treatments would increase from 30 to 90. Such a large experiment would mean difficulties in financing, in obtaining an adequate experimental area, in controlling soil heterogeneity, and so on. Thus, the more practical approach would be to test the 30 clones first in a single-factor experiment, and then use the results to select a few clones for further studies in more detail. For example, the initial single-factor experiment may show that only five clones are outstanding enough to warrant further testing. These five clones could then be put into a factorial experiment with three levels of nitrogen, resulting in an experiment with 15 treatments rather than the 90 treatments needed with a factorial experiment with 30 clones.

The amount of change produced in the process output for a change in the 'level' of a given factor is referred to as the 'main effect' of that factor. Table 1 shows an example of a simple factorial experiment involving two factors with two levels each. The two levels of each factor may be denoted as 'low' and 'high', which are usually symbolized by '-' and '+' in factorial designs, respectively.

Table 1. A Simple 2-Factorial Experiment

	A (-)	**A (+)**
B (-)	20	40
B (+)	30	52

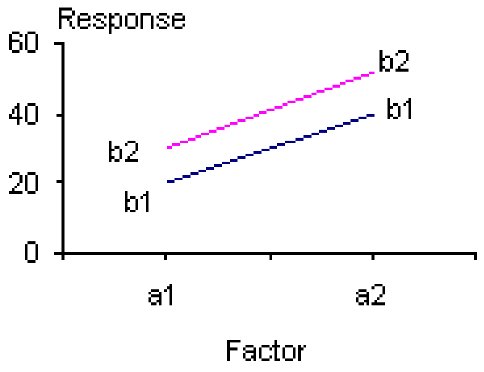
The main effect of a factor is basically the 'average' change in the output response as that factor goes from '-' to '+'. Mathematically, this is the average of two numbers: 1) the change in output when the factor goes from low to high level as the other factor stays low, and 2) the change in output when the factor goes from low to high level as the other factor stays high.

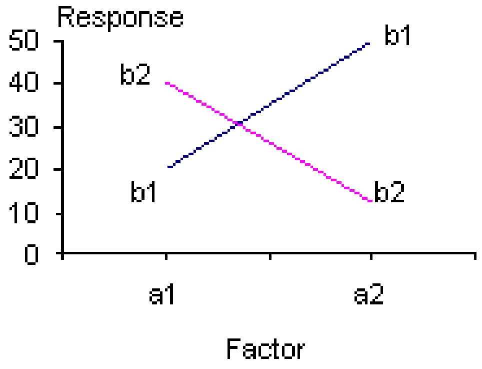
In the example in Table 1, the output of the process is just 20 (lowest output) when both A and B are at their '-' level, while the output is maximum at 52 when both A and B are at their '+' level. The main effect of A is the average of the change in output response when B stays '-' as A goes from '-' to '+', or (40-20) = 20, and the change in output response when B stays '+' as A goes from '-' to '+', or (52-30) = 22. The main effect of A, therefore, is equal to 21.

Similarly, the main effect of B is the average change in output as it goes from '-' to '+' , i.e., the average of 10 and 12, or 11. Thus, the main effect of B in this process is 11. Here, one can see that the factor A exerts a greater influence on the output of process, having a main effect of 21 versus factor B's main effect of only 11. It must be noted that aside from 'main effects', factors can likewise result in 'interaction effects.' Interaction effects are changes in the process output caused by two or more factors that are interacting with each other. Large interactive effects can make the main effects insignificant, such that it becomes more important to pay attention to the interaction of the involved factors than to investigate them individually. In Table 1, as effects of A (B) is not same at all the levels of B (A) hence, A and B are interacting.

Thus, **interaction** is the failure of the differences in response to changes in levels of one factor, to retain the same order and magnitude of performance through out all the levels of other factors OR the factors are said to interact if the effect of one factor changes as the levels of other factor(s) changes.

Graphical representation of lack of interaction between factors and interaction between factors are shown below. In case of two parallel lines, the factors are non-interacting.





If interactions exist which is fairly common, we should plan our experiments in such a way that they can be estimated and tested. It is clear that we cannot do this if we vary only one factor at a time. For this purpose, we must use multilevel, multifactor experiments.

The running of factorial combinations and the mathematical interpretation of the output responses of the process to such combinations is the essence of factorial experiments. It allows to understand which factors affect the process most so that improvements (or corrective actions) may be geared towards these.

We may define factorial experiments as experiments in which the effects (main effects and interactions) of more then one factor are studied together. In general if there are 'n' factors, say, F_1, F_2,..., F_n and i^{th} factor has s_i levels, i=1,...,n, then total number of treatment combinations is $\prod_{i=1}^{n} s_i$. Factorial experiments are of two types.

Experiments in which the number of levels of all the factors are same i.e all s_i's are equal are called **symmetrical factorial experiments** and the experiments in at least two of the s_i's are different are called as **asymmetrical factorial experiments**. Factorial experiments provide an opportunity to study not only the individual effects of each factor but also there interactions. They have the further advantage of economising on experimental resources. When the experiments are conducted factor by factor much more resources are required for the same precision than when they are tried in factorial experiments.

2. Experiments with Factors Each at Two Levels

The simplest of the symmetrical factorial experiments are the experiments with each of the factors at 2 levels. If there are 'n' factors each at 2 levels, it is called as a 2^n factorial where the power stands for the number of factors and the base the level of each factor. Simplest of the symmetrical factorial experiments is the 2^2 factorial experiment i.e. 2 factors say A and B each at two levels say 0 (low) and 1 (high). There will be 4 treatment combinations which can be written as

00 $= a_0 b_0 = 1$; A and B both at first (low) levels

10 $= a_1 b_0 = a$; A at second (high) level and B at first (low) level

01 $= a_0 b_1 = b$; A at first level (low) and B at second (high) level

11 $= a_1 b_1 = ab$; A and B both at second (high) level.

In a 2^2 factorial experiment wherein r replicates were run for each combination treatment, the main and interactive effects of A and B on the output may be mathematically expressed as follows:

$A = [ab + a - b - (1)] / 2r$; (main effect of factor A)

$B = [ab + b - a - (1)] / 2r$; (main effect of factor B)

$AB = [ab + (1) - a - b] / 2r$; (interactive effect of factors A and B)

where r is the number of replicates per treatment combination; a is the total of the outputs of each of the r replicates of the treatment combination a (A is 'high and B is 'low); b is the total output for the n replicates of the treatment combination b (B is 'high' and A is 'low); ab is the total output for the r replicates of the treatment combination ab (both A and B are 'high'); and (1) is the total output for the r replicates of the treatment combination (1) (both A and B are 'low').

Had the two factors been independent, then $[ab + (1) - a - b] / 2n$ will be of the order of zero. If not then this will give an estimate of interdependence of the two factors and it is called the interaction between A and B. It is easy to verify that the interaction of the factor B with factor A is BA which will be same as the interaction AB and hence the interaction does not depend on the order of the factors. It is also easy to verify that the main effect of factor B, a contrast of the treatment totals is orthogonal to each of A and AB.

Table 2. Two-level 2-Factor Full-Factorial

RUN	Comb.	M	A	B	AB
1	(1)	+	-	-	+
2	a	+	+	-	-
3	b	+	-	+	-
4 = 2^2	ab	+	+	+	+

Consider the case of 3 factors A, B, C each at two levels (0 and 1) i.e. 2^3 factorial experiment. There will be 8 treatment combinations which are written as

000 = $a_0\, b_0\, c_0$ = (1); A, B and C all three at first level
100 = $a_1\, b_0\, c_0$ = a ; A at second level and B and C at first level
010 = $a_0\, b_1\, c_0$ = b ; A and C both at first level and B at second level
110 = $a_1\, b_1\, c_0$ = ab; A and B both at second level and C is at first level.
001 = $a_0\, b_0\, c_1$ = c ; A and B both at first level and C at second level.
101 = $a_1\, b_0\, c_1$ = ac; A and C at second level, B at first level
011 = $a_0\, b_1\, c_1$ = bc; A is at first level and B and C both at second level
111 = $a_1\, b_1\, c_1$ = abc; A, B and C all the three at second level

In a three factor experiment there are three main effects A, B, C; 3 first order or two factor interactions AB, AC, BC; and one second order or three factor interaction ABC.

Table 3. Two-level 3-Factor Full-Factorial Experiment Pattern

RUN	Comb.	M	A	B	AB	C	AC	BC	ABC
1	(1)	+	-	-	+	-	+	+	-
2	A	+	+	-	-	-	-	+	+
3	B	+	-	+	-	-	+	-	+
4	Ab	+	+	+	+	-	-	-	-
5	C	+	-	-	+	+	-	-	+
6	Ac	+	+	-	-	+	+	-	-
7	Bc	+	-	+	-	+	-	+	-
8 = 2^3	Abc	+	+	+	+	+	+	+	+

$$\text{Main effect A} = \frac{1}{4}\{[abc] -[bc] +[ac] -[c] + [ab] -[b] + [a] -[1]\}$$

$$= \frac{1}{4}(a-1)\,(b+1)\,(c+1)$$

$$AB = \frac{1}{4}\ [(abc)-(bc) -(ac) +c) - (ab) - (b) - (a)+ (1)\ ]$$

$$ABC = \frac{1}{4}\ [\ (abc) - (bc) - (ac) + (c) - (ab) + (b) + (a) - (1)\ ]$$

or equivalently,

$$AB \quad = \ \frac{1}{4}\ (a-1)\ (b-1)\ (c+1)$$

$$ABC = \frac{1}{4}\ (a-1)\ (b-1)\ (c-1)$$

The method of representing the main effect or interaction as above is due to Yates and is very useful and quite straightforward. For example, if the design is 2^4 then

$$A \ = (1/2^3)\ \ [\ (a-1)\ (b+1)\ (c+1)\ (d+1)\ ]$$

$$ABC \ = \ (1/2^3)\ \ [\ (a-1)\ (b-1)\ (c-1)\ (d+1)]$$

In case of a 2^n factorial experiment, there will be 2^n (=v) treatment combinations with 'n' main effects, $\binom{n}{2}$ first order or two factor interactions, $\binom{n}{3}$ second order or three factor interactions, $\binom{n}{4}$ third order or four factor interactions and so on , $\binom{n}{r}$, (r-1)th order or r factor interactions and $\binom{n}{n}$ (n-1)th order or n factor interaction. Using these v treatment combinations, the experiment may be laid out using any of the suitable experimental designs viz. completely randomised design or block designs or row-column designs, etc.

Steps for Analysis

1. The Sum of Squares (S.S.) due to treatments, replications [in case randomised block design is used], due to rows and columns (in case a row-column design has been used), total S.S. and error S.S. is obtained as per established procedures. No replication S.S. is required in case of a completely randomised design.
2. The treatment sum of squares is divided into different components viz. main effects and interactions each with single d.f. The S.S. due to these factorial effects is obtained by dividing the squares of the factorial effect total by $r.2^n$. For obtaining 2^n-1 factorial effects in a 2^n factorial experiment, the 'n' main effects is obtained by giving the positive signs to those treatment totals where the particular factor is at second level and minus to others and dividing the value so obtained by $r.2^{n-1}$, where r is the number of replications of the treatment combinations. All

interactions can be obtained by multiplying the corresponding coefficients of main effects.

For a 2^2 factorial experiment, the S.S. due to a main effect or the interaction effect is obtained by dividing the square of the effect total by 4r. Thus,

S.S. due to main effect of A = $[A]^2/4r$, with 1 d.f.

S.S. due to main effect of B = $[B]^2/4r$, with 1 d.f

S.S. due to interaction AB = $[AB]^2/4r$, with 1 d.f.

3. Mean squares (M.S) is obtained by dividing each S.S. by corresponding degrees of freedom.
4. After obtaining the different S.S.'s, the usual Analysis of variance (ANOVA) table is prepared and the different effects are tested against error mean square and conclusions drawn.
5. Standard errors (S.E.'s) for main effects and two factor interactions:

S.E of difference between main effect means = $\sqrt{\frac{2MSE}{r.2^{n-1}}}$

S.E of difference between A means at same level of B=S.E of difference between B means at same level of A= $\sqrt{\frac{2MSE}{r.2^{n-2}}}$

In general,

S.E. for difference between means in case of a r-factor interaction = $\sqrt{\frac{2MSE}{r.2^{n-r}}}$

The critical differences are obtained by multiplying the S.E. by the student's t value at α% level of significance at error degrees of freedom.

The ANOVA for a 2^2 factorial experiment with r replications conducted using a RCBD is as follows:

ANOVA

Sources of Variation	**DF**	**S.S.**	**M.S.**	**F**
Between Replications	r-1	SSR	MSR=SSR/(r-1)	MSR/MSE
Between treatments	2^2-1=3	SST	MST=SST/3	MST/MSE
A	1	SSA=$[A]^2/4r$	MSA=SSA	MSA/MSE
B	1	SSB=$[B]^2/4r$	MSB=SSB	MSB/MSE
AB	1	SSAB=$[AB]^2/4r$	MSAB=SSAB	MSAB/MSE
Error	3(r-1)	SSE	MSE=SSE/3(r-1)	
Total	4r-1	TSS		

ANOVA for a 2^3-factorial experiment conducted in RCBD with r replications is given by

ANOVA

Sources of Variation	**DF**	**SS**	**MS**	**F**
Between Replications	r-1	SSR	MSR=SSR/(r-1)	MSR/MSE
Between treatments	2^3 -1=7	SST	MST=SST/7	MST/MSE
A	1	SSA	MSA=SSA	MSA/MSE
B	1	SSB	MSB=SSB	MSB/MSE
C	1	SSC	MSC=SSC	MSC/MSE
AB	1	SSAB	MSAB=SSAB	MSAB/MSE
AC	1	SSAC	MSAC=SSAC	MSAC/MSE
BC	1	SSBC	MSBC=SSBC	MSBC/MSE
ABC	1	SSABC	MSABC=SSABC	MSABC/MSE
Error	$(r-1)(2^3-1)$ $=7(r-1)$	SSE	MSE=SSE/7(r-1)	
Total	$r.2^3-1=8r-1$	TSS		

Similarly ANOVA table for a 2^n factorial experiment can be made.

3. Experiments with Factors Each at Three Levels

When factors are taken at three levels instead of two, the scope of an experiment increases. It becomes more informative. A study to investigate if the change is linear or quadratic is possible when the factors are at three levels. The more the number of levels, the better, yet the number of the levels of the factors cannot be increased too much as the size of the experiment increases too rapidly with them. Consider two factors A and B, each at three levels say 0, 1 and 2 (3^2-factorial experiment). The treatment combinations are

00	$= a_0b_0$	$= 1$	; A and B both at first levels
10	$= a_1b_0$	$= a$	; A is at second level and B is at first level
20	$= a_2b_0$	$= a^2$	; A is at third level and b is at first level
01	$= a_0b_1$	$= b$	; A is at first level and B is at second level
11	$= a_1b_1$	$= ab$	; A and B both at second level
21	$= a_2b_1$	$= a^2b$	; A is at third level and B is at second level
02	$= a_0b_2$	$= b^2$	; A is at first level and B is at third level
12	$= a_1b_2$	$= ab^2$	; A is at second level and B is at third level
22	$= a_2b_2$	$= a^2b^2$	; A and B both at third level

Any standard design can be adopted for the experiment.

The main effects A, B can respectively be divided into linear and quadratic components each with 1 d.f. as A_L, A_Q, B_L and B_Q. Accordingly AB can be partitioned into four components as A_L B_L , A_L B_Q, A_Q B_L, A_Q B_Q.

The coefficients of the treatment combinations to obtain the above effects are given as

Treatment Totals→ Factorial Effects ↓	**[1]**	**[a]**	**[a^2]**	**[b]**	**[ab]**	**[a^2b]**	**[b^2]**	**[ab^2]**	**[a^2 b^2]**	**Divisor**
M	+1	+1	+1	+1	+1	+1	+1	+1	+1	$9r=r\times3^2$
A_L	-1	0	+1	-1	0	+1	-1	0	+1	6r=r×2× 3
A_Q	+1	-2	+1	+1	-2	+1	+1	-2	+1	18r=6×3
B_L	-1	-1	-1	0	0	0	+1	+1	+1	6r=r×2×3
A_L B_L	+1	0	-1	0	0	0	-1	0	+1	4r=r×2×2
A_Q BL	-1	+2	-1	0	0	0	+1	-2	+1	12r=r×6×2
B_Q	+1	+1	+1	-2	-2	-2	+1	+1	+1	18r=r×3×6
A_L B_Q	-1	0	+1	+2	0	-2	-1	0	+1	12r=r×2×6
A_Q B_Q	+1	-2	+1	-2	+4	-2	+1	-2	+1	36r=r×6×6

The rule to write down the coefficients of the linear (quadratic) main effects is to give a coefficient as +1 (+1) to those treatment combinations containing the third level of the corresponding factor, coefficient as 0(-2) to the treatment combinations containing the second level of the corresponding factor and coefficient as -1(+1) to those treatment combinations containing the first level of the corresponding factor. The coefficients of the treatment combinations for two factor interactions are obtained by multiplying the corresponding coefficients of two main effects. The various factorial effect totals are given as

$[A_L] \quad = +1[a^2b^2]+0[ab^2] -1[b^2]+1[a^2b]+0[ab] -1[b]+1[a^2]+0[a] -1[1]$

$[A_Q] \quad = +1[a^2b^2] -2[ab^2]+1[b^2]+1[a^2b] -2[ab]+1[b]+1[a^2] -2[a]+1[1]$

$[B_L] \quad = +1[a^2b^2]+1[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2] -1[a] -1[1]$

$[A_LB_L] = +1[a^2b^2]+0[ab^2] -1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2]+0[a] -1[1]$

$[A_QB_L] = +1[a^2b^2] -2[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2]+2[a] -1[1]$

$[B_Q] \quad = +1[a^2b^2]+1[ab^2]+1[b^2] -2[a^2b] -2[ab] -2[b] -1[a^2] -1[a] -1[1]$

$[A_LB_Q] = +1[a^2b^2]+0[ab^2] -1[b^2] -2[a^2b]+0[ab]+2[b]+1[a^2]+0[a] -1[1]$

$[A_QB_Q] = +1[a^2b^2] -2[ab^2]+1[b^2] -2[a^2b]+4[ab] -2[b]+1[a^2] -2[a]+1[1]$

Factorial effects are given by

$A_L = [A_L]/r.3$ $A_Q= [A_Q]/r.3$ $B_L = [B_L]/r.3$ $A_LB_L = [A_LB_L]/r.3$

$A_QB_L = [A_QB_L]/r.3$ $B_Q = [B_Q]/r.3$ $A_LB_Q = [A_LB_Q]/r.3$ $A_QB_Q = [A_QB_Q]/r.3$

The sum of squares due to various factorial effects is given by

$SSA_L = \frac{[A_L]^2}{r.2.3}$; $SSA_q = \frac{[A_Q]^2}{r.6.3}$; $SSB_L = \frac{[B_L]^2}{r.3.2}$;

$SSA_LB_L = \frac{[A_LB_L]^2}{r.2.2}$;

$SSA_QB_L = \frac{[A_QB_L]^2}{r.6.2}$; $SSB_Q= \frac{[B_Q]^2}{r.3.6}$; $SSA_LB_Q = \frac{[A_LB_Q]^2}{r..2.6}$;

$SSA_QB_Q = \frac{[A_QB_Q]^2}{r.6.6}$;

If a RBD is used with r-replications then the outline of analysis of variance is

ANOVA

Sources of Variation	**D.f**	**SS**	**MS**
Between Replications	r-1	SSR	MSR=SSR/(r-1)
Between treatments	3^2-1=8	SST	MST=SST/8
A	2	SSA	MSA=SSA/2
A_L	1	SSA_L	$MSA_L= SSA_L$
A_Q	1	SSA_Q	$MSA_Q=SSA_Q$
B	2	SSB	MSB=SSB/2
B_L	1	SSB_L	$MSB_L= SSB_L$
B_Q	1	SSB_Q	$MSB_Q=SSB_Q$
AB	4	SSAB	MSAB=SSAB/2
A_LB_L	1	SSA_LB_L	$MSA_LB_L=SSA_LB_L$
A_QB_L	1	SSA_QB_L	$MSA_QB_L=SSA_QB_L$
A_LB_Q	1	SSA_LB_Q	$MSA_LB_Q=SSA_LB_Q$
A_QB_Q	1	SSA_QB_Q	$MSA_QB_Q=SSA_QB_Q$
Error	(r-1)(3^2-`1) =8(r-1)	SSE	MSE=SSE/8(r-1)
Total	r.3^2-1=9r-1	TSS	

In general, for n factors each at 3 levels, the sum of squares due to any linear (quadratic) main effect is obtained by dividing the square of the linear (quadratic) main effect total by r.2.3^{n-1}(r.6.3^{n-1}). Sum of squares due to a 'p' factor interaction is given by taking the square of the total of the particular interaction component divided by r.(a_1 a_2 ...a_p). 3^{n-p}, where a_1, a_2,...,a_p are taken as 2 or 6 depending upon the linear or quadratic effect of particular factor.

4. Confounding in Factorial Experiments

When the number of factors and/or levels of the factors increase, the number of treatment combinations increase very rapidly and it is not possible to accommodate all these treatment combinations in a single homogeneous block. For example, a 2^5 factorial would have 32 treatment combinations and blocks of 32 plots are quite big to ensure homogeneity within them. A new technique is therefore necessary for designing experiments with a large number of treatments. One such device is to take blocks of size less than the number of treatments and have more than one block per replication. The treatment combinations are then divided into as many groups as the number of blocks per replication. The different groups of treatments are allocated to the blocks.

There are many ways of grouping the treatments into as many groups as the number of blocks per replication. It is known that for obtaining the interaction contrast in a factorial experiment where each factor is at two levels, the treatment combinations are divided into two groups. Such two groups representing a suitable interaction can be taken to form the contrasts of two blocks each containing half the total number of treatments. In such case the contrast of the interaction and the contrast between the two block totals are given by the same function. They are, therefore, mixed up and can not be separated. In other words, the interaction has been confounded with the blocks. Evidently the interaction confounded has been lost but the other interactions and main effects can now be estimated with better precision because of reduced block size. This device of reducing the block size by taking one or more interaction contrasts identical with block contrasts is known as **confounding**. Preferably only higher order interactions, that is, interactions with three or more factors are confounded, because their loss is immaterial. As an experimenter is generally interested in main effects and two factor interactions, these should not be confounded as far as possible.

When there are two or more replications, if the same set of interactions are confounded in all the replications, confounding is called **complete** and if different sets of interaction are confounded in different replications, confounding is called **partial**. In complete confounding all the information on confounded interactions are lost. But in partial confounding, the confounded interactions can be recovered from those replications in which they are not confounded.

Advantages of Confounding

It reduces the experimental error considerably by stratifying the experimental material into homogeneous subsets or subgroups. The removal of the variation among incomplete blocks (freed from treatments) within replicates results in smaller error mean square as compared with a RBD, thus making the comparisons among some treatment effects more precise.

Disadvantages of Confounding

- In the confounding scheme, the increased precision is obtained at the cost of sacrifice of information (partial or complete) on certain relatively unimportant interactions.
- The confounded contrasts are replicated fewer times than are the other contrasts and as such there is loss of information on them and they can be estimated with a lower degree of precision as the number of replications for them is reduced.
- An indiscriminate use of confounding may result is complete or partial loss of information on the contrasts or comparisons of greatest importance. As such the experimenter should confound only those treatment combinations or contrasts which are of relatively less or of importance at all.
- The algebraic calculations are usually more difficult and the statistical analysis is complex, especially when some of the units (observations) are missing.

Confounding in 2^3 Experiment

Although 2^3 is a factorial with small number of treatment combinations but for illustration purpose, this example has been considered. Let the three factors be A, B, C each at two levels.

Factorial Effects → **Treat. Combinations ↓**	A	B	C	AB	AC	BC	ABC
(1)	-	-	-	+	+	+	-
(a)	+	-	-	-	-	+	+
(b)	-	+	-	-	+	-	-
(ab)	+	+	-	+	-	-	-
(c)	-	-	+	+	-	-	+
(ac)	+	-	+	-	+	-	-
(bc)	-	+	+	-	-	+	-
(abc)	+	+	+	+	+	+	+

The various factorial effects are as follows:

A = (abc) + (ac) + (ab) + (a) - (bc) - (c) - (b) - (1)

B = (abc) + (bc) + (ab) + (b) - (ac) - (c) - (a) - (1)

C = (abc) + (bc) + (ac) + (c) - (ab) - (b) - (a) - (1)

AB = (abc) + (c) + (ab) + (1) - (bc) - (ac) - (b) - (a)

AC = (abc) + (ac) + (b) + (1) - (bc) - (c) - (ab) - (a)

BC = (abc) + (bc) + (a) + (1) - (ac) - (c) - (ab) - (b)

ABC = (abc) + (c) + (b) + (a) - (bc) - (ac) - (ab) - (1)

Let the highest order interaction ABC be confounded and we decide to use two blocks of 4 units (plots) each per replicate.

Thus in order to confound the interaction ABC with blocks all the treatment combinations with positive sign are allocated at random in one block and those with negative signs in the other block. Thus the following arrangement gives ABC confounded with blocks and hence we loose information on ABC.

Replication I

Block 1:	(1)	(ab)	(ac)	(bc)
Block 2 :	(a)	(b)	(c)	(abc)

It can be observed that the contrast estimating ABC is identical to the contrast estimating block effects.

The other six factorial effects viz. A, B, C, AB, AC, BC each contain two treatments in block 1 (or 2) with the positive signs and two with negative sign so that they are orthogonal with block totals and hence these differences are not influenced among blocks and can thus be estimated and tested as usual without any difficulty. Whereas for confounded interaction, all the treatments in one group are with positive sign and in the other with negative signs.

Similarly if AB is to be confounded, then the two blocks will consists of

Block 1	(abc)	(c)	(ab)	(1)
Block 2	(bc)	(ac)	(b)	(a)

Here AB is confounded with block effects and cannot be estimated independently whereas all other effects A, B, C, AC, Bc and ABC can be estimated independently.

When an interaction is confounded in one replicate and not in another, the experiment is said to be partially confounded. Consider again 2^3 experiment with each replicate divided into two blocks of 4 units each. It is not necessary to confound the same interaction in all the replicates and several factorial effects may be confounded in one single experiment. For example, the following plan confounds the interaction ABC, AB, BC and AC in replications I, II, III and IV respectively.

Rep. I		**Rep. II**		**Rep. III**		**Rep. IV**	
Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**	**Block 7**	**Block 8**
(abc)	(ab)	(abc)	(ac)	(abc)	(ab)	(abc)	(ab)
(a)	(ac)	(c)	(bc)	(bc)	(ac)	(ac)	(bc)
(b)	(bc)	(ab)	(a)	(a)	(b)	(b)	(a)
(c)	(1)	(1)	(b)	(1)	(c)	(1)	(c)

In the above arrangement, the main effects A, B and C are orthogonal with block totals and are entirely free from block effects. The interaction ABC is completely confounded with blocks in replicate 1, but in the other three replications the ABC is orthogonal with blocks and consequently an estimate of ABC may be obtained from replicates II, III and IV. Similarly it is possible to recover information on the other confounded interactions AB (from I, III, IV), BC (from I, II, IV) and AC (from I, II, III). Since the partially confounded interactions are estimated from only a portion of the observations, they are determined with a lower degree of precision than the other effects.

For carrying out the statistical analysis, the various factorial effects and their S.S. are estimated in the usual manner with the modification that for **completely confounded** interactions neither the S.S due to confounded interaction is computed nor it is included in the ANOVA table. The confounded component is contained in the (2p-1) degrees of freedom (D.f.) (in case of p replicates) due to blocks. The partitioning of the d.f for a 2^3 completely confounded factorial is as follows.

Source of Variation	**D.f**
Blocks	2p-1
A	1
B	1
C	1
AB	1
AC	1
BC	1
Error	6(p-1)
Total	8p-1

In general for a 2^n completely confounded factorial in p replications, the different d.f's are given as follows

Source of Variation	**D.f**
Replication	p-1
Blocks within replication	$p(2^{n-r}-1)$
Treatments	$2^n-1-(2^{n-r}-1)$
Error	By subtraction
Total	$p2^n-1$

The treatment d.f has been reduced by $2^{n-r}-1$ as this is the total d.f confounded per block.

In case of partial confounding, we can estimate the effects confounded in one replication from the other replication in which it is not confounded. In (2^n, 2^r) factorial experiment with p replications, following is the splitting of d.f's.

Source of Variation	**D.f**
Replication	p-1
Blocks within replication	$p(2^{n-r}-1)$
Treatments	2^n-1
Error	By subtraction
Total	$p2^n-1$

The S.S. for confounded effects are to be obtained from those replications only in which the given effect is not confounded.

5. Fractional Factorial

In a factorial experiment, as the number of factors to be tested increases, the complete set of factorial treatments may become too large to be tested simultaneously in a single experiment. A logical alternative is an experimental design that allows testing of only a fraction of the total number of treatments. A design uniquely suited for experiments involving large number of factors is the fractional factorial. It provides a systematic way of selecting and testing only a fraction of the complete set of factorial treatment combinations. In exchange, however, there is loss of information on some pre-selected effects. Although this information loss may be serious in experiments with one or two factors, such a loss becomes more tolerable with large number of factors. The number of interaction effects increases rapidly with the number of factors involved, which allows flexibility in the choice of the particular effects to be sacrificed. In fact, in cases where some specific effects are known beforehand to be small or unimportant, use of the fractional factorial results in minimal loss of information.

In practice, the effects that are most commonly sacrificed by use of the fractional factorial are high order interactions - the four-factor or five-factor interactions and at times, even the three-factor interaction. In almost all cases, unless the researcher has prior information to indicate otherwise one should select a set of treatments to be tested so that all main effects and two-factor interactions can be estimated.

In forestry research, the fractional factorial is to be used in exploratory trials where the main objective is to examine the interactions between factors. For such trials, the most appropriate fractional factorials are those that sacrifice only those interactions that involve more than two factors.

With the fractional factorial, the number of effects that can be measured decreases rapidly with the reduction in the number of treatments to be tested. Thus, when the number of effects to be measured is large, the number of treatments to be tested, even with the use of fractional factorial, may still be too large. In such cases, further reduction in the size of the experiment can be achieved by reducing the number of replications. Although the use of fractional factorial without replication is uncommon in forestry experiments, when fractional factorial is applied to exploratory trials, the number of replications required can be reduced to the minimum.

Another desirable feature of fractional factorial is that it allows reduced block size by not requiring a block to contain all treatments to be tested. In this way, the homogeneity of experimental units within the same block can be improved. A reduction in block size is, however, accompanied by loss of information in addition to that already lost through the reduction in number of treatments.

DATA MINING: AN OVERVIEW

Shashi Dahiya
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
shashi.dahiya@icar.gov.in

Introduction

Rapid advances in data collection and storage technology have enables organizations to accumulate vast amounts of data. However, extracting useful information has proven extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small. In other situations, the questions that need to be answered cannot be addressed using existing data analysis techniques, and thus, new methods need to be developed.

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It has also opened up exiting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways. Data Mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation, such as predicting whether a newly arrived customer will spend more than Rs.1000 at a department store.

Data mining, or knowledge discovery, has become an indispensable technology for businesses and researchers in many fields. Drawing on work in such areas as statistics, machine learning, pattern recognition, databases, and high performance computing, data mining extracts useful information from the large data sets now available to industry and science.

Knowledge Discovery in Database

The transformation of data into knowledge has been using mostly manual methods for data analysis and interpretation, which makes the process of pattern extraction of databases too expensive, slow and highly subjective, as well as unthinkable if the volume of data is huge. The interest in automating the analysis process of great volumes of data has been fomenting several research projects in an emergent field called *Knowledge Discovery in Databases* (KDD). KDD is the process of knowledge extraction from great masses of data with the goal of obtaining meaning and consequently understanding of the data, as well as to acquire new knowledge. This process is very complex because it consists of a technology composed of a group of mathematical and technical models of software that are used to find patterns and regularities in the data.

Knowledge discovery in databases (KDD) has been defined as the *process of discovering valid, novel, and potentially useful patterns from data*. Let us examine these terms in more details:

- Data is a set of facts *F* (e.g. cases in databases).
- Pattern is an expression *E* in a language L describing facts in a subset $F_E$of *F*. *E* is called a pattern if it simpler than the enumeration of all facts in F_E.
- Process: Usually in KDD is a multi step process, which involves data preparation, search for patterns, knowledge evaluation, and refinement involving iteration after modification. The process is assumed to be non-trivial-that is, to have some degree of search autonomy.
- Validity: The discovered patterns should be valid on new data with some degree of certainty.
- Novel: The patterns are novel (at least to the system). Novelty can be measured with respect to changes in data (by comparing current values to previous or expected values) or knowledge (how a new finding is related to old ones). In general, it can be measured by a function *N (E, F)*, which can be a Boolean function or a measure of degree of novelty or unexpectedness.
- Potentially useful: The patterns should potentially lead to some useful actions, as measured by some utility function. Such a function U maps expressions in L to a partially or totally ordered measure space M_U: hence u=*U (E,F).*
- Ultimately Understandable: A goal of KDD is to make patterns understandable to humans in order to facilitate a better understanding of the underlying data. While this is difficult to measure precisely, one frequent substitute is the simplicity measure. Several measure of simplicity exist, and they range form the purely syntactic to the semantic. It is assumed that this is measured, if possible, by a function S mapping expressions E in L to a partially or totally ordered space M_S: hence, *s= S (E, F).*

An important notion, called interestingness, is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity. Some KDD systems have an explicit interestingness function $i = I(E, F, C, N, U, S)$ which maps expressions in L to a measure space M_I. Other systems define interestingness indirectly via an ordering of the discovered patterns.

Based on the notions given above, we can now make an attempt to define knowledge.

Knowledge: A pattern E$\in$is called knowledge if for some user-specified threshold i $\in M_I$, $I(E, F, C, N, U, S) > i$.

This definition of knowledge is purely user-oriented and determined by whatever functions and thresholds the user chooses.

To extract knowledge from databases, it is essential that the *Expert* follows some steps or basic stages in order to find a path from the raw data to the desired knowledge. The KDD process organizes these stages in a sequential and iterative form. In this way, it would be interesting if the obtained results of these steps were analyzed in a more interactive and friendly way, seeking a better evaluation of these results. The process of knowledge extraction from databases combines methods and statistical tools, machine learning and databases to find a mathematical and/or logical description, which can be eventually complex, of patterns and regularities in data. The knowledge extraction from a large amount of data should be seen as an interactive and iterative process, and not as a system of automatic analysis.

The interactivity of the KDD process refers to the greater understanding, on the part of the users of the process, of the application domain. This understanding involves the selection of a representative data subset, appropriate pattern classes and good approaches to evaluating the knowledge. For a better understanding the functions of the users that use the KDD process can be divided in three classes:

(a) *Domain Expert*, who should possess a large understanding of the application domain;

(b) *Analyst*, who executes the KDD process and, therefore, he should have a lot of knowledge of the stages that compose this process and

(c) *Final User*, who does not need to have much knowledge of the domain, the *Final User* uses knowledge extracted from the KDD process to aid him in a decision-making process.

KDD Process: Knowledge discovery from data can be understood as a process that contains, at least, the steps of application domain understanding, selection and preprocessing of data, Data Mining, knowledge evaluation and consolidation and use of the knowledge. The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user. Practical view of the KDD process emphasizing the interactive nature of the process outlines the following basic steps:

- **Data Selection**: Where data relevant to the analysis task are retrieved from the database.
- **Data Preprocessing**: To remove noise and inconsistent data which is called cleaning and integration of data that is combining multiple data sources.
- **Data Transformation**: Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining**: An essential process where intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation**: To identify the truly interesting patterns representing knowledge based on some interestingness measures.
- **Knowledge Presentation**: Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

The several steps of KDD have been shown in the following figure.

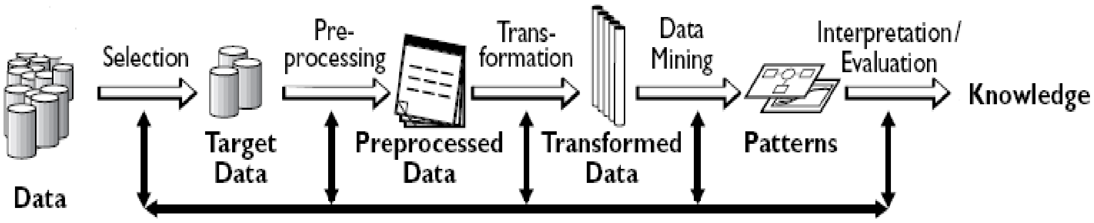


Figure: Various Steps of KDD process

The KDD process begins with the understanding of the application domain, considering aspects such as the objectives of the application and the data sources. Next, a representative sample (e.g. using statistical techniques) is removed from database, preprocessed and submitted to the methods and tools of the Data Mining stage with the objective of finding patterns/models (knowledge) in the data. This knowledge is then evaluated as to its quality and/or usefulness, so that it can be used to support a decision-making process.

The data mining component of the KDD process is mainly concerned with means by which patterns are extracted and enumerated from the data. Knowledge discovery involves the evaluation and possibly interpretation of the patterns to make the decision of what constitutes knowledge and what does not. It also includes of encoding schemes, preprocessing, sampling and projections of the data prior to the data mining step.

Data Mining

Generally, Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data Mining can be defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” and "the science of extracting useful information from large data sets or databases". Although it is usually used in relation to analysis of data, data mining, like artificial intelligence, is an umbrella term and is used with varied meaning in a range of wide contexts. It is usually associated with a business or other organization's need to identify trends.

Data Mining involves the process of analyzing data to show patterns or relationships; sorting through large amounts of data; and picking out pieces of relative information or patterns that occur e.g., picking out statistical information from some data.

The Data-Mining Communities: As data-mining has become recognized as a powerful tool, several different communities have laid claim to the subject:

1. Statistics.
2. AI, where it is called \machine learning."
3. Researchers in clustering algorithms.
4. Visualization researchers.
5. Databases.

In a sense, data mining can be thought of as algorithms for executing very complex queries on non-main-memory data.

Motivating Challenges

Traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by new data sets. The following are some of the specific challenges that motivated the development of data mining:

- **Scalability:** Because of advances in data generation and collection datasets with sizes of gigabytes, terabytes, or even petabytes are becoming common. If data mining algorithms are to handle these massive datasets, then they must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. For instance, out-of-core algorithms may be necessary when processing data sets that cannot fit into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

- **High Dimensionality:** It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. It the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken. Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data. Also, for some data analysis algorithms, the computational complexity increase rapidly as the dimensionality (the number of features) increases.
- **Heterogeneous and Complex Data:** Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. Recent years have also seen the emergence of more complex data objects. Examples of such non-traditional types of data include collections of Web pages containing semi-structured text and hyper lines; DNA data with sequential and three-dimensional structure; and climate data that consists of time series measurements (temperature, pressure, etc.) at various locations on the Earth’s surface. Techniques developed for mining such complex objects should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity, and parent-child relationships between the elements in semi-structures text and XML documents.
- **Data Ownership and Distribution:** Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple

entities. This requires the development of distributed data mining techniques. Among the key challenges faced distributed data mining algorithms include (1) how to reduce the amount of communication needed to perform the distributed computation, (2) how to effectively consolidate the data mining results obtained from multiple sources, and (3) how to address data security issues.

- **Non-Traditional Analysis:** The traditional statistical approach is based on a hypothesize-the test paradigm. In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analysed with respect to the hypothesis. Unfortunately, this process is extremely labor-intensive. Current data analysis tasks often require the generation and evaluated of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Furthermore, the data sets analyzed in data mining are typically nor the result of a carefully designed experiments and often represent opportunistic samples of the data, rather than random samples. Also, the data sets frequently involve non-traditional types of data and data distributions.

Data Preprocessing

Data preprocessing is a broad area and consists of a number of different strategies and techniques that are interrelated in complex ways. We will present some of the most important ideas and approaches, and try to point the interrelationships among them. The preprocessing techniques fall into two categories: selecting data objects and attributes for the analysis or creating/ changing the attributes. In both cases the goal is to improve the data mining analysis with respect to time, cost, and quality. Specifically, following are the important preprocessing techniques:

- **Aggregation:** Sometimes “less is more” and this is the case with aggregation, the combining of two or more objects into a single object.. Consider a dataset consisting of transactions (data objects) recording the daily sales of products in various store locations for different days over the course of a year. One way of aggregate the transactions of this data set is to replace all the transactions of a single store with a single storewide transaction. This reduces the hundreds or thousands of transactions that occur daily at a specific store to a single daily transaction, and the number of data objects is reduced to the number of stores.

An obvious issue is how an aggregate transaction is created; i.e. how the values of each attribute are combined across all the records corresponding to a particular location to create the aggregate transaction that represents the sales of a single store or date. Quantitative attributes, such as price, are typically aggregated by taking a sum or an average. A qualitative attribute, such as item, can either be omitted or summarized as the set of all the items that were sold at that location.

- **Sampling:** Sampling is a commonly used approach for selectinga subset of the data objects to be analyzed. In statistics, it has long been used for both the preliminary investigation of the data and the final data analysis. Sampling can also be very useful in data mining. However, the motivations for sampling in statistics and data mining are often different. Statisticians use sampling because obtaining the entire set of data of interest is too expensive or time consuming to process all the data. In some cases, using a sampling algorithm can reduce the data size to the point where a better, but more expensive algorithm can be used.
- **Dimensionality reduction:** Datasets can have a large number of feature. Consider set documents, where each documents is represented by a vector whose components are the frequencies with which each word occurs in the document. In such cases, there are typically thousands or tens of thousands of attributes (components), one for each word in the vocabulary. As another example, consider a set of time series consisting of the daily closing price of various stocks over a period of 30 days. In this case, the attributes, which are the prices on specific days again number in the thousands.

There is variety of benefits to dimensionality reduction. A key benefit is that many data mining algorithms work better if the dimensionality — the number of attributes in the data—is lower. This is partly because the dimensionality reduction can eliminateirrelevant features and reduce noise and partly because of the curse of dimensionality. Another benefit of dimensionality reduction is that a reduction of dimensionality can lead to a more understandable model because the model may involve fewer attributes. Also, dimensionality reduction may allow the data to be more easily visualized. Even if dimensionality reduction doesn’t reduce the data to two or three dimensions, data is often visualized by looking at pairs or triplets of attributes, and the number of such combinations is greatly reduced.

Finally, the amount of time and memory required by the data mining algorithms is reduced with a reduction in dimensionality.

- **Feature subset selection:** The term dimensionality reduction is often those techniques that reduce the dimensionality of data set by creating new attributes that are a combination of the old attributes. The reduction of dimensionality by selecting new attributes that are a subset of the old is known as feature subset selection or feature selection. While it might seem that such as approach would lose information, this is not the case if redundant and irrelevant features are present. Redundant features duplicate much or all the information contained in one or more other attributes. For example, the purchase price of a product and ge amount of sales tax paid contain much of the same information. Irrelevant features contain almost no useful information for the data mining task at hand. For instance, student's ID numbers are irrelevant to the task of predicting student's grade point averages. Redundant and irrelevant features can reduce classification accuracy and the quality of the clusters that are found.
- **Feature creation:** It is frequently possible to create, from the original attributes, a new set of attributes that captures the important information in a data set much more effectively. Furthermore, the number of new attributes can be smaller than the number of original attributes, allowing us to reap all the benefits of dimensionality reduction. Three related methodologies for creating new attributes are: feature extraction, mapping the data to a new space, and feature construction.
- **Discretization and Binarization:** Some data mining algorithms, especially certain classification algorithms, require that the data be in the form of categorical attributes. Algorithms that fine association patterns require that the data be in the form of binary attributes. Thus, it is often necessary to transform a continuous attribute into a categorical attribute (discretization), and both continuous and discrete attributes may need to be transformed into one or more binary attributes (binarization). Additionally, if a categorical attribute has a large number of values (categories), or some values occur infrequently,

then it may be beneficial for certain data mining tasks to reduce the number of categories by combining some of the values.

- **Variable transformation:** A variable transformation refers to a transformation that is applied to all the values of a variable. In other words, for each subject, the transformation is applied to the value of the variable for that object. For example, if only the magnitude of a variable is important, then the values of the variable can be transformed by taking the absolute value.

What kinds of Data can be Mined?

Data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data, and transactional data. Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the world wide web data). Techniques for mining of these kinds of data may be different. Data mining will certainly continue to embrace new data types as they emerge.

Tasks in Classical Data Mining

The two "high-level" primary goals of data mining in practice tend to be prediction and description. Data Mining tasks are generally divided into two major categories:

Predictive Tasks: the objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the *target* or *dependent variable*, while the attributes used for making the prediction are known as the *explanatory* or *independent variables*.

Descriptive Tasks: Here, the objective is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often explanatory in nature and frequently require post processing techniques to validate and explain and results.

The relative importance of prediction and description for particular data mining applications can vary considerably. However, in context of KDD, description tends to be more important than prediction.

Discovering patterns and rules: Other data mining applications are concerned with pattern detection. One example is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest. Another use is in astronomy, where detection of unusual stars or galaxies may lead to the discovery of previously unknown phenomenon. Yet another is the task of finding combinations of items that occur frequently in transaction databases (e.g., grocery products that are often purchased together). This problem has been the focus of much attention in data mining and has been addressed using algorithmic techniques based on association rules.

A significant challenge here, one that statisticians have traditionally dealt with in the context of outlier detection, is deciding what constitutes truly unusual behavior in the context of normal variability. In high dimensions, this can be particularly difficult. Background knowledge and human interpretation can be invaluable.

To achieve the goals of prediction and description, following data mining tasks are carried out.

- Classification
- Association Rule Mining
- Clustering
- Evolution Analysis
- Outlier Detection
- Dependency Modeling
- Change and Deviation Detection

1. Classification: Classification, which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications. Examples include, detecting spam email messages based upon the message header and content, categorizing cells as malignant or benign based upon the results of MRI scans, and classifying galaxies based upon their shapes.

The input data for a classification task is a collection of records. Each record, also known as an instance or example, is categorized by a tuple (x, y), where x is the attribute set and y is a special attribute, designated as the class label (also known as category or the target attribute). The attributes set in a dataset for classification can be either discrete or continuous but the class label must be a discrete attribute. This is the key characteristic that distinguishes classification from regression, a predictive modeling task in which y is a continuous attribute.

Definition (classification): Classification is the task of learning a target function *f* that maps each attribute set x to one of the predefined class labels y.

The target function is also known informally as a classification model. A classification model is useful for the following purposes.

Descriptive Modeling: A classification model can serve as an explanatory tool to distinguish between objects of different classes. For example, it would be useful-for both biologists and others-to have a descriptive model that summarizes that data shown… and explains what features define a vertebrate as a mammal, reptile, bird, fish, and amphibian.

Predictive Modeling: A classification model can also be used to predict the class label of unknown records. A classification model can be treated as a black box that automatically assigns a class label when presented with the attribute set of an unknown record.

Classification techniques are most suited for predicting or describing data sets with binary or nominal categories. They are less effective for ordinal categories (e.g., to classify a person as a member of high, medium or low income group) because they do not consider the implicit order among the categories. Other forms of relationships, such as subclass-super class relationships among categories (e.g., humans and apes are primates, which in turn is a subclass of mammals) are also ignored.

The classifier-training algorithm uses pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

2. Association Rule Mining: Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in 1993.It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared later. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub-problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence. Suppose one of the large item sets is L_k, $L_k = \{I_1, I_2, \dots , I_k\}$, association rules with this item sets are generated in the following way: the first rule is $\{I_1, I_2, \dots , I_{k-1}\}\Rightarrow\{I_k\}$, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem. The first sub-problem can be further divided into two sub-problems: candidate large item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item- sets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only “interesting” rules, generating only “non redundant” rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.

Methods for association rule mining:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

3. Clustering: Clustering or cluster analysis divides the data into groups (clusters) that are meaningful, useful or both. If meaningful groups are the goal, then the clusters should capture the natural structure of the data. In some cases, however, cluster analysis is only a useful starting point for other purposes, such as data summarization. Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering. There are various clustering methods:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

4. Evolution Analysis: Data evolution analysis describes and models regularities or trends for objects whose behaviors changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct feature of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

5. Outlier Detection: A database may contain data objects that do not comply with the general behavior or model of the data. Theses data objects are outliers. Most data mining methods discard outliers as noise as exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

6. Dependency modeling: Dependency modeling consists of finding a model that describes significant dependencies between variables. Dependency models exist at two levels: (1) the *structural level* of the model specifies (often in graphic form) which variables are locally dependent on each other and (2) the *quantitative level* of the model specifies the strengths of the dependencies using some numeric scale. For example, probabilistic dependency networks use conditional independence to specify the structural aspect of the model and probabilities or correlations to specify the strengths of the dependencies. Probabilistic dependency networks are increasingly finding applications in areas as diverse as the development of probabilistic medical expert systems from databases, information retrieval, and modeling of the human genome.

7. Change and deviation detection: Change and deviation detection focuses on discovering the most significant changes in the data from previously measured or normative values.

Components of Data Mining Algorithms

The data mining algorithms that address various data mining tasks have four basic components:

1. **Model or Pattern Structure:** Determining the underlying structure of functional forms that we seek from the data.
2. **Score Function:** Score functions are for judging the quality of a fitted model. Score Functions quantify how well a model or parameter structure fits a given data set. In an ideal world the choice of score function would precisely reflect the utility (i.e., the true expected benefit) of a particular predictive model. In practice, however, it is often difficult to specify precisely the true utility of a model's predictions. Hence, simple, "generic" score functions, such as least squares and classification accuracy are commonly used.
3. **Optimization and Search Method:** Optimizing the score function and searching over different model and pattern structures. The score function is a measure of how ell aspects of the data match proposed models or patterns. Usually, these models or patters are described in terms of a structure, sometimes with unknown parameter values. The goal of optimization and search is to determine the structure and the parameter values that achieve a minimum (or maximum, depending on the context) value of the score function. The task of finding the "best" values of parameters in models is typically cast as an optimization (for estimation) problem. The task of finding interesting patterns (such as rules) from a large family of potential patterns is typically cast as a combinatorial search problem, and is, often accomplished using heuristic search techniques. In linear regression, a prediction rule is usually found by minimizing a least squares score function (the sum of squared errors between the prediction from a model and the observed values of the predicted variable). Such a score function is amenable to mathematical manipulation, and the model that minimizes it can be found algebraically. In contrast, a score function such as misclassification rate in supervised classification is difficult to minimize analytically.
4. **Data Management Strategy:** Handling the data access efficiently during the search/optimization. The final component in any data mining algorithm is the data management strategy: the ways in which the data stored, indexed, and accessed. Most well-known data analysis algorithms in statistics and machine

learning have been developed under the assumption that all individual data points can be accessed quickly and efficiently in random-access memory(RAM), while main memory technology has improved rapidly, there have been equally rapid improvements in secondary (disk) and tertiary tape) storage technologies, to the extent that many massive data sets still reside largely on disk or tape and will not fit in available RAM. Thus, there will probably be a price to pay for accessing massive data sets, since not all data points can be simultaneously close to the main processor.

Some Challenges

A data mining system which is quick and correct on some small training sets, could behave completely different when applied to a larger database. A data mining system may work perfect for consistent data and may not perform well when a little noise is added to the training set. The most prominent challenges for data mining systems today are:

- Noisy Data
- Difficult Training Set
- Databases are Dynamic
- Databases may be Huge

Noisy Data: In a large database, many of the attribute values will be inexact or incorrect. This may be due to erroneous instruments measuring some property, or human error when registering it. We will distinguish between two forms of noise in the data, both described below:

Corrupted Values: Sometimes some of the values in the training set are altered from what they should have been. This may result in one or more tuples in the database conflict with the rules already established. The system may then regard these extreme values as noise, and ignore them. Alternatively, one may take the values into account possibly changing correct patterns recognized. The problem is that one never knows if the extreme values are correct or not, and the challenge is how to handle ``weird" values in the best manner.

Missing Attribute Values: One or more of the attribute values may be missing both for examples in the training set and for object which are to be classified. If attributes are missing in the training set, the system may either ignore this object totally, try to take it into account by for instance finding what is the missing attribute's most probable value, or use the value ``unknown" as a separate value for the attribute. When an attribute value is missing for an object during classification, the system may check all matching rules and calculate the most probable classification.

Difficult Training Set: Sometimes the training set is not the ultimate training set due to several reasons. These are the following:

Not Representative Data: If the data in the training set is not representative for the objects in the domain, we have a problem. If rules for diagnosing patients are being created and only elderly people are registered in the training set, the result for diagnosing a kid based on these data probably will not be good. Even though this may have serious consequences, we would say that not representative data is mainly a problem of machine learning when the learning is based on few examples. When using large data sets, the rules created probably are representative, as long as the data being classified belongs to the same domain as those in the training set.

No Boundary Cases: To find the real differences between two classes, some boundary cases should be present. If a data mining system for instance is to classify animals, the property counting for a bird might be that it has wings and not that it can fly. This kind of detailed distinction will only be possible if e.g. penguins are registered.

Limited Information: In order to classify an object to a specific class, some condition attributes are investigated. Sometimes, two objects with the same values for condition attributes have a different classification. Then, the objects have some properties which are not among the attributes in the training set, but still make a difference. This is a problem for the system, which does not have any way of distinguish these two types of objects.

Databases are Dynamic: Databases usually change continually. We would like rules which reflect the content of the database at all times, in order to make the best possible classification. Many existing data mining systems require that all the training examples are given at once. If something is changed at a later time, the whole learning process may have to be conducted again. An important challenge for data mining systems is to avoid this, and instead change its current rules according to updates performed.

Databases may be Huge: The size of databases seem to be ever increasing. Most machine learning algorithms have been created for handling only a small training set, for instance a few hundred examples. In order to use similar techniques in databases thousands of times bigger, much care must be taken. Having very much data is advantageous since they probably will show relations really existing, but the number of possible descriptions of such a dataset is enormous. Some possible ways of coping with this problem, are to design algorithms with lower complexity and to use heuristics to find the best classification rules. Simply using a faster computer is seldom a good solution.

References

Breiman, L., Freidman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*. Wadsworth, Belmont CA.

Larose, DT.(2006).*Data Mining Methods and Models*. Wiley-Interscience, New Jersey, USA.

Han, J., Kamber, M., Pei, J. (2012).*Data mining: concepts and techniques.* Morgan Kaufmann, Elsevier, USA.

CLASSIFICATION AND REGRESSION TREE (CART) AND SELF ORGANIZING MAPS (SOM)

Ramasubramanian V.
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
r.subramanian@icar.gov.in

1. Introduction

In certain research studies, development of a reliable decision rule, which can be used to classify new observations into some predefined categories, plays an important role. The existing traditional statistical methods are inappropriate to use in certain specific situations, or of limited utility, in addressing these types of classification problems. There are a number of reasons for these difficulties. First, there are generally many possible “predictor” variables which makes the task of variable selection difficult. Traditional statistical methods are poorly suited for this sort of multiple comparisons. Second, the predictor variables are rarely nicely distributed. Many variables (in agriculture and other real life situations) are not normally distributed and different groups of subjects may have markedly different degrees of variation or variance. Third, complex interactions or patterns may exist in the data. For example, the value of one variable (e.g., age) may substantially affect the importance of another variable (e.g., weight). These types of interactions are generally difficult to model and virtually impossible to model when the number of interactions and variables becomes substantial. Fourth, the results of traditional methods may be difficult to use. For example, a multivariate logistic regression model yields a probability for different classes of the dependent variable, which can be calculated using the regression coefficients and the values of the explanatory variable. But practitioners generally do not think in terms of probability but, rather in terms of categories, such as “presence” versus “absence.” Regardless of the statistical methodology being used, the creation of a decision rule requires a relatively large dataset.

Classification methods include the conventional clustering methods (e.g. K-means), discriminant function method and SOFMs while predictive models include decision trees (e.g., CART - Classification And Regression Trees), neural networks (the most popular type of architectures being MLP – MultiLayer Perceptron) and statistical models (e.g. MLR - Multiple Linear Regression, Logistic regression etc.). Decision trees are nothing but classification systems that predict or classify future observations based on a set of decision rules and are sometimes called rule induction methods because the reasoning process behind them is clearly evident when browsing the trees. Neural network models are used when the underlying relationship between the different variables in the system are unknown (which are complex and typically non-linear). Self-Organizing Feature Maps (SOFMs) also known as Kohonen neural networks which comes under the category of unsupervised learning which are used when the study or main or dependent variable is a categorical variable and hence such networks are used for classification purposes.

The Kohonen architecture of neural networks is a special type of architecture and is totally different from other types and solely meant for classification rather than prediction. Kohonen network offers a considerably different approach to ANNs and are designed primarily for unsupervised learning rather than for supervised problems. The very first thing to be aware of while employing any classification method or prediction model is of ascertaining whether the nature of the problem requires a 'supervised' or an 'unsupervised' approach. The supervised problem occurs when there is a known membership class or output associated with each input in the 'training' data set i.e. the set upon which the method or model will be fitted or employed. The unsupervised problem means that one deals with a set of data which have no specific associated classes or outputs attached.

In this write-up, two chief methods viz., CART and SOM in the context of classification (i.e. when the main or study or dependent variable is categorical) are discussed in detail.

1. Classification And Regression Tree (CART)

CART analysis is a tree-building technique which is different from traditional data analysis methods. In a number of studies, CART has been found to be quite effective for creating decision rules which perform as well or better than rules developed using more traditional methods aiding development of DSS (Decision Support Systems). In addition, CART is often able touncover complex interactions between predictors which may be difficult or impossible using raditional multivariate techniques. It is now possible to perform a CART analysis with a simple understanding of each of the multiple steps involved in its procedure. Classification tree methods such as CART are convenient way to produce a prediction rule from a set of observations described in terms of a vector of features and a response value. The aim is to define a general prediction rule which can be used to assign a response value to the cases solely on the bases of their predictor (explanatory) variables. Tree-structured classifications are not based on assumptions of normality and user-specified model statements, as are some conventional methods such as discriminant analysis and ordinary least square regression.

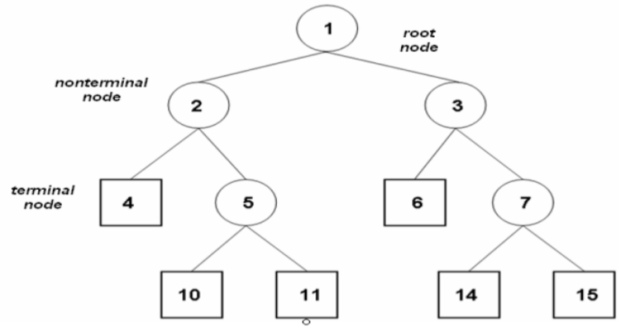
Tree based classification and regression procedure have greatly increased in popularity during the recent years. Tree based decision methods are statistical systems that mine data to predict or classify future observations based on a set of decision rules and are sometimes called rule induction methods because the reasoning process behind them is clearly evident when browsing the trees. The CART methodology have found favour among researchers for application in several areas such as agriculture, medicine, forestry, natural resources management etc. as alternatives to the conventional approaches such as discriminant function method, multiple linear regression, logistic regression etc. In CART, the observations are successively separated into two subsets based on associated variables significantly related to the response variable; this approach has an advantage of providing easily comprehensible decision strategies. CART can be applied either as a classification tree or as a regressive tree depending on whether the response variable is categorical or continuous. Tree based methods are not based on any stringent assumptions. These methods can handle large number of variables, are resistant to outliers, non-parametric, more versatile, can handle categorical variables, though computationally more intensive. They can be applied to data sets having both a large number of cases and a large number of variables, and are extremely robust to outliers. These are not based on assumptions such as normality and user-specified model statements, as are some conventional methods such as discriminant analysis or ordinary least square (OLS) regression. Yet, unlike the case for other nonparametric methods for classification and regression, such as kernel-based methods and nearest neighbor methods, the resulting tree-structured predictors can be relatively simple functions of the predictor variables which are easy to use.

CART can be a good choice for the analysts as they give fairly accurate results quickly, than traditional methods. If more conventional methods are called for, trees can still be helpful if there are a lot of variables, as they can be used to identify important variables and interactions. These are also invariant to the monotonic transformations of the explanatory variables and do not require the selection of the variable in advance as in regression analysis.

Agriculture being a highly uncertain occupation, classification and prediction in the field of agriculture aid planners to take proactive measures. Keeping in view the requirements to develop a sound classificatory system and that the potentials of the tree based methods for this purpose has not fully been explored, it will be of interest to employ these methodologies upon a suitable data set in the field of agriculture. More importantly, since the real world data often does not satisfy the usual assumptions like that of normality, homoscedasticity etc it can be taken up as a motivation to find such a classificatory rule where assumptions of such rules fail. Apart from all these, tree based methods are one among the promising data mining tools that provide easily comprehensible decision strategy.

Tree based applications originated in the 1960s with the development of AID (Automatic Interaction Detector) by Morgan and Sonquistin the 1960s as regression trees. Further modifications in this technique was carried out to result in THAID (THeta AID) by Morgan and Messenger (1973) to produce classification trees and CHAID (CHi AID) by Kass in the late 1970s.Breiman*et al.*(1984) developed CART (Classification and Regression Trees) which is a sophisticated program for fitting trees to data. Breiman, again in 1994, developed the bagging predictors which is a method of generating multiple versions of a predictor and using them to get an aggregated predictor. A good account of the CART methodology can be found in many recent books, say, Izenman (2008).An application of classification trees in the field of agriculture can be found in Sadhu *et al.* (2014).

Theconventional CART methodologyis outlined briefly. Following is a schematic representation of aconventional CART tree structure:



The unique starting point of,say, a classification tree, is called a root node and consists of the entire learning set $\mathcal{L}$ at the top of the tree. A node is a subset of the set of variables, and it can be terminal or nonterminal node. A nonterminal (or parent) node is a node that splits into two left and right child nodes (binary split). Such a binary split is determined by a condition on the value of a single variable, where the condition is either satisfied or not satisfied by the observed value of that variable. All observations in $\mathcal{L}$ that have reached a particular (parent) node and satisfy the condition for that variable drop down to one of the two *child* nodes; the remaining observations at that (parent) node that do not satisfy the condition drop down to the other *child* node. A node that does not split is called a terminal node and is assigned a class label. Each observation in $\mathcal{L}$ falls into one of the terminal nodes. When an observation of unknown class is “dropped down” the tree and ends up at a terminal node, it is assigned the class corresponding to the class label attached to that node. There may be more than one terminal node with the same class label. To produce a tree-structured model using recursive binary partitioning, CART determines the best split of the learning set $\mathcal{L}$ to start with and thereafter the best splits of its subsets on the basis of various issues such as identifying which variable should be used to create the split, and determining the precise rule for the split, determining when a node of the tree is a terminal one, and assigning a predicted class to each terminal node. The assignment of predicted classes to the terminal nodes is relatively simple, as is determining how to make the splits, whereas determining the right-sized tree is not so straightforward. After growing a fully expanded tree, a tree of optimum size is obtained. In a particular type of tree building called 'exhaustive search', at each stage of recursive partitioning, all of the allowable ways of splitting a subset of $\mathcal{L}$ are considered, and the one which leads to the greatest increase in node purity is chosen. This can be accomplished using what is called an "impurity function", which is nothing but a function of the proportion of the learning sample belonging to the possible classes of the response variable. To choose the best split over all variables, first the best split for a given variable has to be determined. To assess the goodness of a potential split, the value of the 'impurity function' such as Gini diversity index and the Entropy function can be calculated using the cases in the learning sample corresponding to the parent node, and subtract from this the weighted average of the impurity for the two *child* nodes, with the weights proportional to the number of cases of the learning sample corresponding to each of the *child* nodes, to get the decrease in the overall impurity that would result from the split. To select the way to split a subset of$\mathcal{L}$in the tree growing procedure, all allowable ways of splitting can be considered, and the one which will result in the greatest decrease in node impurity (or, in other words, greatest increase in the node purity) can be chosen.

In order to grow a tree, the starting point is the root node, which consists of the learning set$\mathcal{L}$. Using the "goodness of split" criterion for a single variable, the tree algorithm finds the best split at the root node for each of the variables. The best split s at the root node is then defined as the one that has the largest value of this goodness of split criterion over all single-variable best splits at that node. Next is to split each of the *child* nodes of the root node in the same way. The above computations are repeated for each of the *child* nodes except that this time only the observations in that specific *child* node are considered for the calculations rather than all the observations. When these splits are completed, the splitting is continued with the subsequent nodes. This sequential splitting procedure of building a tree layer-by-layer is hence called recursive partitioning. If every parent node splits in two *child* nodes, the result is called a binary tree. If the binary tree is grown until none of the nodes can be split any further, then the tree is said to be saturated. Usually, first a very large tree is grown, splitting subsets in the current partition of $\mathcal{L}$ even if a split does not lead to an appreciable decrease in impurity. Then a sequence of smaller trees can be created by "pruning" the initial large tree, where in the pruning process, splits that were made are removed and a tree having a fewer number of nodes is produced. The crucial part of creating a good tree-structured classification model is determining how complex the tree should be. If nodes continue to be created until no two distinct values of the independent variables for the cases in the learning sample belong to the same node, the tree may be over fitting the learning sample and not be a good classifier of future cases. On the other hand, if a tree has only a few terminal nodes, then it may be that it is not making enough use of information in the learning sample, and classification accuracy for future cases will suffer. Initially, in the tree-growing procedure, the predictive accuracy typically increases as more nodes are created and the partition gets finer. But it is usually seen that at some point the misclassification rate for future cases will start to get worse as the tree becomes more complex. In order to compare the prediction accuracy of various tree-structured models, there needs to be a way to estimate a given tree's misclassification rate for the future observations, a measure named 'resubstitution estimate' of the misclassification rate is obtained by using the tree to classify the members of the learning sample (that were used to create the tree), and observing the proportion that are misclassified. More often, a better estimate of a tree's misclassification rate can be obtained using an independent "test set", which is a collection of cases coming from the same population or distribution as the learning set. Like the learning set, for the test set the true class for each case is known in addition to the values for the predictor variables. The test set estimate of the misclassification rate is just the proportion of the test set cases that are misclassified when predicted classes are obtained using the tree created from the learning set. The learning set and the test set are both composed of cases for which the true class is known in addition to the values for the predictor variables. Generally, about one third of the available cases should be set aside to serve as a test set, and the rest of the cases should be used as learning set. But sometimes a smaller fraction, such as one tenth, is also used and then resorting to 10-fold cross validation. A specific way to create a useful sequence of different-sized trees is to use “minimum cost-complexity pruning”. In this process, a nested sequence of subtrees of the initial large tree is created by “weakest-link cutting”. With weakest-link cutting (pruning), all of the nodes that arise from a specific nonterminal node are pruned off (leaving that specific node itself as terminal node), and the specific node selected is the one for which the corresponding pruned nodes provide the smallest per node decrease in the resubstitution misclassification rate. If two or more choices for a cut in the pruning process would produce the same per node decrease in the resubstitution misclassification rate, then pruning off the largest number of nodes is preferred. The sequence of subtrees produced by the pruning procedure serves as the set of candidate subtrees for the model, and to obtain the classification tree, all that remains to be done is to select the one which will hopefully have the smallest misclassification rate for future observations. The selection of final tree is based on estimated misclassification rates, obtained using a test set or by cross validation.

1. Self Organizing Map (SOM)

In SOM, the training data set contains only input variables and no outputs. It is a 'self-organizing' system, which automatically adapts itself in such a way that similar input objects are associated with the topological close neurons in the ANN. The phrase 'topological close neurons' means that neurons that are physically located close to each other will react similar to similar inputs, while the neurons that are far apart in the lay-out of the ANN will react quite different to similar inputs. A practical treatment on SOFM based Kohonen networks can be found in Haykin (1996).

The principal goal is to transform an incoming input pattern of arbitrary dimension into a two dimensional discrete map. Neurons in the network are arranged in a two dimensional grid and there happens a competition among these neurons to represent the input pattern. The ‘winning’ neurons and the similar pattern neurons i.e. the neighboring neurons are placed in contiguous locations in output space. The neurons learn to pin-point the location of the neuron in the ANN that is most 'similar' to the input vector. Here, the phrase 'location of the most similar neuron' has to be taken in a very broad sense. It can mean the location of the closest neuron with the smallest or with the largest Euclidean distance to the input vector, or it can mean the neuron with the largest output in the entire network for this particular input vector etc. In other words, in the Kohonen network, a ‘rule’ deciding which of all neurons will be selected after the input vector enters the ANN is mandatory. During the training in the Kohonen’s ANN, the multidimensional neurons self-organise themselves in the two-dimensional plane in such a way that the objects from the multidimensional measurement space are mapped into the plane of neurons with respect to some internal property correlated to the m-dimensional measurement space of objects.

Bullinaria (2004) has explained the above discussion in the following manner. Neurons are placed at the nodes of a lattice that is usually two-dimensional and undergo the following three steps:

(i) Competition

Neurons become selectively tuned to various input patterns (stimuli). Such “winning” neurons become ordered w.r. to each other in such a way that a meaningful coordinate system for different input features is created over the lattice. The competitive learning is characterized by formation of a topographic map of the inputs in which spatial locations of the neurons in the lattice are indicative of intrinsic features contained in the inputs, hence the name SOFM.

(ii) Cooperation

The winning neurons determines the spatial location of a topographic neighbourhood of excited neurons, thereby providing the basis for cooperation

(iii) Adaptation

The excited neurons adapts their individual values of its functional form in relation to the input pattern through suitable adjustments applied to their synaptic weights. Thus the response of the winning neuron to the subsequent application of a similar input pattern is enhanced

The correction of weights is carried out after the input of each input object in the following four steps:

(i) the neuron with the most ’distinguished’ response of all (in a sense explained above) is selected and named the ’central’ or the ’most excited’ neuron

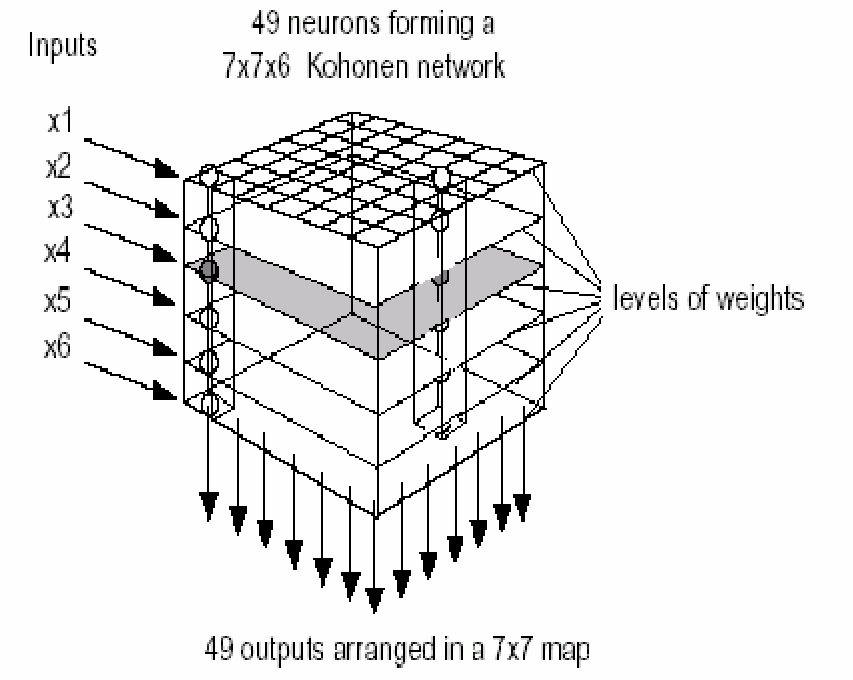
(ii) the maximal neighbourhood around this central neuron is determined.

(iii) the ‘correction factor’ is calculated for each neighbourhood ring separately (the correction changes according to the distance and time of training)

(iv) the ‘weights’ in neurons of each neighbourhood are corrected according to a pre-specified equation

The most important difference is that the neurons in the error back propagation learning (in that of the most famous multi-layer perceptron type of architectured neural network) tries to yield quantitatively an answer as close as possible to the target, while in the Kohonen approach the neurons learn to pin-point the location of the neuron in the ANN that is most ’similar’ to the input vector.

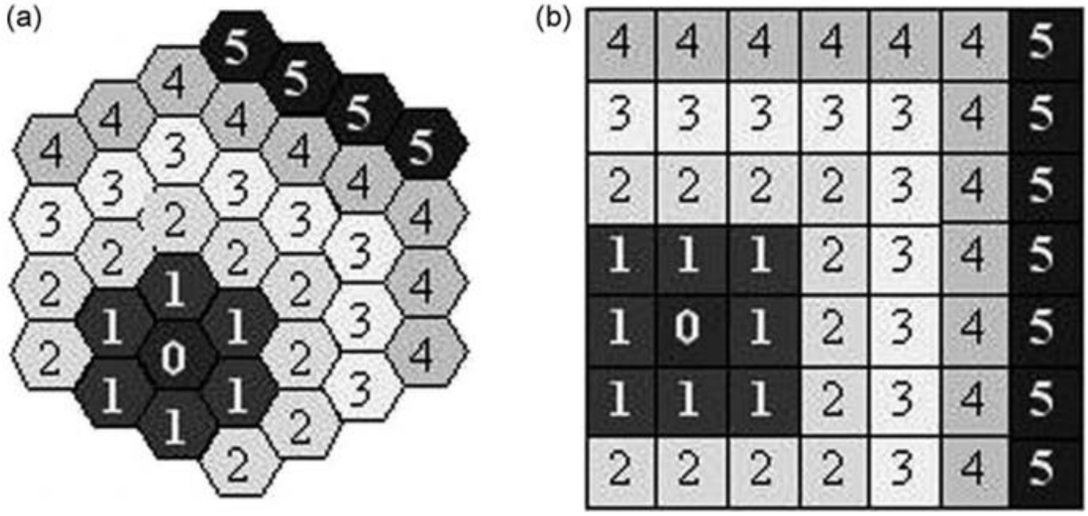
In order to make things clear, let us consider the following figure wherein there are six input variables along with a two-dimensional map of order 7x7. The neurons are in the columns associating the input variables with the (i, j)-th neuron in the output map, with weights at various levels corresponding to the inputs. That is, because the Kohonen ANN has only one layer of neurons, the specific input variable, let us say the i-th variable x_i is always received in all neurons of the ANN by the weight placed in the i-th position. If the neurons are presented as columns of weights then all i-th weights in all neurons can be regarded as the weights of the i-th level (Zupan, 1994).



Because the Kohonen ANN has only one layer of neurons the specific input variable, let us say the i-th variable, x_i, is always received in all neurons of the ANN, by the weight placed at the i-th position. If the neurons are presented as columns of weights then all i-th weights in all neurons can be regarded as the weights of the i-th level. This is especially important because the neurons are usually ordered in a two-dimensional formation.

Thus the main goal of Kohonen is to perform a non-linear mapping from an high-dimensional variable space to a low-dimensional (usually 2D) target space so that the distance and proximity relations between the samples or, in a single word, the topology, are preserved. The target space used in Kohonen mapping is a two-dimensional array of neurons fully connected to the input layer, onto which the samples are mapped. Introducing the preservation of topology, results in specifying for each node in the Kohonen layer, a defined number of neurons as nearest neighbors, second-nearest neighbors and so on.

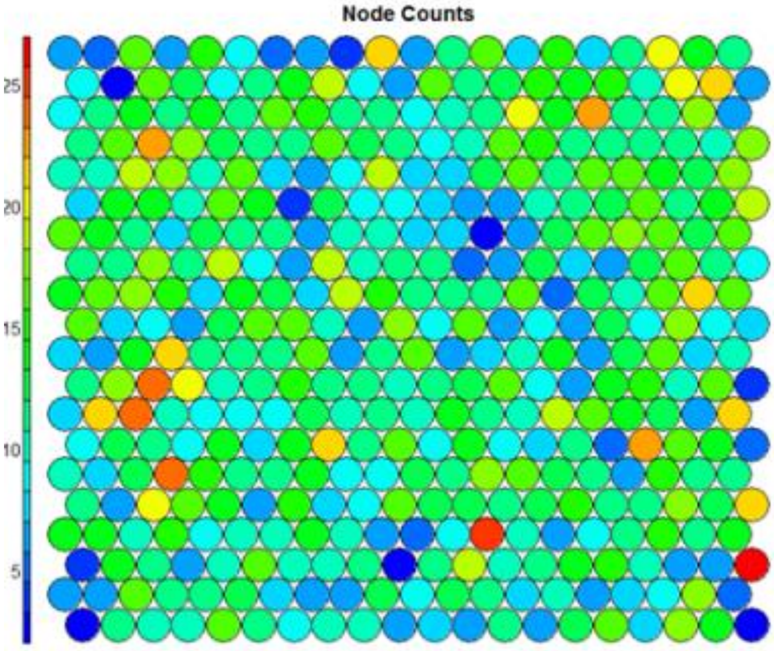
The layout of neurons in the Kohonen ANN is an important feature to be discussed (Marini *et al.*, 2007). The neighborhood of a neuron is usually considered to be hexagonal [see (a) in figure below] or square [see (b) in figure below] which means that each neuron has eight or six nearest neighbors, respectively.



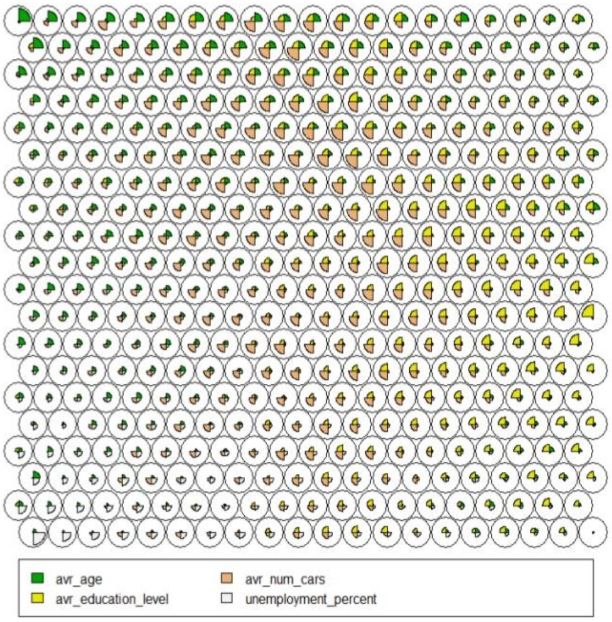
4	4	4	4	4	4	5
3	3	3	3	3	4	5
2	2	2	2	3	4	5
1	1	1	2	3	4	5
1	0	1	2	3	4	5
1	1	1	2	3	4	5
2	2	2	2	3	4	5

The main issue in Kohonen learning is that similar input vectors excite neurons which are very close in the 2D layer. From an algorithmic point of view, Kohonen mapping implements competitive learning, i.e. only one neuron in the 2D layer is selected after each input is presented to the network (winner takes-all). The winning neuron c is selected as the one having the weight vector most similar to the input pattern. After the winning neuron in the Kohonen layer is selected, the weights of each other neuron in the Kohonen layer are updated on the basis of the difference between their old value and the values of the input vector; this correction is scaled according to the topological distance from the winner.

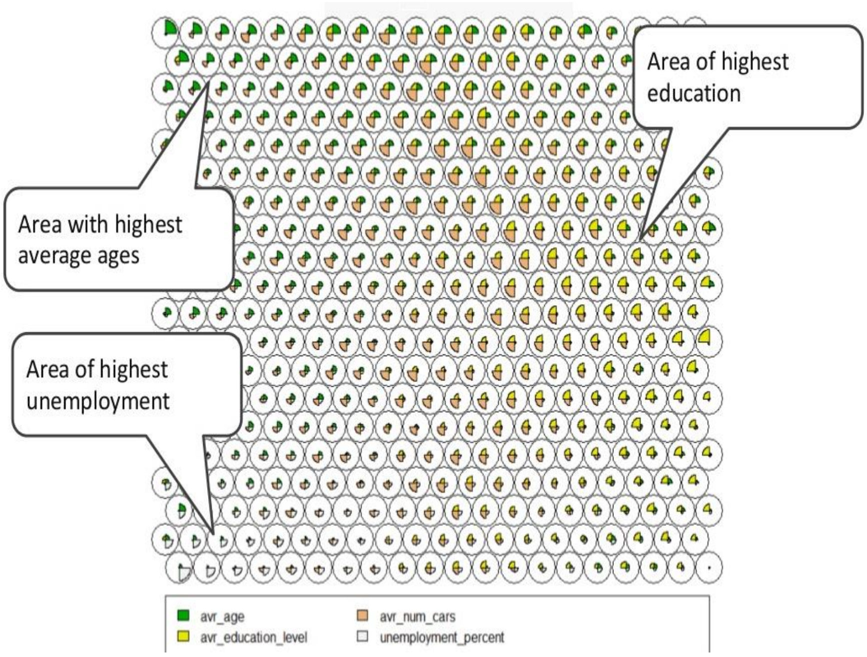
Lynn (2014) have extensively discussed about the theKohonen package available in the open source and freely available R software. This Kohonen R package allows us to visualise the count of how many samples are mapped to each node on the map. This metric can be used as a measure of map quality – ideally the sample distribution is relatively uniform. Large values in some map areas suggests that a larger map would be benificial. Empty nodes indicate that the map size is too big for the number of samples. He suggest that one should aim for at least 5-10 samples per node when choosing map size. One such output where node counts are visualized is given subsequently.



The node weight vectors, or “codes”, are made up of normalised values of the original variables used to generate the SOM. Each node’s weight vector is representative / similar of the samples mapped to that node. By visualising the weight vectors across the map, we can see patterns in the distribution of samples and variables. Such a visualisation of the weight vectors can be done using a “fan diagram”, where individual fan representations of the magnitude of each variable in the weight vector is shown for each node. One such fan diagram is given below.



- Fan diagram shows distribution of variables across map.
- Can see patterns by examining dominant colours etc.
- This type of representation is useful for SOMs when the number of variables is less than ~ 5
- Good to get a grasp of general patterns in SOM



References

Breiman, L., Freidman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*. Wadsworth, Belmont CA.

Bullinaria, J.A. (2004). https://www.cs.bham.ac.uk/~jxb/NN/l16.pdf, accessed on 29th February, 2020.

Haykin, S. (1996). *Neural networks: A comprehensive foundation*, Pearson Education, Asia.

Izenman, A.J. (2008). *Modern multivariate statistical techniques: Regression, classification and manifold learning*. Springer, New York.

Lynn, S. (2014). https://www.slideshare.net/shanelynn/2014-0117-dublin-r-selforganising-maps-for-customer-segmentation-shane-lynn, accessed on 23rd December, 2019.

Marini, F., Magri, A. L., Bucci, R. and Magri, A.D. (2007). Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils, *AnalyticaChimicaActa*, **599**, 232–240.

Morgan, J.N. and Messenger, R.C. (1973). THAID: a sequential search program for the analysis of nominal scale dependent variables. Institute for Social Research, University of Michigan, Ann Arbor, MI.

Sadhu, S.K., Ramasubramanian, V., Rai, A. and Kumar, A. (2014). Decision tree based models for classification in agricultural ergonomics, *Statistics and Applications*, **12(1&2)**, 21-33.

Zupan, J. (1994). Introduction to Artificial Neural Network (ANN) methods: What they are and how to use them, *Acta Chimica Slovenica* ,**41 (3)**, 327-352.

CLUSTER ANALYSIS USING R

Alka Arora
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
Alka.Arora@icar.gov.in

1. Introduction

Clustering algorithms maps the data items into clusters, where clusters are natural grouping of data items based on similarity methods. Unlike classification and prediction which analyzes class-label data objects, clustering analyzes data objects without class-labels and tries to generate such labels. Clustering has many applications. In business/ marketing, clustering can help in identifying different customer groups and appropriate marketing campaign can be carried out targeting different groups. In agriculture, it can be used to derive plant and animal taxonomies, characterization of diseases and varieties, in bioinformatics- categorization of genes with similar functionally. Further it can be used to group similar documents on the web for faster discovery of content. It can be used to group geographical locations based on crime, amenities, weather etc. As data mining function, cluster analysis is used to gain insight into distribution of data, to observe the characteristics of each cluster and to focus on a particular set of clusters for further analysis.

1. Similarity Measures

Similarity is fundamental to majority of clustering algorithms. *Similarity is quantity that reflects the strength of relationship between two objects or two features.* This quantity is usually having range of either -1 to +1 or normalized into 0 to 1. If the similarity between feature *i* and feature *j* is denoted by s_{ij} , we can measure this quantity in several ways depending on the scale of measurement (or data type) that we have. Dissimilarity is opposite to similarity. There are many types of distance and similarity measures.

Similarity and dissimilarity can be measured for two objects based on several features/ variables. After the distance or similarity of each variable is determined, we can aggregate all features/ variables together into single Similarity (or dissimilarity) index between the two objects.

2.1 Distance for binary variables

We often face variables that only binary value such as Yes and No, or Agree and Disagree, True and False, Success and Failure, 0 and 1, Absence or Present, Positive and Negative, etc. Similarity of dissimilarity (distance) of two objects that represented by binary variables can be measured in term of number of occurrence (frequency) of positive and negative in each object.

For example:

Feature of Fruit	**Sphere shape**	**Sweet**	**Sour**	**Crunchy**
Object i =Apple	Yes	Yes	Yes	Yes
Object j =Banana	No	Yes	No	No

The coordinate of Apple is (1,1,1,1) and coordinate of Banana is (0,1,0,0). Because each object is represented by 4 variables, we say that these objects have 4 dimensions.

Let p = number of variables that positive for both objects .

q = number of variables that positive for the i th objects and negative for the j th object

r= number of variables that negative for the i th objects and positive for the j th object

s= number of variables that negative for both objects

t= p+q+r+s = total number of variables.

		Object j	
		Yes	**No**
object i	**Yes**	p	q
	No	r	s

For our example above, we have measured Apple and Banana have $p=1$, $q=3$ and $r=0$, $s=0$. Thus, $t= p+q+r+s=4$.

The most common use of binary dissimilarity (distance) is

Simple Matching distance $d_{ij} = \dfrac{q+r}{t}$

Jaccard's distance $d_{ij} = \dfrac{q+r}{p+q+r}$

Hamming distance $d_{ij} = q+r$

Example: Simple matching distance between Apple and Banana is 3/4.

Jaccard's distance between Apple and Banana is 3/4.

Hamming distance between Apple and Banana is 3.

2.2 Distance for quantitative variables

Variable which have quantitative values.

		Features k		
	cost	time	weight	incentive
Object A	0	3	4	5
Object B	7	6	3	-1

We can represent the two objects as points in 4 dimension. Point A has coordinate (0, 3, 4, 5) and point B has coordinate (7, 6, 3, -1). Dissimilarity (or similarity) between the two objects are based on these coordinates.

Euclidean Distance: Euclidean Distance is the most common use of distance. In most cases when people said about distance, they will refer to Euclidean distance. Euclidean distance or simply 'distance' examines the *root of square differences* between coordinates of a pair of objects.

Formula $$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$$

$$\begin{aligned} d_{BA} &= \sqrt{(0-7)^2 + (3-6)^2 + (4-3)^2 + (5+1)^2} \\ &= \sqrt{49+9+1+36} = 9.747 \end{aligned}$$

Euclidean distance is a special case of Minkowski distance with $\lambda = 2$

City block (Manhattan) distance : It is also known as *Manhattan* distance, *boxcar* distance, *absolute value* distance. It examines the *absolute differences* between coordinates of a pair of objects. City block distance is a special case of Minkowski distance with $\lambda = 1$

Formula: $$d_{ij} = \sum_{k=1}^{n} \left|x_{ik} - x_{jk}\right|$$

The City Block Distance between point A and B is

$$\begin{aligned} d_{BA} &= |0-7| + |3-6| + |4-3| + |5+1| \\ &= 7+3+1+6 = 17 \end{aligned}$$

Chebyshev Distance : Chebyshev distance is also called Maximum value distance. It examines the *absolute magnitude of the differences* between coordinates of a pair of objects. This distance can be used for both ordinal and quantitative variables.

Formula $d_{ij} = \max_{k} \left| x_{ik} - x_{jk} \right|$ and B is

$$d_{BA} = \max\{|0-7|,|3-6|,|4-3|,|5+1|\}$$
$$= \max\{7,3,1,6\} = 7$$

Minkowski Distance: This is the generalized metric distance. When $\lambda = 1$ it becomes city block distance and when $\lambda = 2$, it becomes Euclidean distance. Chebyshev distance is a special case of Minkowski distance with $\lambda = \infty$ (taking a limit). This distance can be used for both ordinal and quantitative variables.

Formula $d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^{\lambda}}$

2. Clustering Algorithms

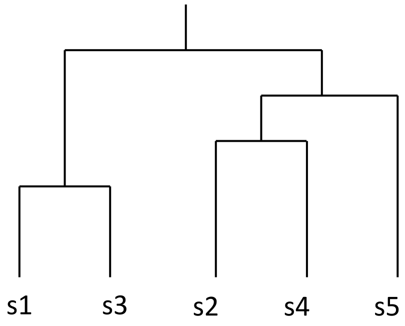
There are many clustering algorithms available in literature, choice of appropriate algorithm depends on the data type and desired results. We will be focusing on commonly used clustering algorithms.

3.1 Hierarchical Algorithms

A hierarchical method creates a hierarchical decomposition of data objects in the form of tree like diagram which is called a dendogram. There are two approaches to building a cluster hierarchy.

Agglomerative approach also called bottom up approach starts with each object forming a separate group and successively merges the objects close to one another, until all the groups are merged into one.

Divisive approach also called top-down approach starts with all the objects in same cluster, until each object is in one cluster.



Process flow of agglomerative hierarchical clustering method is given below:

- Convert object features to distance matrix.

- Set each object as a cluster (thus if we have 6 objects, we will have 6 clusters in the beginning)
- Iterate until number of cluster is 1
    1. Merge two closest clusters
    2. Update distance matrix

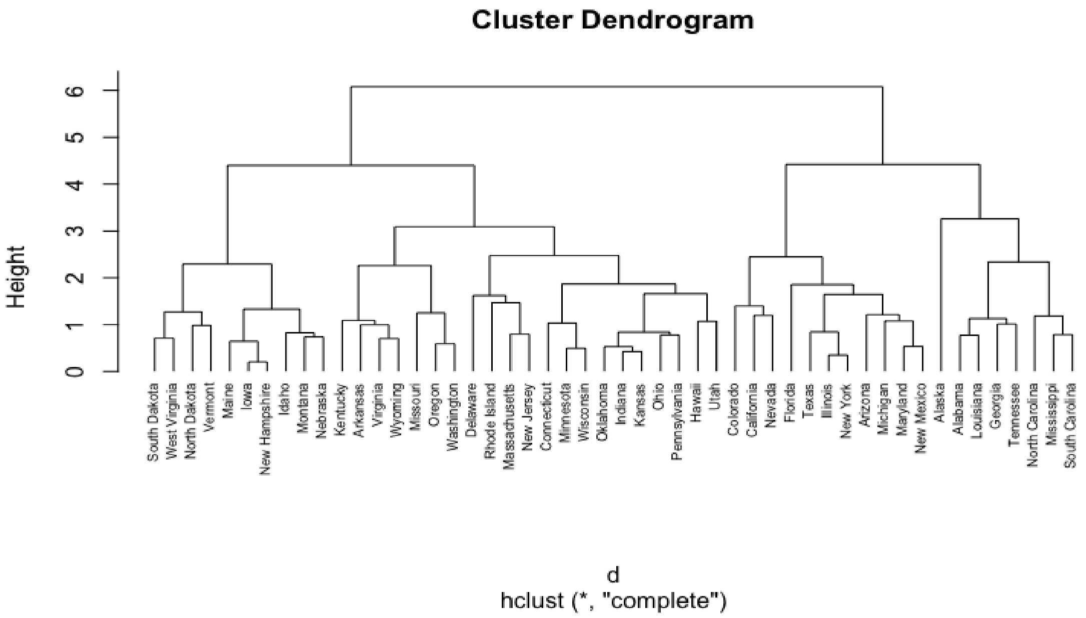
First distance matrix is computed using any valid distance measure between pairs of objects. The choice of which clusters to merge is determined by a linkage criterion, which is a function of the pairwise distances between observations. Commonly used linkage criteria are mentioned below:

- Complete Linkage: The maximum distance between elements of each cluster $\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}.$
- Single Linkage: The minimum distance between elements of each cluster $\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}.$
- Average Linkage /UPGMA: The mean distance between elements of each cluster
$$\frac{1}{|\mathcal{A}|\cdot|\mathcal{B}|}\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} d(x,y).$$

3.1.1 Hierarchical Clustering (HC) using R:

In R, function hclust() performs hierarchical clustering. First the dissimilarity values are computed with dist function. Feed these values into hclust and specify the agglomeration method to be used (i.e. “complete”, “average”, “single”, “ward.D”). Then plot the dendrogram.

```
# Dissimilarity matrix
d <- dist(df, method = "euclidean")
# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )
# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)
```



Alternatively, you can use the agnes function. These functions behave very similarly; however, with the agnes function you can also get the agglomerative coefficient, which measures the amount of clustering structure found (values closer to 1 suggest strong clustering structure).

```
# Compute with agnes
hc2 <- agnes(df, method = "complete")
# Agglomerative coefficient
hc2$ac
## [1] 0.8531583
```

This allows us to find certain hierarchical clustering methods that can identify stronger clustering structures. Here we see that Ward's method identifies the strongest clustering structure of the four methods assessed.

```
# methods to assess
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")
# function to compute coefficient
ac <- function(x) {
  agnes(df, method = x)$ac
}
map_dbl(m, ac)
##   average    single  complete      ward
## 0.7379371 0.6276128 0.8531583 0.9346210
hc3 <- agnes(df, method = "ward")
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of agnes")
```

Similarly, HC can be performed using function diana. diana works similar to agnes; however, there is no method to provide.

```
# compute divisive hierarchical clustering
hc4 <- diana(df)
# Divise coefficient; amount of clustering structure found
hc4$dc
## [1] 0.8514345
# plot dendrogram
pltree(hc4, cex = 0.6, hang = -1, main = "Dendrogram of diana")
```

Working with Dendrograms

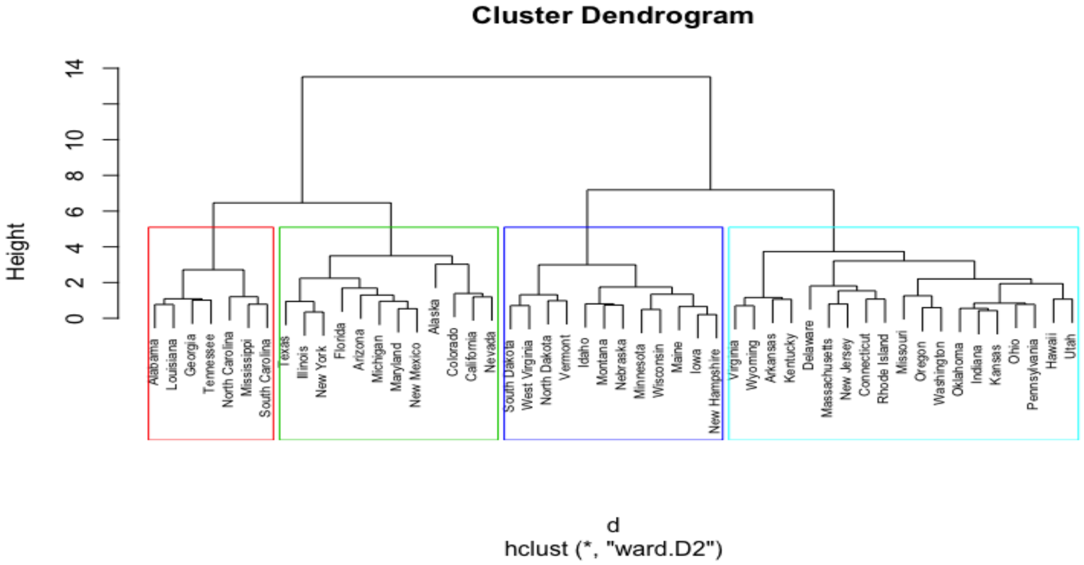
In the dendrogram displayed above, each leaf corresponds to one observation. As we move up the tree, observations that are similar to each other are combined into branches, which are themselves fused at a higher height. The height of the fusion, provided on the vertical axis, indicates the (dis)similarity between two observations.

The height of the cut to the dendrogram controls the number of clusters obtained. we can cut the dendrogram with cutree ():

```
# Ward's method
hc5 <- hclust(d, method = "ward.D2" )
# Cut tree into 4 groups
sub_grp <- cutree(hc5, k = 4)
# Number of members in each cluster
table(sub_grp)
## sub_grp
##  1  2  3  4
##  7 12 19 12
```

It’s also possible to draw the dendrogram with a border around the 4 clusters. The argument border is used to specify the border colors for the rectangles:

```
plot(hc5, cex = 0.6)
rect.hclust(hc5, k = 4, border = 2:5)
```



3.2 Partitional Algorithms

It basically involves segmenting data objects into k partitions, optimizing some criteria, over t iterations. These methods are popularly known as iterative relocation methods.

3.2.1 K-means Algorithm

K-means is the most popularly used algorithm in this category. It randomly selects k objects as cluster mean or center. It works towards optimizing square error criteria function, defined as:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - m_i \right\|^2 \text{, where } m_i \text{ is the mean of cluster } C_i \text{.}$$

Main steps of k-means algorithm are:

1) Assign initial means m_i

2) Assign each data object x to the cluster C_i for the closest mean

3) Compute new mean for each cluster

4)Iterate until criteria function converges, that is, there are no more new assignments.

The k-means algorithm is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data.

3.2.2 k-means clustering in R :

We can compute k-means in R with the kmeans function. In this example, data is grouped into two clusters (centers = 2). The kmeans function also has an nstart option that attempts multiple initial configurations and reports on the best one. For example, adding nstart = 25 will generate 25 initial configurations.

```
k2 <- kmeans(df, centers = 2, nstart = 25)
str(k2)
## List of 9
##  $ cluster     : Named int [1:50] 1 1 1 2 1 1 2 2 1 1 ...
##   ..- attr(*, "names")= chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ centers     : num [1:2, 1:4] 1.005 -0.67 1.014 -0.676 0.198 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ totss       : num 196
##  $ withinss    : num [1:2] 46.7 56.1
##  $ tot.withinss: num 103
##  $ betweenss   : num 93.1
##  $ size        : int [1:2] 20 30
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

The output of kmeans is a list with several bits of information. The most important being:

cluster: A vector of integers (from 1:k) indicating the cluster to which each point is allocated.

centers: A matrix of cluster centers.

totss: The total sum of squares.

withinss: Vector of within-cluster sum of squares, one component per cluster.

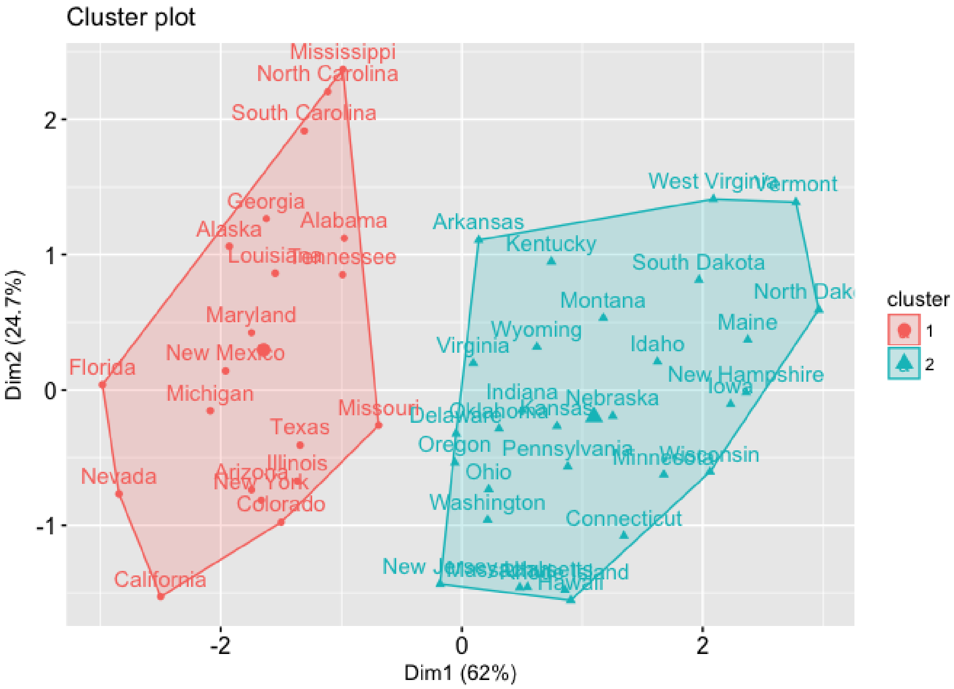
tot.withinss: Total within-cluster sum of squares, i.e. sum(withinss).

betweenss: The between-cluster sum of squares, i.e. $totss-tot.withinss$.

size: The number of points in each cluster.

We can also view the results by using fviz_cluster. This provides a nice illustration of the clusters. If there are more than two dimensions (variables). fviz_cluster will perform principal component analysis (PCA) and plot the data points according to the first two principal components that explain the majority of the variance.

```
fviz_cluster(k2, data = df)
```



References

A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, *ACM Computing Surveys*, 31(3): 264-323, 1999, ISSN: 0360-0300.

B. Mirkin, *Clustering for Data Mining: Data Recovery Approach*, Chapman & Hall/CRC, 2005.

I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*, Morgan Kaufmann publishers, 1999.

J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann Publisher, 2006, ISBN 1-55860-901-6

L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, 1990.

http://en.wikipedia.org/wiki/K-means_clustering

http://people.revoledu.com/kardi/tutorial

R. Xu, D. Wunsch, Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, May 2005

S. Mitra, T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics*, John Wiley & Sons, 2004, ISBN 9812-53-063-0.

https://uc-r.github.io/hc_clustering

https://uc-r.github.io/kmeans_clustering

https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/hclust

ARCH/ GARCH FAMILY OF NON-LINEAR MODELS

Ranjit Kumar Paul
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
ranjitstat@gmail.com, ranjit.paul@icar.gov.in

Introduction

The most widely used technique for analysis of time-series data is; undoubtedly, the Box Jenkins' Autoregressive integrated moving average (ARIMA) methodology (Box *et al.*, 2007). However, it is based on some crucial assumptions, like linearity and homoscedastic prediction error variances. In reality, underlying relationships among variables are highly complex and cannot be described satisfactorily through a linear modelling approach. There are many features, like existence of threshold value, which can be described only through a nonlinear approach. During the last few decades a new area of "Nonlinear time-series modelling" is fast coming up. Here, there are basically two approaches, viz. Parametric or Nonparametric. Evidently, if in a particular situation, we are quite sure about the functional form, we should use the former; otherwise the latter may be employed.

When dealing with nonlinearities, Campbell *et al.* (1997) made the distinction between:

- *Linear Time-Series:* shocks are assumed to be uncorrelated but not necessarily identically and independently distributed (*iid*).
- *Nonlinear Time-Series:* shocks are assumed to be *iid*, but there is a nonlinear function relating the observed time-series $\{X_t\}_{t=0}^{\alpha}$ and the underlying shocks, $\{\varepsilon_t\}_{t=0}^{\alpha}$.

A nonlinear process is described as

$$X_t = g(\varepsilon_{t-1}, \varepsilon_{t-2}, ...) + \varepsilon_t h(\varepsilon_{t-1}, \varepsilon_{t-2}, ...).\ E[X_t / \psi_{t-1}] = g(\varepsilon_{t-1}, \varepsilon_{t-2}, ...)$$

$$Var[X_t / \psi_{t-1}] = E\left[\{(X_t - E(X_t)) / \psi_{t-1}\}^2\right] = \{h(\varepsilon_{t-1}, \varepsilon_{t-2}, ...) / \psi_{t-1}\}^2$$

where function $g(\cdot)$ corresponds to conditional mean of X_t, and function $h(\cdot)$ is coefficient of proportionality between innovation in X_t and shock ε_t. The general form above leads to a natural division in Nonlinear time-series literature in two branches:

- Models Nonlinear in Mean: $g(\cdot)$ is nonlinear;
- Models Nonlinear in Variance: $h(\cdot)$ is nonlinear.

The most promising parametric nonlinear time series models like ARCH and GARCH models are described below.

Autoregressive Conditional Heteroscedastic (ARCH) Model

The most promising parametric nonlinear time-series model has been the Autoregressive conditional heteroscedastic (ARCH) model, which was introduced by Engle (1982). It allows the conditional variance to change over time as a function of squared past errors leaving the unconditional variance constant. The presence of ARCH type effects in financial and macro-economic time-series is a well established fact. The combination of ARCH specification for conditional variance and the Autoregressive (AR) specification for conditional mean has many appealing features, including a better specification of the forecast error variance. Ghosh and Prajneshu (2003) employed AR(*p*)-ARCH(*q*)-in-Mean model for carrying out modelling and forecasting of volatile monthly onion price data. The AR-ARCH model has also been used as the basic "building blocks" for Markov switching and mixture models (See e.g. Lanne and Saikkonen 2003 and Wong and Li 2001).

The ARCH (*q*) model for series $\{\varepsilon_t\}$ is defined by specifying the conditional distribution of ε_t given information available up to time $t-1$. Let ψ_{t-1} denote this information. It consists of the knowledge of all available values of the series, and anything which can be computed from these values, *e.g.* innovations, and squared observations. In principle, it may even include knowledge of the values of other related time-series, and anything else which might be useful for forecasting and is available by time $t-1$.

We say that the process $\{\varepsilon_t\}$ is ARCH (*q*), if the conditional distribution of $\{\varepsilon_t\}$ given available information ψ_{t-1} is

$$\varepsilon_t \mid \psi_{t-1} \sim N(0, h_t) \text{ and } h_t = a_0 + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 \qquad (1)$$

where $a_0 > 0$, $a_i \geq 0$ for all *i* and $\sum_{i=1}^{q} a_i < 1$

Properties of the ARCH model (Tsay,2005)

To study the properties of ARCH model, consider the simple ARCH (*1*) model. The conditional variance equation of the this model is defined as

$$\varepsilon_t = \eta_t h_t^{1/2},$$

η_t is white noise and conditional variance h_t satisfies

$$h_t = a_0 + a_1\varepsilon_{t-1}^2$$

where $a_0 > 0$, $a_1 \geq 0$.The important properties of ARCH models are mentioned below:

(i) The unconditional mean of ε_t remains zero because,

$$E(\varepsilon_t) = E[E(\varepsilon_t | \Psi_{t-1})] = E[\sqrt{h_t}E(\varepsilon_t)] = 0$$

(ii) The unconditional variance of ε_t can be defined as

$$\operatorname{var}(\varepsilon_t) = E(\varepsilon_t^2) = E[E(\varepsilon_t^2 \mid \psi_{t-1})] = E(a_0 + a_1\varepsilon_{t-1}^2) = a_0 + a_1E(\varepsilon_{t-1}^2).$$

If ε_t is a stationary process with $E(\varepsilon_t) = 0$, $\operatorname{var}(\varepsilon_t) = \operatorname{var}(\varepsilon_{t-1}) = E(\varepsilon_{t-1}^2)$. Therefore, $\operatorname{var}(\varepsilon_t) = a_0 + a_1 \operatorname{var}(\varepsilon_t)$ and so $var(\varepsilon_t) = a_0 / (1 - a_1)$. Since variance of ε_t must be positive, therefore $0 \leq a_1 < 1$.

(iii) In some applications, higher order moments of ε_t are required to exist and, hence, a_1 must satisfy some additional constraints. For instance, to study its tail behavior, we require that the fourth moment of ε_t is finite.

Heavy tails are a common aspect of financial data, and hence the ARCH models are very popular in this field. Besides that, Bera and Higgins (1993) mention the following reasons for the ARCH success:

- ARCH models are simple and easy to handle.
- ARCH models take care of clustered errors.
- ARCH models take care of nonlinearities.
- ARCH models take care of changes in the econometrician's ability to forecast.

Forecasting

Forecasts of the ARCH model can be obtained recursively as those of an AR model. Consider an ARCH (q) model. At the forecast origin t, the one-step ahead forecast is

$$h_t(1) = a_0 + a_1\varepsilon_t^2 + \ldots + a_q\varepsilon_{t+1-q}^2 \quad (2)$$

The two-step ahead forecast is $h_t(2) = a_0 + a_1h_t(1) + a_2\varepsilon_t^2 + \ldots + a_q\varepsilon_{t+2-q}^2$, and l- step ahead forecast is $h_t(l) = a_0 + \sum_{i=1}^{q} a_i h_t(l-i)$ where $h_t(l-i) = \varepsilon_{t+l-i}^2$ if $l - i \leq 0$.

However, ARCH model has some drawbacks. Firstly, when the order of ARCH model is very large, estimation of a large number of parameters is required. Secondly, conditional variance of ARCH(q) model has the property that unconditional autocorrelation function (Acf) of squared residuals; if it exists, decays very rapidly compared to what is typically observed, unless maximum lag q is large. To overcome these difficulties, Bollerslev (1986) proposed the Generalized ARCH (GARCH) model in which conditional variance is also a linear function of its own lags. This model is also a weighted average of past squared residuals, but it has declining weights that never go completely to zero. It gives parsimonious models that are easy to estimate and, even in its simplest form, has proven surprisingly successful in predicting conditional variances. Angelidis *et al*. (2004) evaluated the performance of GARCH models in modelling the daily Value-at-Risk (VaR) of perfectly distributed portfolios in five stock indices, using a number of distributional assumptions and sample sizes. Paul *et al*. (2009, 2014) applied GARCH model for forecasting of spices export and wheat yield respectively.

Generalized ARCH(GARCH) Model

To overcome the weaknesses of ARCH model, Bollerslev (1986) and Taylor (1986) proposed the Generalized ARCH (GARCH) model independently of each other, in which conditional variance is also a linear function of its own lags and has the following form

$$\varepsilon_t = \xi_t h_t^{1/2} \quad h_t = a_0 + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j\, h_{t-j} \qquad (3)$$

where ξ_t ~ IID(*0,1*). A sufficient condition for the conditional variance to be positive is

$$a_0 > 0,\ a_i \geq 0,\ i = 1,2,...,q.\ \ b_j \geq 0,\ \ j = 1,2,...,p$$

The GARCH (p, q) process is weakly stationary if and only if $\sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1$.

The conditional variance defined by (3) has the property that the unconditional autocorrelation function of ε_t^2 ; if it exists, can decay slowly. For the ARCH family, the decay rate is too rapid compared to what is typically observed in financial time-series, unless the maximum lag q is long. As (3) is a more parsimonious model of the conditional variance than a high-order ARCH model, most users prefer it to the simpler ARCH alternative.

The most popular GARCH model in applications is the GARCH(*1,1*) model. To express GARCH model in terms of ARMA model, denote $\eta_t = \varepsilon_t^2 - h_t$. Then from equation (3)

$$\varepsilon_t^2 = a_0 + \sum_{i=1}^{Max(p,q)} (a_i + b_i)\varepsilon_{t-i}^2 + \eta_t + \sum_{j=1}^{p} b_j\, \eta_{t-j} \tag{4}$$

Thus a GARCH model can be regarded as an extension of the ARMA approach to squared series { ε_t^2 }. Using the unconditional mean of an ARMA model, we have

$$\mathrm{E}(\varepsilon_t^2) = \frac{a_0}{1 - \sum_{i=1}^{Max(p,q)} (a_i + b_i)} \tag{5}$$

provided that the denominator of the prior fraction is positive.

Properties of GARCH model

The most widely used GARCH specification asserts that the best predictor of the variance in the next period is a weighted average of the long-run average variance, the variance predicted for this period, and the new information in this period that is captured by the most recent squared residual. Such an updating rule is a simple description of adaptive or learning behavior and can be thought of as Bayesian updating.

The properties of GARCH models can easily be studied by focusing on the simplest GARCH(*1,1*) model with

$$\varepsilon_t = \xi_t h_t^{1/2} \quad h_t = a_0 + a_1\, \varepsilon_{t-1}^2 + b_1\, h_{t-1}, \tag{6}$$

where ξ_t ~ IID(*0,1*) and $0 \le a_1, b_1 \le 1, (a_1 + b_1) < 1$.

The GARCH model that has been described is typically called the GARCH(*1,1*) model. The (*1,1*) in parentheses is a standard notation in which the first number refers to how many autoregressive lags, or ARCH terms, appear in the equation, while the second number refers to how many moving average lags are specified, which here is often called the number of GARCH terms. Sometimes models with more than one lag are needed to find good variance forecasts.

First a large ε_{t-1}^2 or h_{t-1} gives rise to a large h_t. This means that a large ε_{t-1}^2 tends to followed by another large ε_t^2, generating again the well known behavior of volatility clustering in financial time-series.

Second it can be seen that if $1-2a_1^2-(a_1+b_1)^2>0$, then

$$\frac{E(\varepsilon_t^4)}{[E(\varepsilon_t)]^2}=\frac{3\left[1-(a_1+b_1)^2\right]}{1-(a_1+b_1)^2-2a_1^2}>3$$

Consequently, similar to ARCH models, the tail distribution of a GARCH(*1,1*) process is heavier than that of a normal distribution.

Third, the model provides a simple parametric function that can be used to describe the volatility evolution.

Forecasting volatility by GARCH model

Forecasts of a GARCH model can be obtained using methods similar to those of an ARMA model. Although this model is directly set up to forecast for just one period, it turns out that based on the one-period forecast, a two-period forecast can be made. Ultimately, by repeating this step, long-horizon forecasts can be constructed. For the GARCH(*1,1*), the two-step forecast is a little closer to the long-run average variance than is the one-step forecast, and, ultimately, the distant-horizon forecast is the same for all time periods as long as $(a_1+b_1)<1$. This is just the unconditional variance. Thus, the GARCH models are mean reverting and conditionally heteroscedastic, but have a constant unconditional variance.

Consider the GARCH(*1,1*) model in (6) and assume that the forecast origin is *t*, the one-step ahead forecast is $h_t(1)=a_0+a_1\varepsilon_t^2+b_1h_t$

For multi-step ahead forecasts, use $\varepsilon_t^2=\xi_t^2h_t$ and rewrite the volatility equation in (6) as

$$h_{t+1}=a_0+(a_1+b_1)h_t+a_1h_t\left(\varepsilon_t^2-1\right)$$

For two-step ahead forecasts $h_{t+2}=a_0+(a_1+b_1)h_{t+1}+a_1h_{t+1}\left(\varepsilon_{t+1}^2-1\right)$ Since

$$E\left(\left(\varepsilon_{t+1}^2-1\right)/\psi_t\right)=0,$$

The two-step ahead volatility forecast at the forecast origin *t* satisfies the equation

$$h_t(2)=a_0+(a_1+b_1)h_t(1)$$

In general we have $h_t(l)=a_0+(a_1+b_1)h_t(l-1)$, *l>1*

This result is exactly the same as that of an ARMA(*1,1*) model. By repeated substitution in the equation (7), the one- step ahead forecast can be written as

$$h_t(l)=\frac{a_0\left[1-(a_1+b_1)^{l-1}\right]}{1-a_1-b_1}+(a_1+b_1)^{l-1}h_t(1)$$

Therefore, $h_t(l) \rightarrow \dfrac{a_0}{1-a_1-b_1}, as\, l \rightarrow \infty$,provided that $a_1 + b_1$<1.

Consequently, the multi-step ahead volatility forecast of a GARCH(*1,1*) model converge to the unconditional variance of ε_t as the forecast horizon increases to infinity provided that Var(ε_t)exists.

In order to estimate the parameters of GARCH model, three types of estimator are available in literature. They are the conditional maximum likelihood estimator, Whitle's estimator and the least absolute deviation estimator.

Conditional maximum likelihood estimator

Similar to the estimation for ARMA models, the most frequently used estimators for ARCH/GARCH models are those derived from a (conditional) Gaussian likelihood function.

The loglikelihood function of a sample of *T* observations, apart from constant, is

$$L_T(\theta) = T^{-1} \sum_{t=1}^{T} \left(log\ h_t + \varepsilon_t^2 h_t^{-1} \right), \text{ where } h_t = a_0 + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j\, h_{t-j}$$

For a general GARCH model the conditional variance (h_t) cannot be expressed in terms of a finite number of the past observations. Some truncation is inevitable. By induction, it is possible to derive

$$h_t = \frac{a_0}{1-\sum_{i=1}^{q} a_i} + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 + \sum_{i=1}^{q} a_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{p} \ldots \sum_{j_k=1}^{p} b_{j_1} .. b_{j_k}\, \varepsilon_{t-i-j_1-\ldots-j_k}^2$$

where the multiple sum vanishes if *q = 0*. It is to be noted that the multiple sum above converges with probability *1* since each a_i and b_i is nonnegative, and since the expected value of the multiple series is finite. In practice the above expression of h_t is replaced by truncation version

$$\tilde{h}_t = \frac{a_0}{1-\sum_{i=1}^{q} a_i} + \sum_{i=1}^{q} a_i\, \varepsilon_{t-i}^2 + \sum_{i=1}^{q} a_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{p} \ldots \sum_{j_k=1}^{p} b_{j_1} .. b_{j_k}\, \varepsilon_{t-i-j_1-\ldots-j_k}^2\, I\left(t-i-j_1-\ldots-j_k \geq 1\right)$$

where*t*>*q*.

In general, suppose that *f*(.) is the probability density function of ε_t. However, generally, maximum likelihood estimators are derived by minimizing

$$L_T(\theta) = T^{-1} \overset{T}{\underset{t=v}{\Sigma}} \left(log \sqrt{\tilde{h}_t} - log\ f\left(\frac{\varepsilon_t}{\sqrt{\tilde{h}_t}} \right) \right)$$

where $\tilde{h}_t$ is the truncated version of h_t (Fan and Yao, 2003).

Whitle's estimator

For GARCH(*p*,*q*) defined by (3), the conditional variance can be written a

$$h_t = \frac{a_0}{1 - \sum_{i=1}^{q} a_i} + \sum_{i=1}^{\infty} d_i\, \varepsilon_{t-i}^2 \text{ where} d_i \geq 0 \text{and } \sum_{i=1}^{\infty} d_i = \frac{\sum_{j=1}^{p} b_j}{1 - \sum_{i=1}^{q} a_i}$$

Suppose that $\{\varepsilon_t\}$ is fourth-order stationary in the sense that its first four moments are all time-invariant. $x_t = \varepsilon_t^2$ then $\{x_t\}$ is a stationary AR(∞) process satisfying

$$x_t = \frac{a_0}{1 - \sum_{i=1}^{q} a_i} + \sum_{i=1}^{\infty} d_i\, x_{t-i} + e_t \text{ where} e_t \text{ is a martingale difference}$$

$$e_t = \left(\varepsilon_t^2 - 1\right) \left\{ \frac{a_0}{\left(1 - \sum_{i=1}^{q} a_i\right)} + \sum_{i=1}^{\infty} d_i\, x_{t-i} \right\} \text{with} \sigma_e^2 = \mathrm{Var}(e_t) < \infty. \text{ Therefore, the spectral}$$

density of the process $\{x_t\}$ is $g(\omega) = \frac{\sigma_e^2}{2\pi} \left| 1 - \sum_{i=1}^{\infty} d_i\, e^{ij\omega} \right|^{-2}$

Whitle's estimators for a_i and b_i are obtained by minimizing $\sum_{j=1}^{T-1} I_T(\omega_j) / g(\omega_j)$

whereI_T(.) is the periodogram of $\{x_t\}$ and $\omega_j = 2\pi j / 2$.

Whittle's estimator suffer from the lack of efficiency, as e_t is unlikely to be normal even when η_t is normal.

Least absolute deviations estimator

Both the estimators discussed above are derived from maximizing a Gaussian likelihood or an approximate Gaussian likelihood. In time-series they are known as L_2 - estimators. Empirical evidence suggests that some financial time-series exhibit heavy-tailed than those of a normal distribution would be more appropriate. Based on this consideration, Peng and Yao (2003) proposed Least absolute deviations estimation (LADE) which minimizes

$\sum_{t=v}^{T}\left|log\varepsilon_t^2 - log(h_t)\right|$ where$v = p+1$, if $q = 0$ and $v > p+1$, if $q > 0$.

The idea behind this implies implicitly a reparameterization of model (3) such that E(ξ_t) = *0* and the median (instead of variance) of η_t^2 is equal to *1*. Peng and Yao (2002) showed that under very mild conditions, the least absolute deviations estimators are asymptotically normal with the standard convergence rate $T^{1/2}$ regardless of whether the distribution of η_t has heavy tails or not. This is in marked contrast to the conditional maximum likelihood estimators, which will suffer from slow convergence when ξ_t is heavy-tailed.

Fan and Yao (2003) and Straumann (2005) have given a good description of various estimation procedures for conditionally heteroscedastic time- series models.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for GARCH model with Gaussian distributed errors are computed by:

$$AIC = \sum_{t=1}^{T}\left(log\tilde{h}_t + \varepsilon_t^2\tilde{h}_t^{-1}\right) + 2(p + q + 1) \qquad (7)$$

and

$$BIC = \sum_{t=1}^{T}\left(log\tilde{h}_t + \varepsilon_t^2\tilde{h}_t^{-1}\right) + 2(p + q + 1)\ log(T - v + 1) \qquad (8)$$

where$T$ is the total number of observations.

Evidently, the likelihood equations are extremely complicated. Fortunately, the estimates can be obtained by using a software package, like EViews, SAS, SPLUS GARCH, GAUSS, TSP, R, MATLAB, and RATS.

Testing for ARCH Effects

Let $\varepsilon_t = y_t - \phi y_{t-1}$ be the residual series. The squared series $\{\varepsilon_t^2\}$ is then used to check for conditional heteroscedasticity, which is also known as the ARCH effects. To this end, two tests, briefly discussed below, are available. The first one is to apply the usual Ljung-Box statistic $Q(m)$ to the $\{\varepsilon_t^2\}$series. The null hypothesis is that the first m lags of autocorrelation functions of the $\{\varepsilon_t^2\}$ series are zero. The second test for conditional heteroscedasticity is the Lagrange multiplier test of Engle (1982). This test is equivalent to usual F-statistic for testing $H_0 : a_i = 0,\ i = 1, 2, \ldots, q$ in the linear regression

$$\varepsilon_t^2 = a_0 + a_1\varepsilon_{t-1}^2 + \ldots + a_q\varepsilon_{t-q}^2 + e_t,\ t = q+1, \ldots, T \qquad (9)$$

where e_t denotes the error term, q is the prespecified positive integer, and T is the sample size.

Let $SSR_0 = \sum_{t=q+1}^{T}\left(\varepsilon_t^2 - \varpi\right)^2$, where $\varpi = \sum_{t=q+1}^{T} \varepsilon_t^2 / T$ is the sample mean of $\left\{\varepsilon_t^2\right\}$, and $SSR_1 = \sum_{t=q+1}^{T} \hat{e}_t^2$, where $\hat{e}_t$ is the least squares residual of (9). Then, under H_0,

$$F = \frac{\left(SSR_0 - SSR_1\right)/q}{SSR_1\left(T - q - 1\right)} \tag{10}$$

is asymptotically distributed as chi-squared distribution with q degrees of freedom. The decision rule is to reject H_0 if $F > \chi_q^2(\alpha)$, where $\chi_q^2(\alpha)$ is the upper $100(1-\alpha)$th percentile of χ_q^2 or, alternatively, the p-value of F is less than α .

Illustration(Paul *et al*., 2009)

Paul *et al*. (2009) found that AR(*1*)-GARCH (*1*,*1*) model was better than ARIMA model for modeling and forecasting of all-India data of monthly export of spices during the period April, 2000 to November, 2006.First of all they fitted ARIMA model. The appropriate model was chosen on the. ARIMA(*1*,*1*,*1*) model is selected for modelling and forecasting of the export of spices based on minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. The estimates of parameters of above model are reported in Table 1. The graph of fitted model along with data points is exhibited in Fig. 1. Evidently, the fitted ARIMA(*1*,*1*,*1)* model is not able to capture successfully the volatility present at various time-epochs, like October, 2001; May, 2002; March, 2004; and March, 2006.

Table 1. Estimates of parameters along with their standard errors for fitted ARIMA(*1,1,1)* model

Parameter	Estimate	Standard error
AR1	-0.100	0.159
MA1	0.696	0.119
Constant	1.468	0.966

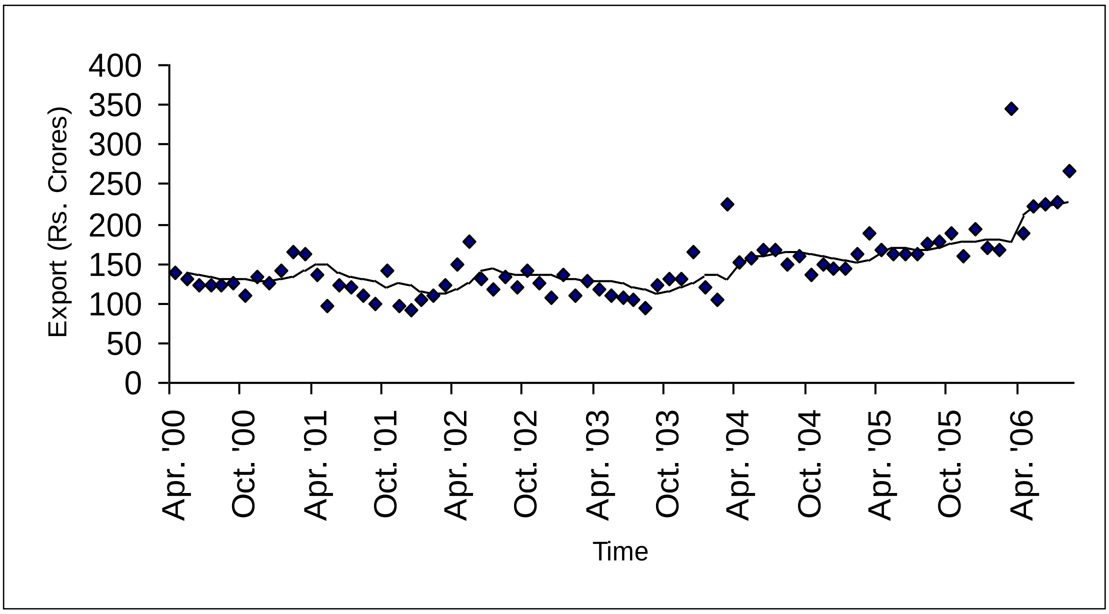


Fig. 1. Fitted ARIMA(*1*,*1*,*1*) model along with data points

Fitting of GARCH Model

On investigating autocorrelation of the squared residuals of the fitted ARIMA(*1*,*1*,*1*) model it was found that the autocorrelation was highest at lag 24, which was 0.265. The ARCH-LM test statistic at lag 24 computed using equation (10) was 37.48, which was significant at 5% level of significance. But it is not reasonable to apply ARCH model of order 24 in view of the enormously large number of parameters. Therefore, the parsimonious GARCH model is applied. The AR(*1*)-GARCH(*1*,*1*) model is selected on the basis of minimum AIC and BIC values. The estimates of parameters of the above model along with their corresponding standard errors in brackets ( ) are

$y_t = 157.99 + 0.829 y_{t-1} + \varepsilon_t$

(33.692) (0.087)

where $\varepsilon_t = h_t^{1/2} \xi_t$, and h_t satisfies the variance equation

$h_t = 1427.855 + 0.354 \varepsilon_{t-1}^2 + 0.509 h_{t-1}$

(237.058) (0.277) (0.206)

Using eqs. (7) and (8), the AIC and BIC values for fitted AR(*1*) – GARCH(*1*,*1*) model, are respectively computed as 479.77 and 521.97. To study the appropriateness of the fitted GARCH model, the autocorrelation function of the standardized residuals and squared standardized residuals are computed and it is found that, in both situations, the autocorrelation function is insignificant at 5% level of significance, thereby confirming that the mean and variance equations are correctly specified.The graph of fitted model along with data points is exhibited in Fig. 2. Obviously, the fitted GARCH model is able to capture the volatility present in the data set.EViews software package was employed for fitting of these models.

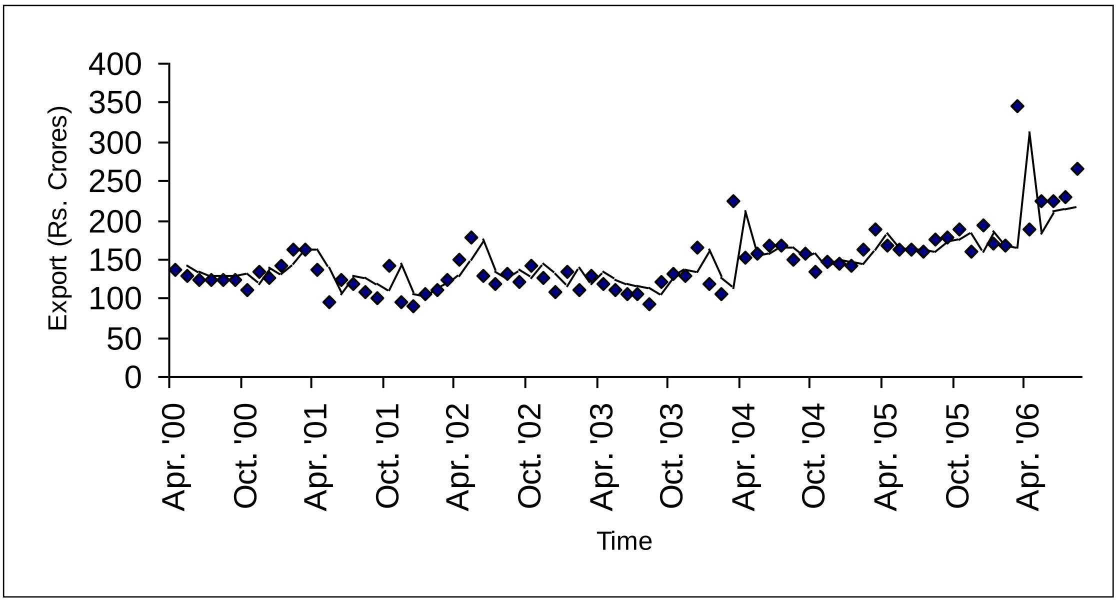


Fig. 4. Fitted AR(*1*) – GARCH(*1,1*) model along with data points

Forecasting

One-step ahead forecasts of export of spices along with their corresponding standard errors inside the brackets ( ) for the months of September, 2006 to November, 2006 in respect of above fitted models are reported in Table 2. A perusal indicates that, for fitted GARCH model, all the forecast values lie within one standard error of forecasts. However, this attractive feature does not hold for fitted ARIMA model.

The Mean square prediction error (MSPE) values and Mean absolute prediction error (MAPE) values for fitted GARCH model are respectively computed as 18.14 and 15.00, which are found to be lower than the corresponding ones for fitted ARIMA model, viz. 33.17 and 29.02 respectively.

Table 2. One-step ahead forecasts of export of spices (Rs. Crores) for fitted models

Months	**Actual Price**	**Forecasts by**	
		ARIMA(*1,1,1*)	**AR(*1*)-GARCH(*1,1*)**
Sep. ’06	270.91	235.67(29.58)	247.14 (40.93)
Oct. ’06	232.59	240.27 (30.12)	231.89 (48.17)
Nov. ’06	286.21	241.50 (31.16)	265.68 (53.31)

To sum up, it may be concluded that the AR(*1*)-GARCH(*1,1*) model has performed better than the ARIMA(*1,1,1*) model for present data for both modelling as well as forecasting purposes.

References

Angelidis, T., Benos, A. and Degiannakis, S. (2004). The use of GARCH models in VaR estimation. *Stat. Meth.*, **1**, 105-128.

Bera, A. K., and Higgins, M. L. (1993), "ARCH Models: Properties, Estimation and Testing," *J. Econ.Surv.*,**7**, 307-366.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econ.*, **31** 307-327.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2007). *Time-Series Analysis: Forecasting and Control*. 3rd edition. Pearson education, India.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*, Princeton, New Jersey: Princeton University Press.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica,* **50**, 987-1008.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, U.S.A.

Ghosh, H. and Prajneshu. (2003).Nonlinear time-series modelling of volatile onion price data using AR(p)-ARCH(q)-in-mean. *Cal. Stat. Assn. Bull.*, **54**, 231 – 47

Lanne, M. and Saikkonen, P. (2003). Modeling the U.S. short-term interest rate by mixture autoregressive processes. *J. Fin. Econ.*, **1**, 96 – 125.

Paul, R. K., Ghosh, H., and Prajneshu (2009). GARCH Nonlinear Time Series Analysis for Modelling and Forecasting of India s Volatile Spices Export Data.J. Ind. Soc. Agri. Stat.**62** (2) 123-132

Paul, R. K., Ghosh, H. and Prajneshu (2014). Development of out-of-sample forecast formulae for ARIMAX-GARCH model and their application. *Journal of the Indian Society of Agricultural Statistics*, **68**(1),85-92.

Peng, L. and Yao, Q. (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika*, **90**, 967-975.

Straumann, D. (2005). *Estimation in conditionally heteroscedastic time series models*. Springer, Germany.

Taylor, S. J. (1986). *Modeling financial time series*. Wiley, New York.

Tsay, R. S. (2005). *Analysis of financial time series*. 2nd Ed. John Wiley, U.S.A.

Wong, C. S. and Li, W. K. (2001). On a mixture autoregressive conditional heteroscedastic model. *J. Amer. Stat. Assoc.*, **96**, 992-995.

ENSEMBLE METHODS

Shashi Dahiya
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
shashi.dahiya@icar.gov.in

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. They combine multiple algorithms to produce better classification performance.It is a machine learning approach to combine multiple other models in the prediction process. The combined models increase the accuracy of the results significantly.Those models are referred to as base estimators. It is a solution to overcome the following technical challenges of building a single estimator:High variance: The model is very sensitive to the provided inputs to the learned features.

- Low accuracy: One model or one algorithm to fit the entire training data might not be good enough to meet expectations.
- Features noise and bias: The model relies heavily on one or a few features while making a prediction.

Bagging is used to reduce the variance of weak learners. Boosting is used to reduce the bias of weak learners. Stacking is used to improve the overall accuracy of strong learners.

Ensemble Algorithm

A single algorithm may not make the perfect prediction for a given dataset. Machine learning algorithms have their limitations and producing a model with high accuracy is challenging. If we build and **combine** multiple models, the overall accuracy could get boosted. The combination can be implemented by aggregating the output from each model with two objectives: reducing the model error and maintaining its generalization. The way to implement such aggregation can be achieved using some techniques. Some textbooks refer to such architecture as *meta-algorithms*.

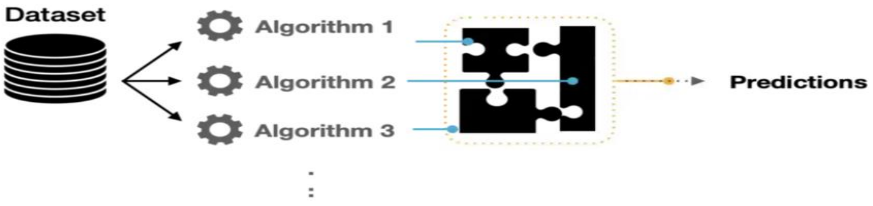


Figure 1: Diversifying the model predictions using multiple algorithms.

Ensemble Learning

Building ensemble models is not only focused on the variance of the algorithm used. For instance, we could build multiple C45 models where each model is learning a specific pattern specialized in predicting one aspect. Those models are called **weak learners** that can be used to obtain a meta-model. In this architecture of ensemble learners, the inputs are passed to each weak learner while collecting their predictions. The combined prediction can be used to build a final ensemble model.

One important aspect to mention is those weak learners can have different ways of mapping the features with variant decision boundaries.

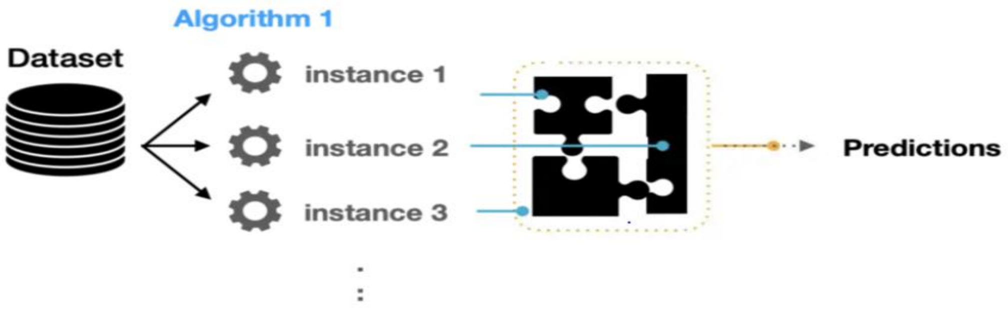


Figure 2: Aggregated predictions using multiple weak learners of the same algorithm.

Ensemble Techniques

Bagging

We use bagging for combining weak learners of high variance. Bagging aims to produce a model with lower variance than the individual weak models. These weak learners are homogenous, meaning they are of the same type.

Bagging is also known as Bootstrap aggregating. It consists of two steps: bootstrapping and aggregation.

Bootstrapping

Involves resampling subsets of data with replacement from an initial dataset. In other words, subsets of data are taken from the initial dataset. These subsets of data are called bootstrapped datasets or, simply, bootstraps. Resampled ‘with replacement’ means an individual data point can be sampled multiple times. Each bootstrap dataset is used to train a weak learner.

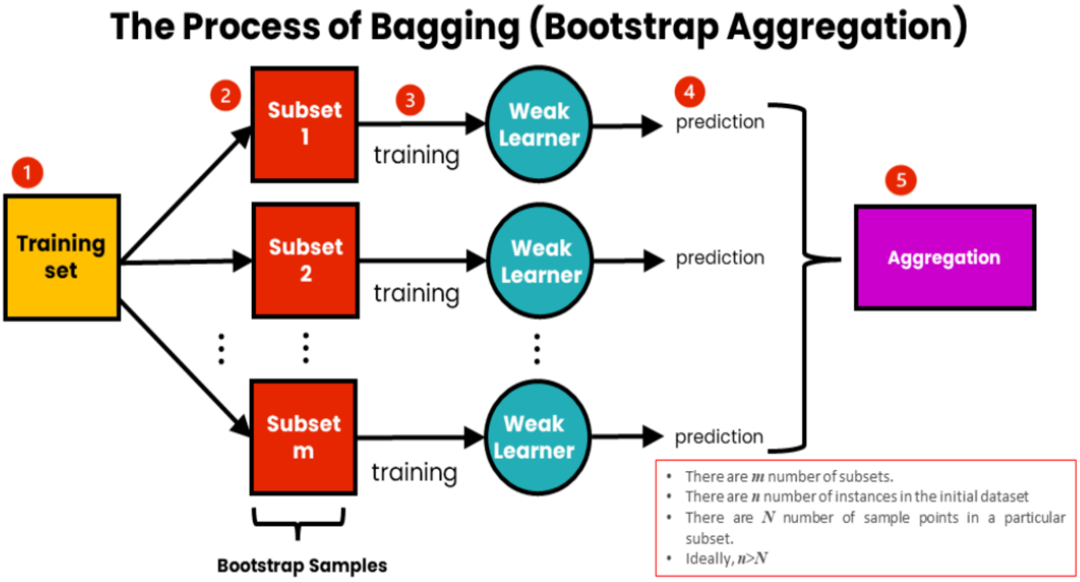
Aggregating

The individual weak learners are trained independently from each other. Each learner makes independent predictions. The results of those predictions are aggregated at the end to get the overall prediction. The predictions are aggregated using either max voting or averaging.

Max Voting is commonly used for classification problems. It consists of taking the mode of the predictions (the most occurring prediction). It is called voting because like in election voting, the premise is that ‘the majority rules’. Each model makes a prediction. A prediction from each model counts as a single ‘vote’. The most occurring ‘vote’ is chosen as the representative for the combined model.

Averaging is generally used for regression problems. It involves taking the average of the predictions. The resulting average is used as the overall prediction for the combined model.

It is one of the most straightforward and most intuitive ensemble-based algorithms that create separate samples of the training dataset. Each training dataset is used to train a different classification.



The Process of Bagging (Bootstrap Aggregation)

- There are m number of subsets.
- There are n number of instances in the initial dataset
- There are N number of sample points in a particular subset.
- Ideally, n>N

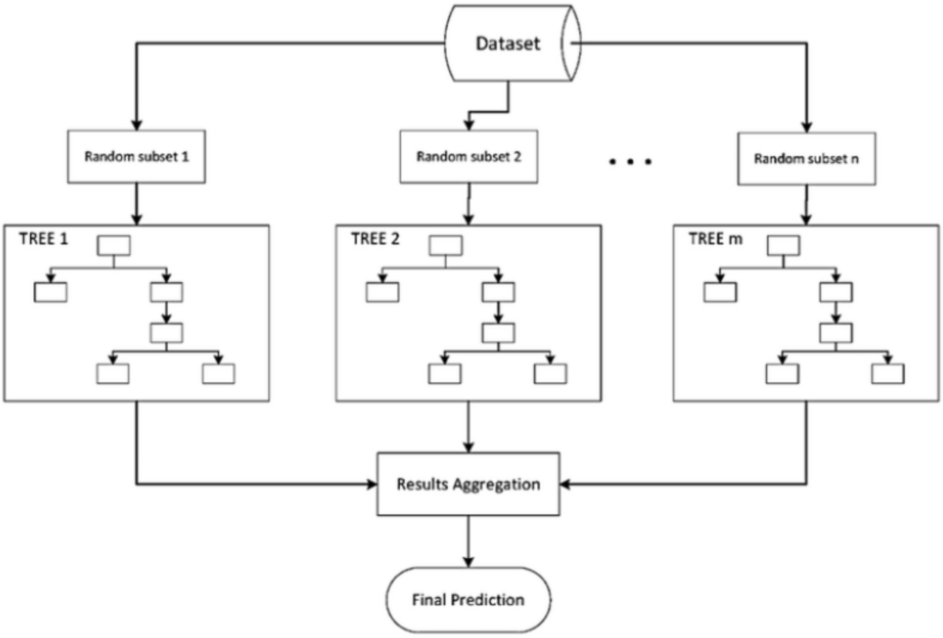
➢ Bagging

The idea of bagging is based on making the training data available to an iterative process of learning. Each model learns the error produced by the previous model using a slightly different subset of the training dataset. Bagging reduces variance and minimizes overfitting. One example of such a technique is the Random Forest algorithm.

The steps of Bagging are as follows:

1. We have an initial training dataset containing n-number of instances.

2. We create a m-number of subsets of data from the training set. We take a subset of N sample points from the initial dataset for each subset. Each subset is taken with replacement. This means that a specific data point can be sampled more than once.
3. For each subset of data, we train the corresponding weak learners independently. These models are homogeneous, meaning that they are of the same type.
4. Each model makes a prediction.
5. The predictions are aggregated into a single prediction. For this, either max voting or averaging is used.



Given a Dataset, bootstrapped subsamples are pulled. A Decision Tree is formed on each bootstrapped sample. The results of each tree are aggregated to yield the strongest, most accurate predictor.

Bagging Algorithm:

Input:

Data Set D = {(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)}

Number of iteration T

Process:

Step 1: for i = 1 to T

(a) Through sampling data points with replacement, create a dataset sample S_m.

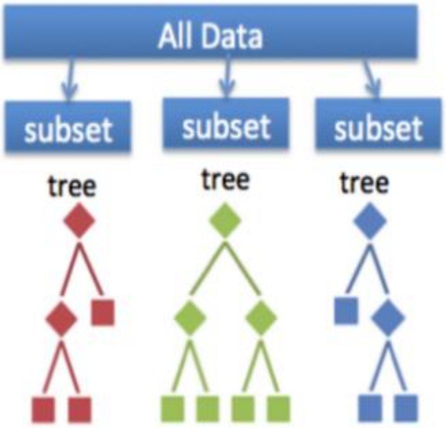
(b) From each dataset sample, S_m learns a classifier C_m.

Step 2: for every test example.

(a) Try all classifiers C_m.

(b) Estimate the class that earns the largest number of votes.

- **Random Forest:** Random Forest is another ensemble machine learning algorithm that follows the bagging technique. It is an extension of the bagging estimator algorithm. The base estimators in random forest are decision trees. Unlike bagging meta estimator, random forest randomly selects a set of features which are used to decide the best split at each node of the decision tree. It uses subset of training samples as well as subset of features to build multiple split trees. Multiple decision trees are built to fit each training set. The distribution of samples/features is typically implemented in a random mode.



A random forest takes a random subset of features from the data, and creates n random trees from each subset. Trees are aggregated together at end.

Looking at it step-by-step, this is what a random forest model does:

1. Random subsets are created from the original dataset (bootstrapping).
2. At each node in the decision tree, only a random set of features are considered to decide the best split.
3. A decision tree model is fitted on each of the subsets.
4. The final prediction is calculated by averaging the predictions from all decision trees.

Note: The decision trees in random forest can be built on a subset of data and features. Particularly, the sklearn model of random forest uses all features for *decision tree and a subset of features are randomly selected for splitting at each node.*

To sum up, Random forest **r**andomly selects data points and features, and builds multiple trees (Forest)**.**

- **Extra-Trees Ensemble:** is another ensemble technique where the predictions are combined from many decision trees. Similar to Random Forest, it combines a large number of decision trees. However, the Extra-trees use the whole sample while choosing the splits randomly.
- **Boosting:**

We use boosting for combining weak learners with high bias. Boosting aims to produce a model with a lower bias than that of the individual models. Like in bagging, the weak learners are homogeneous.

Boosting involves sequentially training weak learners. Here, each subsequent learner improves the errors of previous learners in the sequence. A sample of data is first taken from the initial dataset. This sample is used to train the first model, and the model makes its prediction. The samples can either be correctly or incorrectly predicted. The samples that are wrongly predicted are reused for training the next model. In this way, subsequent models can improve on the errors of previous models.

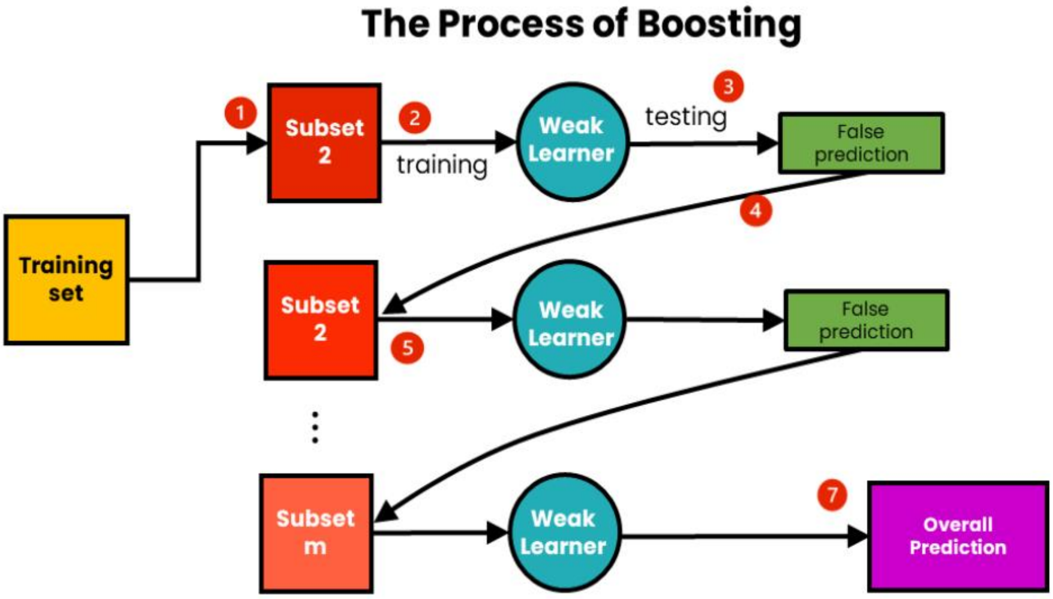
Unlike bagging, which aggregates prediction results at the end, boosting aggregates the results at each step. They are aggregated using weighted averaging.

Weighted averaging involves giving all models different weights depending on their predictive power. In other words, it gives more weight to the model with the highest predictive power. This is because the learner with the highest predictive power is considered the most important.

Boosting works with the following steps:

1. We sample m-number of subsets from an initial training dataset.
2. Using the first subset, we train the first weak learner.
3. We test the trained weak learner using the training data. As a result of the testing, some data points will be incorrectly predicted.
4. Each data point with the wrong prediction is sent into the second subset of data, and this subset is updated.
5. Using this updated subset, we train and test the second weak learner.

6. We continue with the following subset until the total number of subsets is reached.
7. We now have the total prediction. The overall prediction has already been aggregated at each step, so there is no need to calculate it.



Algorithm:

Input:

Data set D = {(X1, Y1), (X2, Y2), ..., (Xn, Yn)}

Number of iteration T

Process:

Step 1: Initialize Weight: Each case receives the same weight.

Wi = 1/N, where i = 1, 2, 3 … N.

Step 2: Construct a classifier using current weight, Compute its error:

$$Em = \frac{\sum wi \times I\{Yi \neq gm(xi)\}}{\sum wi}$$

Step 3: Get a classifier influence and update example weight.

$$am = \log\left(\frac{1 - Em}{Em}\right)$$

Step 4: Go to step 2.

- ➢ **Adaptive Boosting (AdaBoost):** is an ensemble of algorithms, where we build models on the top of several weak learners. As we mentioned earlier, those learners are called weak because they are typically simple with limited prediction capabilities. It is one of the simplest boosting algorithms. Usually, decision trees are used for modelling. Multiple sequential models are created, each correcting the errors from the last model. AdaBoost assigns weights to the observations which are incorrectly predicted and the subsequent model works to predict these values correctly.

The adaptation capability of AdaBoost made this technique one of the earliest successful binary classifiers. **Sequential** decision trees were the core of such adaptability where each tree is adjusting its weights based on prior knowledge of accuracies. Hence, we perform the training in such a technique in sequential rather than parallel process. In this technique, the process of training and measuring the error in estimates can be repeated for a given number of iteration or when the error rate is not changing significantly.

AdaBoost was the first boosting technique and is still now widely used in several domains. AdaBoost, in theory, is not prone to overfitting. Stage-wise estimation may slow down the learning process since parameters aren't jointly optimized. AdaBoost may be used to increase the accuracy of the weak classifiers, allowing it to be more flexible. It requires no normalization and has a low generalization error rate. However, training the algorithm takes enormous time. The method is also susceptible to noisy data and outliers. Therefore, removing them before employing them is strongly advised.

Looking at it step-by-step, this is what a AdaBoost model does:

1. Initially, all observations in the dataset are given equal weights.
2. A model is built on a subset of data.
3. Using this model, predictions are made on the whole dataset.
4. Errors are calculated by comparing the predictions and actual values.
5. While creating the next model, higher weights are given to the data points which were predicted incorrectly.
6. Weights can be determined using the error value. For instance, higher the error more is the weight assigned to the observation.
7. This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

- **Gradient Boosting:** Gradient Boosting or GBM is another ensemble machine learning algorithm that works for both regression and classification problems. GBM uses the boosting technique, combining a number of weak learners to form a strong learner. Regression trees used as a base learner, each subsequent tree in series is built on the errors calculated by the previous tree.Gradient boosting algorithms are great techniques that have high predictive performance. Xgboost, LightGBM, and CatBoost are popular boosting algorithms that can be used for regression and classification problems. Their popularity has significantly increased after their proven ability to win some Kaggle competitions.

- **Stacking**

Stacking, also known as Stacked Generalization,is use to improve the prediction accuracy of strong learners. Stacking aims to create a single robust model from multiple heterogeneous strong learners.

Stacking differs from bagging and boosting in that:

- It combines strong learners
- It combines heterogeneous models
- It consists of creating a Metamodel. A metamodel is a model created using a new dataset.

Individual heterogeneous models are trained using an initial dataset. These models make predictions and form a single new dataset using those predictions. This new data set is used to train the metamodel, which makes the final prediction. The prediction is combined using weighted averaging.

Because stacking combines strong learners, it can combine bagged or boosted models.

Stackingis a method similar to boosting. It is an interesting way of combining different models where multiple different algorithms are applied to the training dataset to create a model. The Meta classifier is used to predict unseen data accurately. They produce more robust predictors. It is a process of learning how to create such a stronger model from all weak learners' predictions.

It is an ensemble technique that combines multiple classifications or regression models via a meta-classifier or a meta-regressor. The base-level models are trained on a complete training set, then the meta-model is trained on the features that are outputs of the base-level model. The base-level often consists of different learning algorithms and therefore stacking ensembles are often heterogeneous.

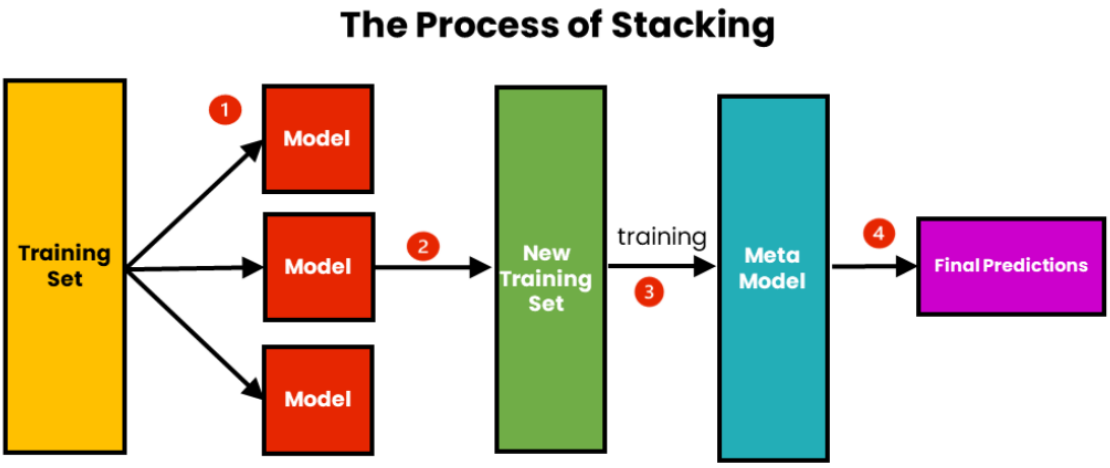
The models(Base-Model) in stacking are typically different (e.g. not all decision trees) and fit on the same dataset. Also, a single model(Meta-model) is used to learn how to best combine the predictions from the contributing models.

The architecture of a stacking model involves two or more base models, often referred to as level-0 models and a meta-model. Meta-model, also referred to as a level-1 model combines the predictions of the base models.

The steps of Stacking are as follows:

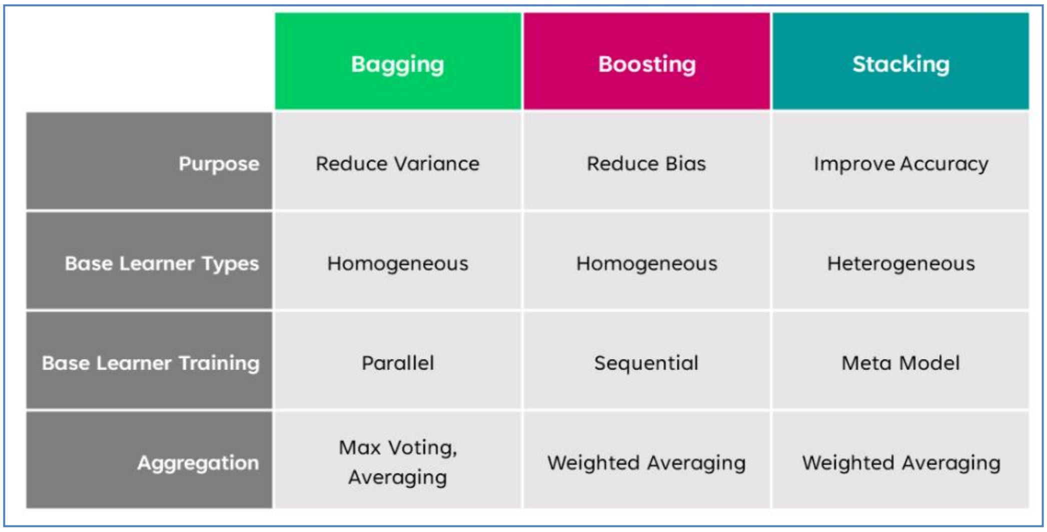
1. We use initial training data to train m-number of algorithms.
2. Using the output of each algorithm, we create a new training set.
3. Using the new training set, we create a meta-model algorithm.
4. Using the results of the meta-model, we make the final prediction. The results are combined using weighted averaging.

The outputs from the base models used as input to the meta-model may be real values in the case of regression, and probability values, probability like values, or class labels in the case of classification.



Please note that what is being learned here (as features) is the prediction from each model.

When to use Bagging, Boosting and Stacking?



	Bagging	Boosting	Stacking
Purpose	Reduce Variance	Reduce Bias	Improve Accuracy
Base Learner Types	Homogeneous	Homogeneous	Heterogeneous
Base Learner Training	Parallel	Sequential	Meta Model
Aggregation	Max Voting, Averaging	Weighted Averaging	Weighted Averaging

- If you want to reduce the overfitting or variance of your model, you use bagging. If you are looking to reduce underfitting or bias, you use boosting. If you want to increase predictive accuracy, use stacking.
- Bagging and boosting both works with homogeneous weak learners. Stacking works using heterogeneous solid learners.
- All three of these methods can work with either classification or regression problems.
- One disadvantage of boosting is that it is prone to variance or overfitting. It is thus not advisable to use boosting for reducing variance. Boosting will do a worse job in reducing variance as compared to bagging.
- On the other hand, the converse is true. It is not advisable to use bagging to reduce bias or underfitting. This is because bagging is more prone to bias and does not help reduce bias.
- Stacked models have the advantage of better prediction accuracy than bagging or boosting. But because they combine bagged or boosted models, they have the disadvantage of needing much more time and computational power. If you are looking for faster results, it’s advisable not to use stacking. However, stacking is the way to go if you’re looking for high accuracy.

References

- Larose, DT.(2006).*Data Mining Methods and Models*. Wiley-Interscience, New Jersey, USA.
- Han, J., Kamber, M., Pei, J. (2012).*Data mining: concepts and techniques.* Morgan Kaufmann, Elsevier, USA.

ASSOCIATION RULES MINING USING R

Dr. Anshu Bharadwaj
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
Anshu.Bharadwaj@icar.gov.in

1. Introduction

Mining association rules is one of the most useful data mining applications. Association rules, were first introduced in 1993 [Agrawal1993], and are used to identify relationships among a set of items in a database. These relationships are not based on inherent properties of the data themselves (as in the case of functional dependencies), but are rather based on co-occurrence of the data items. Association rules are mainly used to analyze transactional data. The association rules are useful in management, to increase the effectiveness and/or reduce the cost associated with advertising, marketing, inventory, stock location on the floor etc. Association rules also provide assistance in other applications such as prediction by identifying what events occur before a set of particular events. An association rule may be one of the following types: Boolean, Spatial, Temporal, Generalized, Quantitative, Interval, and Multiple Min-Support Association etc or a mix of them.

Formally the association rule as stated in [Agrawal1993] and [Cheung1996] is,

Let D be a transaction database and $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct items (attributes) of D, where each transaction (record) T is a set of items such that $T \subseteq I$ and has unique identifier. A transaction T is said to contain a set of item A if and only if $A \subseteq T$. An *association rule* is of the form of an implication expression $A \Rightarrow B$, where A, $B \subset$I, are sets of items called *itemsets*, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction data D with *support (s)* where s is the ratio (in percent) of the records that contain $A \cup B$ (i.e. both A and B) to the total number of records in the database, i.e. the probability $P(A \cup B)$. The rule $A \Rightarrow B$ has *confidence* (c) in the D, the ratio (in percent) of the number of records that contain $\mathrm{A} \cup \mathrm{B}$ to the number of records that contain A. This is taken to be the conditional probability $P(B \mid A)$. Mining of association rules from a database consists of finding all rules that meet the user-specified thresholds of support and confidence termed as minimum support and minimum confidence. The problem of mining association rules has been decomposed into the following two subproblems [Agrawal1994]:

1) To find all sets of items which occur with a frequency that is greater than or equal to the user-specified threshold support, say *s*.
2) To generate the rules using the frequent itemsets, which have confidence greater than or equal to the user-specified threshold confidence, say c.

The Association relationships are not based on inherent properties of the data themselves but rather based on co-occurrence of the data items. Application of association rules spans across a wide range of domains such as, business, finance, health, geographical information system, weather forecast and many such areas of real life application. The association rules in management may be handy to increase the effectiveness and/or reduce the cost associated with advertising, marketing, inventory, stock location on the floor etc. Association rules could assist in prediction of an event co-occurrence of a set of events. Association rules are generally categorized in following types: Boolean, Spatial, Temporal, Generalized, Quantitative, and Interval or may be mixed of them. The above definition of association rule is also known as Boolean Association Rule.

Association rule mining is:

- Unsupervised learning
- Used for pattern discovery
- Each rule has form: A -> B, or Left -> Right

For example: “70% of customers who purchase 2% milk will also purchase whole wheat bread.”

Data mining using association rules is the process of looking for strong rules:

1. Find the large itemsets (i.e. most frequent combinations of items)
2. Generate association rules for the above itemsets.

2. Performance Evaluation Measure of Association Rules

How to measure the strength of an association rule? Using support/confidence

Support: Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both A and B, i.e.

Support = Probability(A and B)

Support = (# of transactions involving A and B) / (total number of transactions).

Confidence: Confidence is the strength of implication of a rule; it is the percentage of transactions that contain B if they contain A, ie.

Confidence = Probability (B if A) = P(B/A)

Confidence =

(# of transactions involving A and B) / (total number of transactions that have A).

3. The Apriori Algorithm

The Apriori Algorithm is an influential algorithm for miningfrequent itemsets for boolean association rules. Some keyconcepts for Apriori algorithm are:

- Frequent Itemsets: The sets of item which hasminimum support (denoted by Li for ith-Itemset).
- Apriori Property: Any subset of frequent itemset mustbe frequent.
- Join Operation: To find Lk , a set of candidate kitemsets is generated by joining Lk-1 with itself.

Very first algorithm proposed for association rules miningwas the Apriori for frequent itemset mining. The mostpopular algorithm for pattern mining is without a doubt Apriori.It is designed to be applied on a transaction database to discover patterns in transactions made by customers in stores. But it can also be applied in several other applications. A transaction is defined a set of distinct items (symbols).

Aprioritakes as input

(1) a minsup threshold set by the user and

(2) atransaction database containing a set of transactions.

Apriorioutputs all frequent itemsets, i.e. groups of items shared by noless than minsup transactions in the input database. Forexample, consider the following transaction data base containing four transactions. Given a minsup of twotransactions, frequent itemsets are“bread, butter”, “breadmilk”, “bread”, “milk” and “butter”.

T1: bread, butter, spinach

T2: butter, salmon

T3: bread, milk, butter

T4: cereal, bread, milk

The Apriori algorithm employs the downward closureproperty if an item set is not frequent, any superset of it cannotbe frequent either. The Apriori algorithm performs a breadthfirstsearch in the search space by generating candidate k+1-itemsets from frequent k itemsets.

The frequency of an item set is computed by counting its occurrence in each transaction. Apriori is an significantalgorithm for mining frequent itemsets for Boolean associationrules. Since the Algorithm uses prior knowledge of frequentitem set it has been given the name Apriori. Apriori is aniterative level wise search Algorithm, where k- itemsets areused to explore (k+1)-itemsets. First, the set of frequents 1- itemsets is found.

This set is denoted by L1. L1 is used to find L2, the set offrequent 2-itemsets , which is used to find L3 and so on , untilno more frequent k-itemsets can be found. The finding of eachLk requires one full scan of database.

There are twosteps for understanding that how Lk-1 is usedto find Lk:-

1) The join step: To find Lk , a set of candidate k-itemsets isgenerated by joining Lk-1 with itself. This set ofcandidates is denoted Ck.

2) The prune step: Ck is a superset of Lk , that is , itsmembers may or may not be frequent , but all of thefrequent k-itemsets are included in Ck .

A scan of the database to determine the count of eachcandidate in Ck would result in the determination of Lk. Ck,however, can be huge, and so this could involve heavycomputation.

To reduce the size of Ck , the Apriori property is used as follows:

i. Any (k-1)-item set that is not frequent cannot be asubset of frequent k-item set.
ii. Hence, if (k-1) subset of a candidate k item set is notin Lk-1 then the candidate cannot be frequent eitherand so can be removed from C.

Based on the Apriori property that all subsets of a frequentitemset must also be frequent, we can determine that four lattercandidates cannot possibly be frequent. How?

For example, let’s take {I1, I2, I3}. The 2-item subsets of itare {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1,I2, I3} are members of L2, We will keep {I1, I2, I3} in C3.

Let’s take another example of {I2, I3, I5} which shows howthe pruning is performed. The 2-item subsets are {I2, I3}, {I2,I5} & {I3,I5}.

BUT, {I3, I5} is not a member of L2 and hence it is notfrequent violating Apriori Property. Thus, we will have toremove {I2, I3, I5} from C3.

Therefore, C3 = {{I1, I2, I3}, {I1, I2, I5}} after checking forall members of result of Join operation for Pruning.

Example : The Titanic Dataset

The Titanic dataset in the datasets package is a 4-dimensional table with summarized information on the fate of passengers on the Titanic according to social class, sex, age and survival. I To make it suitable for association rule mining, we reconstruct the raw data as titanic.raw, where each row represents a person. The reconstructed raw data can also be downloaded at http://www.rdatamining.com/data/titanic.raw.rdata.

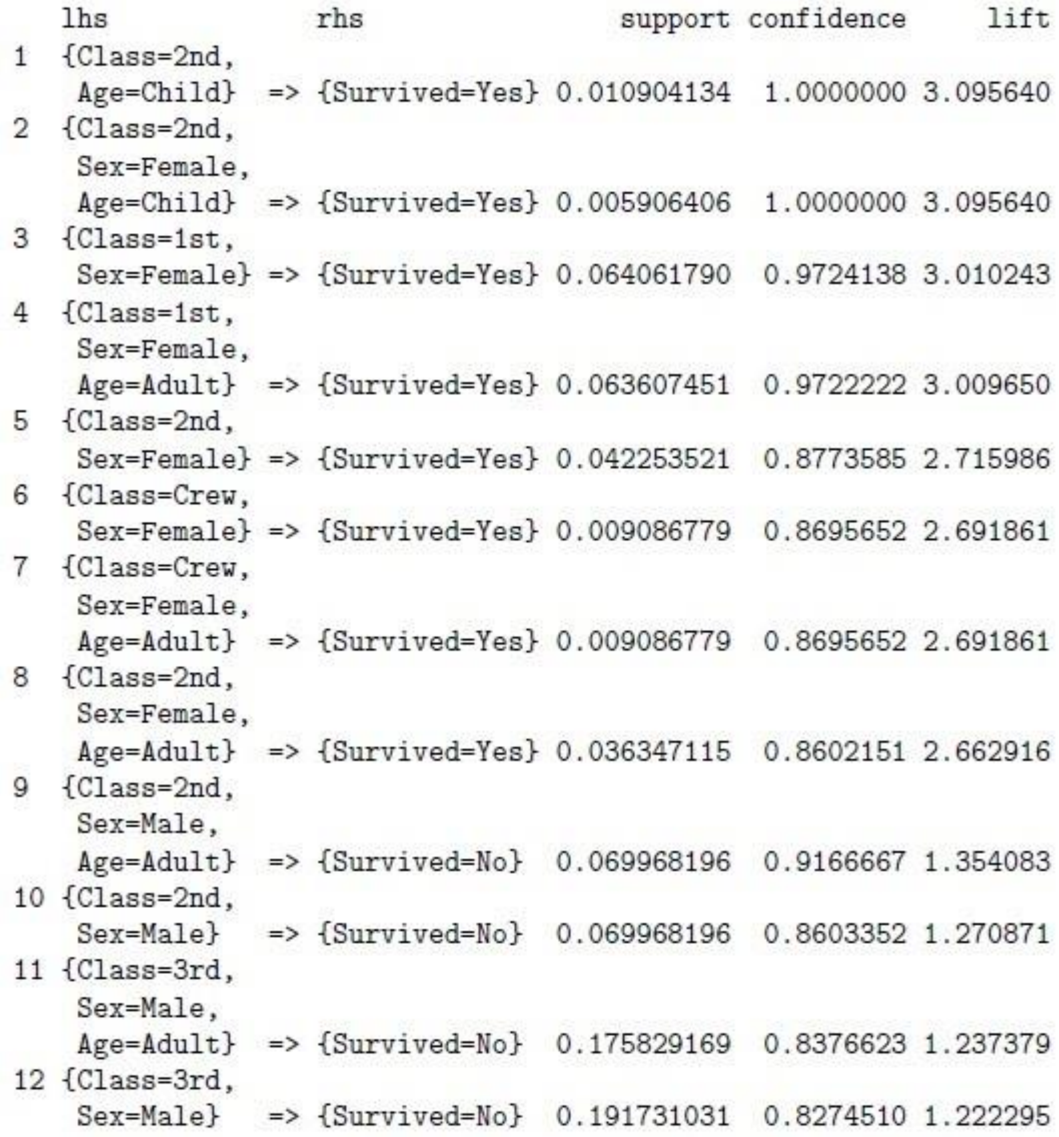
```
> str(titanic.raw)
'data.frame': 2201 obs. of 4 variables:
$ Class : Factor w/ 4 levels "1st","2nd","3rd",..: 3 3 3 3 3 3 3 3 3 3 ...
$ Sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
$ Age : Factor w/ 2 levels "Adult","Child": 2 2 2 2 2 2 2 2 2 2 ...
$ Survived: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Association Rule Mining

```
> library(arules)
> # find association rules with default settings
> rules <- apriori(titanic.raw)
> inspect(rules)
  lhs              rhs         support   confidence lift
1 {}            => {Age=Adult} 0.9504771 0.9504771  1.0000000
2 {Class=2nd}    => {Age=Adult} 0.1185825 0.9157895  0.9635051
3 {Class=1st}    => {Age=Adult} 0.1449341 0.9815385  1.0326798
4 {Sex=Female}   => {Age=Adult} 0.1930940 0.9042553  0.9513700
5 {Class=3rd}    => {Age=Adult} 0.2848705 0.8881020  0.9343750
6 {Survived=Yes} => {Age=Adult} 0.2971377 0.9198312  0.9677574
7 {Class=Crew}   => {Sex=Male}  0.3916402 0.9740113  1.2384742
```

We then set rhs=c("Survived=No", "Survived=Yes") in appearance to make sure that only "Survived=No" and "Survived=Yes" will appear in the rhs of rules.

```
> # rules with rhs containing "Survived" only
> rules <- apriori(titanic.raw,
  + parameter = list(minlen=2, supp=0.005, conf=0.8),
  + appearance = list(rhs=c("Survived=No", "Survived=Yes"),
  + default="lhs"),
  + control = list(verbose=F))
> rules.sorted <- sort(rules, by="lift")
> inspect(rules.sorted)
```

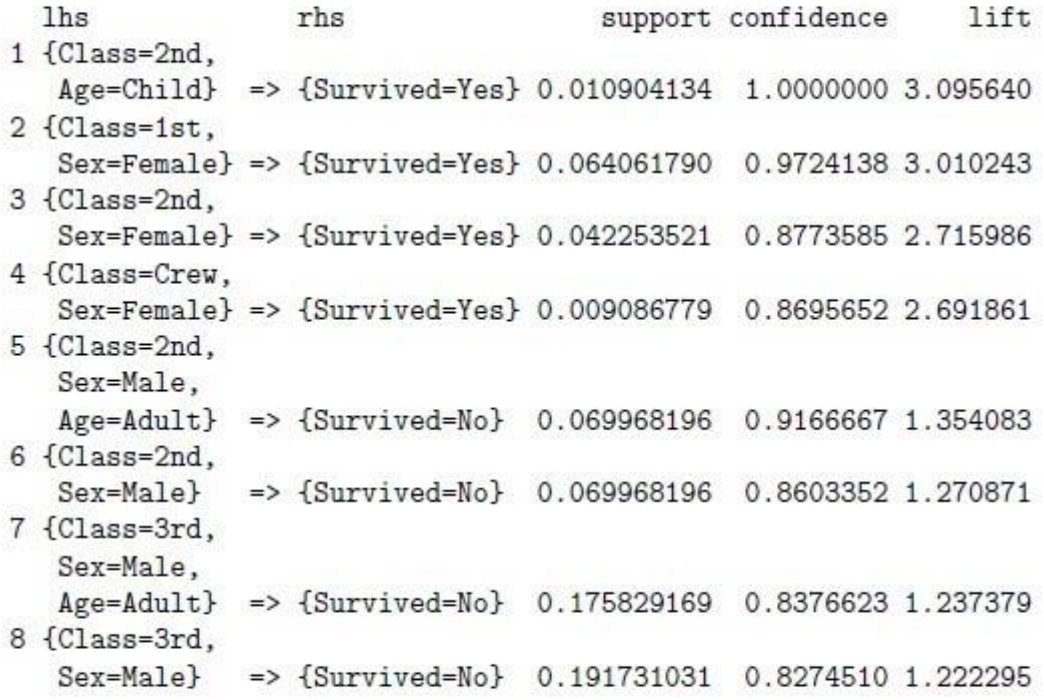


```
   lhs               rhs              support     confidence  lift
1  {Class=2nd,
    Age=Child}    => {Survived=Yes} 0.010904134  1.0000000 3.095640
2  {Class=2nd,
    Sex=Female,
    Age=Child}    => {Survived=Yes} 0.005906406  1.0000000 3.095640
3  {Class=1st,
    Sex=Female}   => {Survived=Yes} 0.064061790  0.9724138 3.010243
4  {Class=1st,
    Sex=Female,
    Age=Adult}    => {Survived=Yes} 0.063607451  0.9722222 3.009650
5  {Class=2nd,
    Sex=Female}   => {Survived=Yes} 0.042253521  0.8773585 2.715986
6  {Class=Crew,
    Sex=Female}   => {Survived=Yes} 0.009086779  0.8695652 2.691861
7  {Class=Crew,
    Sex=Female,
    Age=Adult}    => {Survived=Yes} 0.009086779  0.8695652 2.691861
8  {Class=2nd,
    Sex=Female,
    Age=Adult}    => {Survived=Yes} 0.036347115  0.8602151 2.662916
9  {Class=2nd,
    Sex=Male,
    Age=Adult}    => {Survived=No}  0.069968196  0.9166667 1.354083
10 {Class=2nd,
    Sex=Male}     => {Survived=No}  0.069968196  0.8603352 1.270871
11 {Class=3rd,
    Sex=Male,
    Age=Adult}    => {Survived=No}  0.175829169  0.8376623 1.237379
12 {Class=3rd,
    Sex=Male}     => {Survived=No}  0.191731031  0.8274510 1.222295
```

Pruning Redundant Rules

In the above result, rule 2 provides no extra knowledge in addition to rule 1, since rules 1 tells us that all 2nd-class children survived. Generally speaking, when a rule (such as rule 2) is a super rule of another rule (such as rule 1) and the former has the same or a lower lift, the former rule (rule 2) is considered to be redundant. Below we prune redundant rules.

```
> # find redundant rules
> subset.matrix <- is.subset(rules.sorted, rules.sorted)
> subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
> redundant <- colSums(subset.matrix, na.rm=T) >= 1
> which(redundant)
[1] 2 4 7 8
> # remove redundant rules
> rules.pruned <- rules.sorted[!redundant]
> inspect(rules.pruned)
```

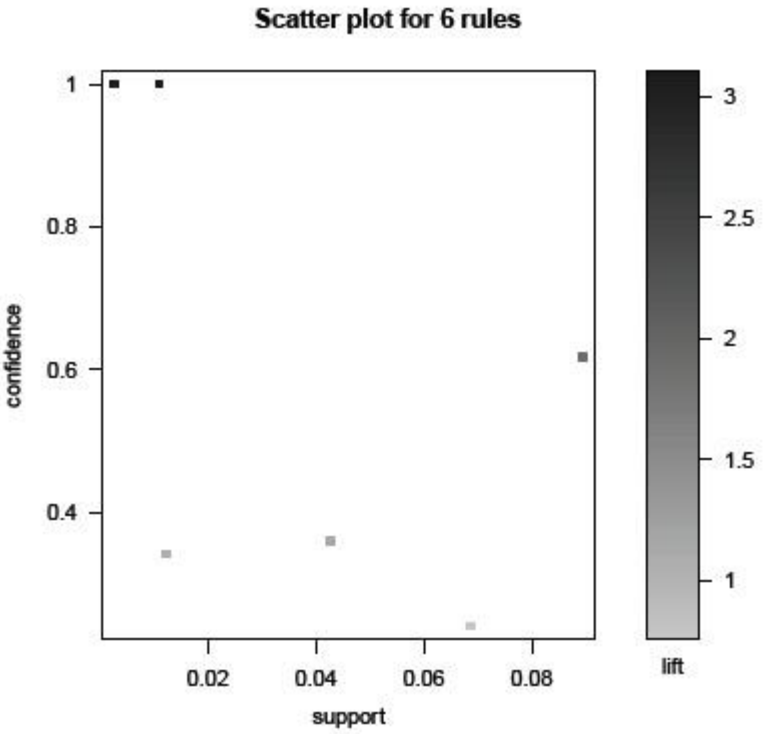


```
  lhs               rhs                support confidence     lift
1 {Class=2nd,
   Age=Child}  => {Survived=Yes} 0.010904134  1.0000000 3.095640
2 {Class=1st,
   Sex=Female} => {Survived=Yes} 0.064061790  0.9724138 3.010243
3 {Class=2nd,
   Sex=Female} => {Survived=Yes} 0.042253521  0.8773585 2.715986
4 {Class=Crew,
   Sex=Female} => {Survived=Yes} 0.009086779  0.8695652 2.691861
5 {Class=2nd,
   Sex=Male,
   Age=Adult}  => {Survived=No}  0.069968196  0.9166667 1.354083
6 {Class=2nd,
   Sex=Male}   => {Survived=No}  0.069968196  0.8603352 1.270871
7 {Class=3rd,
   Sex=Male,
   Age=Adult}  => {Survived=No}  0.175829169  0.8376623 1.237379
8 {Class=3rd,
   Sex=Male}   => {Survived=No}  0.191731031  0.8274510 1.222295
```

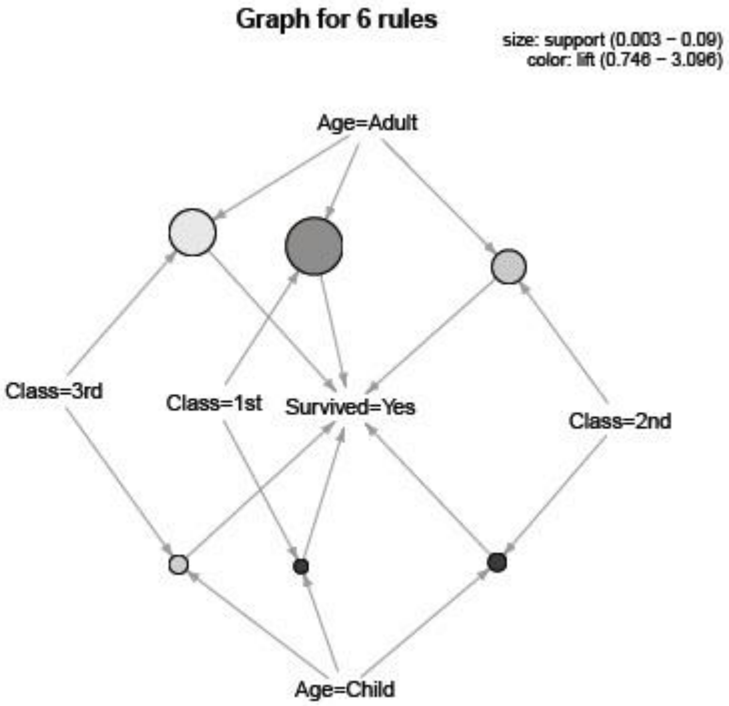
Visualizing Association Rules

Package arules Viz supports visualization of association rules with scatter plot, balloon plot, graph, parallel coordinates plot, etc.

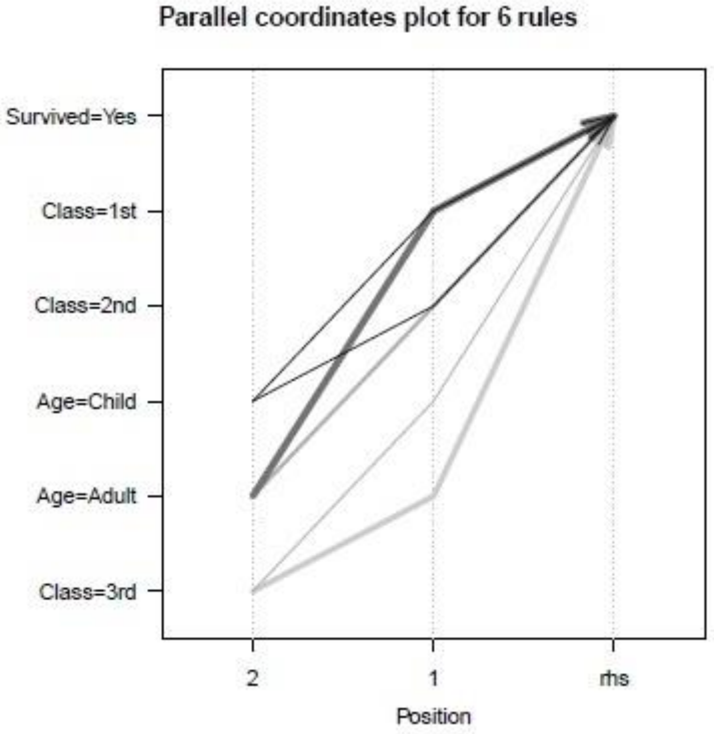
```
> library(arulesViz)
> plot(rules)
```



```
> plot(rules, method="graph", control=list(type="items"))
```



```
> plot(rules, method="paracoord", control=list(reorder=TRUE))
```



4. Frequent Pattern (FP) Growth Method

The FP-Growth Algorithm is an alternative way to find frequent itemsets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the itemset association information.

In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step it divides the compressed database into a set of conditional databases, each one associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity.

In large databases, it's not possible to hold the FP-tree in the main memory. A strategy to cope with this problem is to firstly partition the database into a set of smaller databases (called projected databases), and then construct an FP-tree from each of these smaller databases.

4.1 FP-Tree structure

The frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database [4].

Han defines the FP-tree as the tree structure io below [1]:

1. One root labeled as "null" with a set of item-prefix subtrees as children, and a frequent-item-header table (presented in the left side of Figure 1);
2. Each node in the item-prefix subtree consists of three fields:
    1. Item-name: registers which item is represented by the node;
    2. Count: the number of transactions represented by the portion of the path reaching the node;
    3. Node-link: links to the next node in the FP-tree carrying the same item-name, or null if there is none.
1. Each entry in the frequent-item-header table consists of two fields:
    1. Item-name: as the same to the node;
    2. Head of node-link: a pointer to the first node in the FP-tree carrying the item-name.

Additionally the frequent-item-header table can have the count support for an item. The Figure below show an example of a FP-tree.

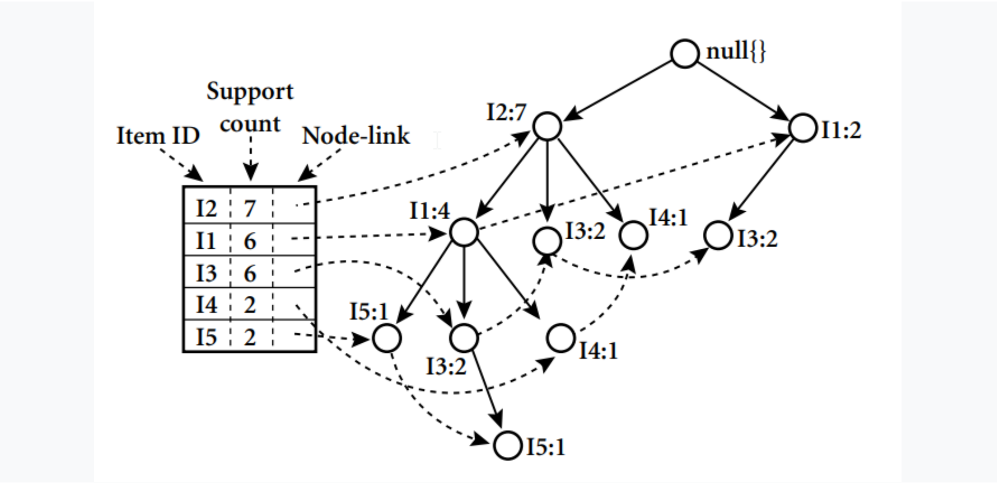
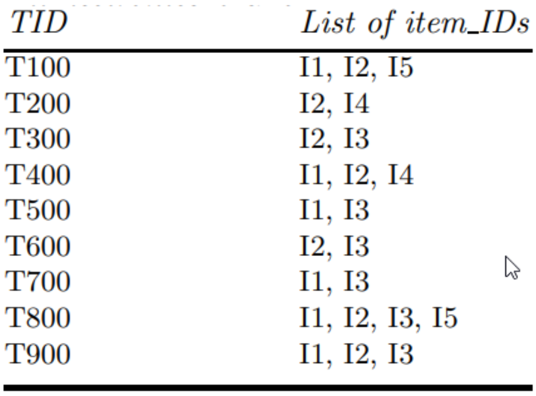


Figure 1: An FP-tree registers compressed, frequent pattern information

Table 1: Transactional data for an AllElectronics branch.

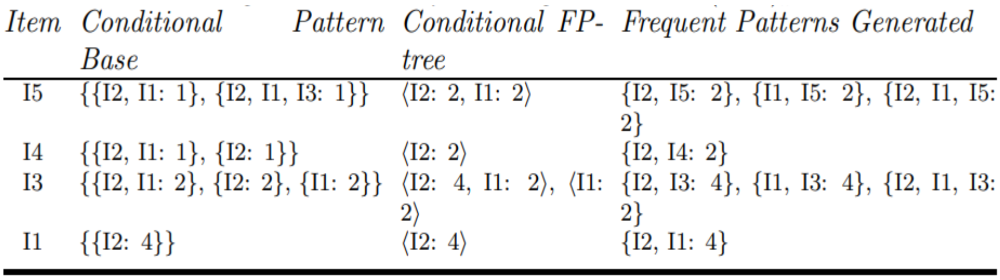


TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

The first scan of the database is the same as Apriori, which derives the set of frequent items (1-itemsets) and their support counts (frequencies). Let the minimum support count be 2. The set of frequent items is sorted in the order of descending support count. This resulting set or list is denoted by L. Thus, we have L ={{I2: 7}, {I1: 6}, {I3: 6}, {I4: 2}, {I5: 2}}. An FP-tree is then constructed as follows. First, create the root of the tree, labeled with "null." Scan database D a second time. The items in each transaction are processed inL order (i.e., sorted according to descending support count), and a branchis created for each transaction. For example, the scan of thefirst transaction, "T100: I1, I2, I5," which contains three items (I2, I1, I5 in L order), leads to the construction of the first branch of the tree with three nodes,hI2: 1i,hI1:1i, and hI5: 1i, where I2islinked as a child to the root, I1islinked to I2, and I5islinked to I1. The second transaction, T200, contains theitems I2 and I4inLorder, whichwould result in a branch where I2 is linked to the root and I4 is linked to I2. However, this branch would share a common prefix, I2, with the existing path for T100. Therefore, we insteadincrement the count of the I2 node by 1, and create a new node,hI4: 1i, which is linked as a child to hI2: 2i. In general, when considering the branch to be addedfor a transaction, the count of each node along a common prefix is incremented by 1, and nodes for the items following the prefix are created and linked accordingly. To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links. The tree obtained after scanning all of the transactions is shown in Figure 6.7 with the associated node-links. In this way, the problem of mining frequent patterns in databases is transformed to that of mining the FP-tree. The FP-tree is mined as follows. Start from each frequent length-1 pattern (as an initial suffix pattern), construct its conditional pattern base (a "sub-database," which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern), then construct its (conditional) FP-tree, and perform mining recursively on such a tree. The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree.

Mining of the FP-tree is summarized in Table 2 and detailed as follows. We first consider I5, which is the last item in L, rather than the first. The reason for starting at the end of the list will become apparent as we explain the FP-tree mining process. I5 occurs in two branches of the FP-tree of Figure 2. (The occurrences of I5 can easily be found by following its chain of node-links.) The paths formed by these branches are hI2, I1, I5: 1i and hI2, I1, I3, I5: 1i. Therefore, considering I5 as a suffix, its corresponding two prefix paths are hI2, I1: 1i and hI2, I1, I3: 1i, which form its conditional pattern base. Using this conditional pattern base as a transaction database, we build an I5-conditional FP-tree, which contains only a single path, hI2: 2, I1: 2i; I3 is not included because its support count of 1 is less than the minimum support count. The single path generates all the combinations of frequent patterns: {I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}. For I4, its two prefix paths form the conditional pattern base, {{I2 I1: 1}, {I2: 1}}, which generates a single-node conditional FP-tree, hI2: 2i, and derives one frequent pattern, {I2, I4: 2}

Table 2: Mining the FP-tree by creating conditional (sub-)pattern bases



Item	*Conditional Pattern Base*	*Conditional FP-tree*	*Frequent Patterns Generated*
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	⟨I2: 2, I1: 2⟩	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	⟨I2: 2⟩	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	⟨I2: 4, I1: 2⟩, ⟨I1: 2⟩	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	⟨I2: 4⟩	{I2, I1: 4}

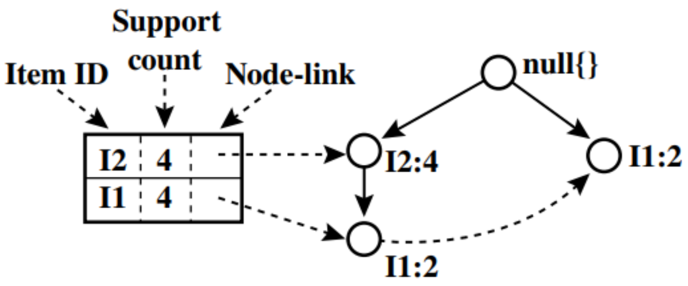


Figure 2: The conditional FP-tree associated with the conditional node I3

Similar to the above analysis, I3's conditional pattern base is {{I2, I1: 2}, {I2: 2}, {I1: 2}}. Its conditional FP-tree has two branches, hI2: 4, I1: 2i and hI1: 2i, as shown in Figure 6.8, which generates the set of patterns {{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}}. Finally, I1's conditional pattern base is {{I2: 4}}, whose FP-tree contains only one node, hI2: 4i, which generates one frequent pattern, {I2, I1: 4}. This mining process is summarized in Figure 6.9. The FP-growth method transforms the problem of finding long frequent patterns to searching for shorter ones in much smaller conditional databases recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs.

5. Basic Association Rules: Problems, Solutions and New Applications

Most of the research efforts in the scope of association rules have been oriented to simplify the rule set and to improve performance of algorithm. But these are not the only problems that can be found and when rules are generated and applied in different domains. Troubleshooting for them should also take into consideration the purpose of association model and data they come from. Some of the major drawbacks of association rule algorithms are as follows:

- Obtaining huge number of rules
- Obtaining non interesting rules

- Low algorithm performance
- Cannot incorporate domain/ user defined knowledge
- Not suitable for supervised learning

Some of the recent studies have focused on overcoming these limitations. Many algorithms for obtaining a reduced number of rules with high **support** and **confidence** have been produced. However these measures are insufficient to determine if discovered associations are really useful. An important property of discovered association rules is that they should be **interesting** and **useful**. Though interestingness of rule is a subjective aspect, many researchers have tried to come up with some ways of **measuring of interest**. It has been suggested that the rules are interesting if they are **unexpected** (unknown to user) and **actionable** (users can do something with them to their advantage). Further some other measures namely: **any-confidence, all confidence and bond** has been suggested as alternative measures of interestingness. Some authors have considered alternative measures of interest as : **gini index, entropy gain or chisquared for database or a measure of implication called conviction.** Most of the approaches for finding interesting rules require user participation to articulate his knowledge or to express what rules are interesting for him. Systems have been developed to analyze the discovered rules against user's knowledge. Discovered rules can be pruned to remove redundant and insignificant rules and further user's evaluation can be used to rank the rules. Unexpected patterns discovered may represent "holes" in domain knowledge which needs to be resolved. These patterns can thus be used to refine already existing beliefs.

Traditionally, association analysis has been considered as an unsupervised technique, so it has been applied for knowledge discovery tasks. Recent studies have shown that knowledge discovery algorithms such as association rule mining can be successfully applied for prediction in classification problems. In such cases the algorithms used for generating association rules must be tailored to peculiarities of predictions in order to build effective classifiers. Some work has been done, where association mining algorithms have been extended so that they can be used for classification/ prediction. A proposal of this category is Classification Based on Association (CBA) algorithm. The algorithm consists of two parts, a rule generator for finding association rules and a classifier builder based on these rules. Main contribution of this algorithm is possibility of making prediction on any attribute in database. Moreover, new incomplete observations can be classified.

In conclusion we can say that association rule mining is an important area of data mining research and a comparatively a younger member of data mining community. In addition to finding co-occurrence relation between items, which is basic objective, the algorithm has been applied for diverse applications. Many extensions of standard methods have been proposed. A major research area on association rules is interestingness of discovered rules. In fact its potential has still to be tapped, so that it can be tailored to solve different types of data mining problems.

STATISTICAL ANALYSIS USING SPSS SOFTWARE

Raju Kumar
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
raju.kumar@icar.gov.in

1. Introduction

SPSS (Statistical Package for the Social Sciences) is a widely used software program for statistical analysis and data management. It provides a comprehensive set of tools and features that enable researchers, data analysts, and students to perform various data-related tasks efficiently. SPSS is known for its user-friendly interface and powerful capabilities, making it a popular choice in both academia and industry.

Originally developed in 1968 by Norman H. Nie, C. Hadlai "Tex" Hull, and Dale H. Bent. The original SPSS manual (Nie*et al.*, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis. Originally it is an acronym of *Statistical Package for the Social Science* but now it stands for *Statistical Product and Service Solutions*. The current versions are officially named IBM SPSS Statistics. Long produced by SPSS Inc., it was acquired by IBM in 2009. During 2009 and 2010 it was called *PASW* (*Predictive Analytics Software*) *Statistics*.SPSS has evolved over the years and is now owned by IBM Corporation. The software has undergone several versions, with each release bringing new functionalities and enhancements to meet the ever-growing demands of statistical analysis.

SPSS allows users to import, manipulate, and analyze data from a wide range of sources, including spreadsheets, databases, and other statistical formats. The software supports both structured and unstructured data, making it versatile for different types of research and analysis. Whether you are working with survey data, experimental data, or observational data, SPSS provides the necessary tools to handle and explore your datasets effectively.

One of the key strengths of SPSS is its extensive range of statistical procedures. The software offers a vast array of statistical techniques, ranging from basic descriptive statistics to advanced multivariate analysis. Users can easily generate frequencies, descriptive statistics, cross-tabulations, and explore relationships between variables. Moreover, SPSS provides options for regression analysis, analysis of variance (ANOVA), factor analysis, cluster analysis, and many other techniques that allow for in-depth data exploration and hypothesis testing.

SPSS also provides a variety of graphical tools for visualizing data. Users can create charts, histograms, scatterplots, and other visual representations to better understand their data and communicate findings effectively. The software supports customization options, enabling users to format and design visuals to suit their specific needs.

In addition to its analytical capabilities, SPSS offers data management features to assist users in preparing and cleaning datasets. With SPSS, users can merge, subset, transform, and recode variables, ensuring data quality and consistency. This helps researchers save time and effort in data preparation, allowing them to focus more on analysis and interpretation.

SPSS is known for its user-friendly interface, making it accessible to users with varying levels of statistical knowledge and programming skills. The software offers a menu-driven interface, where users can perform tasks by selecting options from dropdown menus. However, for more advanced users, SPSS also supports a syntax-based approach, allowing for greater flexibility and automation in data analysis.

Furthermore, SPSS provides options for integration with other statistical software and programming languages. Users can import and export data in various formats, such as Excel, CSV, and SQL, facilitating seamless data exchange between different software tools. SPSS also supports integration with R and Python, allowing users to leverage the power of these programming languages for custom analyses and extensions.

In conclusion, SPSS is a powerful and versatile software program for statistical analysis and data management. With its user-friendly interface, extensive statistical procedures, and data visualization capabilities, SPSS enables researchers and data analysts to explore, analyze, and interpret data efficiently. Its wide range of features and compatibility with other software tools make SPSS a valuable asset in various fields, including social sciences, market research, healthcare, and more.

Some versions of SPSS released in recent years are

- SPSS Statistics 17.0.1 - December 2008
- PASW Statistics 17.0.3 - September 2009
- PASW Statistics 18.0, 18.0.1, 18.0.2, 18.0.3
- IBM SPSS Statistics 19.0 - August 2010
- IBM SPSS Statistics 19.0.1, 20.0, 20.0.1, 21.0, 22.0, 23.0, 24.0,25.0,26.0,27,28,29

Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services).

2.Opening SPSS

Depending on how the computer you are working on is structured, you can open SPSS in one of two ways.

1. If there is an SPSS shortcut like this on the desktop, simply put the cursor on it and double click the left mouse button.

2. Click the left mouse button on the button on your screen, then put your cursor on **Programs** or **All Programs** and left click the mouse. Select **SPSS 17.0 for Windows or IBM SPSS STATISTICS20 by** clicking the left mouse button. Either approach will launch the program.

3. Key Featuresof SPSS

Some of the key features of SPSS are

- It is easy to learn and use with its pull-down menu features
- It includes a full range of data management system and editing tools
- It offers comprehensive range of plotting, reporting and presentation features.
- It provides in-depth statistical analysis capabilities

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary stored in the datafile) are features of the base software. There are varieties of statistics included in the base software. Some of the important statistics are:

Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, DescriptiveRatio Statistics etc.

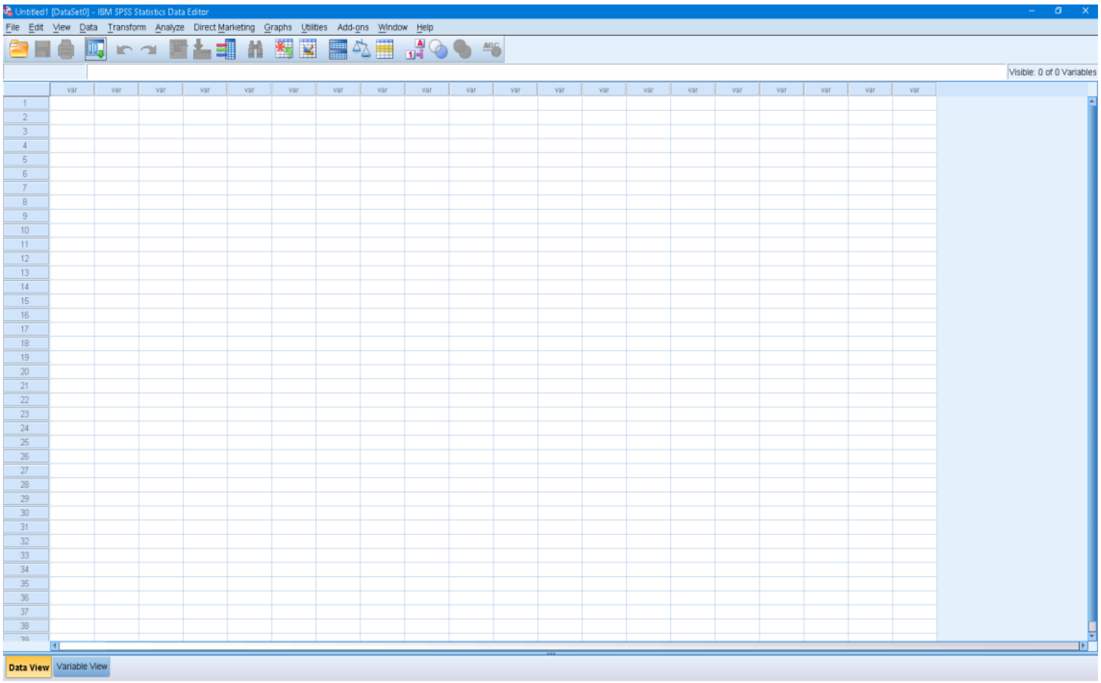
Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), nonparametric tests etc.

Prediction for numerical outcomes: Linear regression, Multiple Regression

Prediction for identifying groups: Factor analysis, Cluster analysis (two-step, K-means,hierarchical),Discriminant analysis etc.

4. Layout of SPSS

Data Editor: This graphical user interface displays the contents of the data file. One can create new data files or modify existing ones. The Data Editor window opens automatically when an SPSS session is started. The Data Editorwindow has two views that can be selected from the lower left hand side of the screen. Data Viewis where you see the data you are using. Variable Viewis where you can specify the format of your data when you are creating a file or where you can check the format of a pre-existing file. The data in the Data Editoris saved in a file with the extension .sav.The data editor offers a simple and efficient spreadsheet-like facility for entering data and browsing the working data file. To invoke SPSS in the windows environment, select the appropriate **SPSS** icon.


Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 0 of 0 Variables
var
Data View
Variable View

One can have only one data file open at a time. This editor has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS window.

- ✓ **Data view**: Displays the actual data values or defined value labels. The 'Data View' shows a spreadsheet view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells. One can modify data values in the Data view in many ways like change data values; cut, copy and paste data values; add and delete cases;
- ✓ **Variable view**: Displays variable definition information contained or metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics. One can modify variable properties in the Variable view for example, add and delete variables, change the order of variables etc.

Cells in both views can be manually edited, defining the file structure and allowing data entry without using command syntax. This may be sufficient for small datasets. Larger datasets such as statistical surveys aremore often created in data entry software, or entered during computer-assisted personal interviewing, by scanning and using optical character recognition and optical mark recognition software, or by direct capture from online questionnaires. These datasets are then read into SPSS. Extension of the saved data file will be ".sav".

Viewer: All results, tables, and charts performed by different statistical analysis are displayed in the Viewer. Extension of the saved output file will be ".spv". One can use the Viewer to browse results, show or hide selected tables and charts, change the display order of results by moving selected items or move items between the Viewer and other applications. The output presented in Viewer can be edited and saved for later use. A Viewer window opens automatically the first time a procedure is run that generates output. The Viewer is divided into two panes:

- ✓ The left pane contains an outline view of the contents. One can click an item in the outline to go directly to the corresponding table or chart.
- ✓ The right pane contains statistical tables, charts, and text output.

Syntax Editor: The pull-down menu interface generates command syntax: this can be displayed in the output. These command syntax can also be pasted into a syntax file in a syntax window using the "paste" button present in each menu. One can then edit the command syntax toutilize special features of SPSS not available through dialog boxes. These commands can be saved in a file for use in subsequent SPSS sessions. Extension of the saved syntax file will be ".sps". Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses. Additionally, some complex applications can only be programmed in syntax that are not accessible through the menu structure.

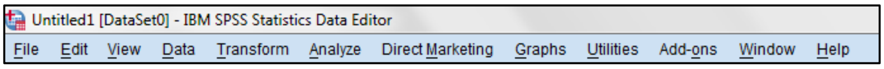
Pivot Table Editor: The results from most statistical procedures are displayed in pivot tables. These pivot tables outputs can be modified in many ways with pivot table editor. One can edit text, swap data in rows and columns, create multidimensional tables, and selectively hide and show results. Changing the layout of the table does not affect the results. Instead, it's a way to display information in a different or more desirable manner.

Text Output Editor: Text output not displayed in pivot tables can be modified with the Text Output Editor. One can edit the output and change font characteristics (type, style, colour, size).

Chart Editor: High-resolution charts and plots can be modified in chart windows. One can change the colours, select different type of fonts and sizes, switch the horizontal and vertical axes, rotate 3-D scatterplots, and even change the chart type.

Script Window: It provides the opportunity to write full-blown programs, in a BASIC-like language. It is a text editor for syntax composition. Extension of the saved script file will be ".sbs"

Many features of SPSS Statistics are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. Many of the tasks that are to be performed with SPSS start with **menu** selections. Each window has its own menu bar with menu selections appropriate for that window type. The various menu options available in SPSS are



Most menu selections open dialog boxes. One can use dialog boxes to select variables and options for analysis. Since most procedures provide a great deal of flexibility, not all of the possible choices can be contained in a single dialog box. The main dialog box usually contains the minimum information required to run a procedure. Additional specifications are made in sub-dialog boxes. All these above mentioned options have further sub-options. To see what applications there are, we simply move the cursor to a particular option and press, when a drop-down menu will appear. To cancel a drop-down menu, place the cursor anywhere outside the option and press the left button.

The three dots after an option term (...) on a drop-down menu, such as **Define Variable**...option in Data option, signifies that a dialog box will appear when this option is chosen. To cancel a dialog box, select the **Cancel** button in the dialog box. A right-facingarrowhead after an option term indicates that a further submenu will appear to the right of thedrop-down menu. An option with neither of these signs means that there are no further dropdownmenus to select. There are five standard command pushbuttons in most dialog boxes.

OK:It runs the procedure. After the variables and additional specifications are selected, clickOK to run the procedure.

Paste:It generates command syntax from the dialog box selections and pastes the syntax into a syntax window.

Reset:It deselects any variables in the selected variable list and resets all specifications in the dialog box.

Cancel:It cancels any changes in the dialog box settings since the last time it was opened and closes the dialog box.

Help:It contains information about the current dialog box.

5. Entering and Editing Data

The easiest way of entering data in SPSS is to type it directly into the matrix of columns and numbered rows in the **Data Editor** window. The columns represent variables and the rows represent cases. The variables can be defined in the variable view. Variable name must be no longer than eight characters and the name must begin with a letter.

Saving data

To be able to retrieve a file, the file must be saved with a proper name. The default extension name for saving files is **sav**. To save this file on a floppy disk, we carry out the following sequence:

> →**File** →**Save As...** [opens**Save Data As** dialog box]→box under **File Name:**delete the asterisk and type file name →**OK**

The output file can also be printed and saved. The extension name for output file is .**spo**.

Retrieving a saved file

To retrieve this file at a later stage when it is no longer the current file, use the following procedure:

> **File→Open→Data...**[opens the **Open Data File** dialog box] →choose drive from options listed →type name under **File Name:** →file name → **OK**

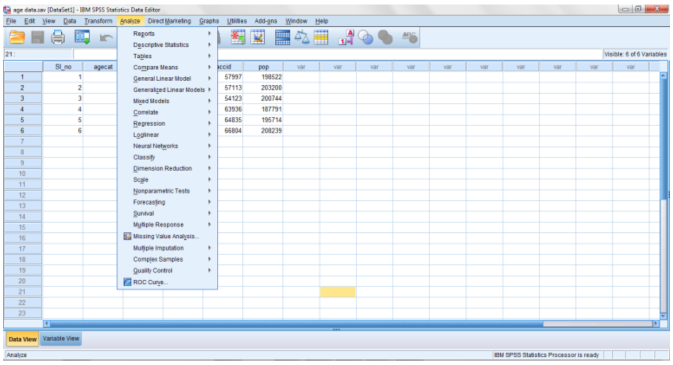
Basic Steps in Data Analysis

- **Get your data into SPSS**. You can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter your data directly in the Data Editor.
- **Select a procedure**. Select a procedure from the menus to calculate statistics or to create a chart.
- **Select the variables for the analysis**. The variables in the data file are displayed in a dialog box for the procedure.
- **Run the procedure**. Results are displayed in the Viewer.

6. Statistical Procedures

After entering the data set in **Data Editor** or reading an ASCII data file, we are now ready to analyse it. The **Analyse** option has the following sub options:

Reports, Descriptive Statistics, Tables, Compare means, General Linear model, Mixed Models, Correlate, Regression, Loglinear, Neural Networks, Classify, Dimension Reduction, Scale, Non parametric tests, Forecasting, Time Series, Survival, Multiple response, Missing value analysis, Multiple imputation, Complex samples, Quality control, ROC curve.



	Sl_no	agecat	ccid	pop
1	1		57997	198522
2	2		57113	203200
3	3		54123	200744
4	4		63936	187791
5	5		64835	195714
6	6		66804	208239

6.1Reports:

This submenu provides techniques for reporting the results. The various sub-sub menus under this are as follows:

Codebook reports the dictionary information such as variable names, variable labels, value labels, missing values and summary statistics for all or specified variables and multiple response sets in the active dataset. For nominal and ordinal variables and multiple response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation, and quartiles.

OLAP (Online Analytical Processing) **Cubes** procedure calculates totals, means, and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.

Case Summaries calculates subgroup statistics for variables within categories of one or more grouping variables. All levels of the grouping variable are cross tabulated. One can choose the order in which the statistics are displayed. Summary statistics for each variable across all categories are also displayed. With large datasets, one can choose to list only the first n cases.

Report Summaries in Rows produces reports in which different summary statistics are laid out in rows. Case listings are also available from this command, with or without summary statistics.

Report Summaries in Columns produces reports in which different summary statistics are laid out in separate columns.

6.2 Descriptive Statistics:

This submenu provides techniques for summarizing data with statistics, charts, and reports. The various sub-sub menus under this are as follows:

Frequencies provides information about the relative frequency of the occurrence of each category of a variable. This can be used it to obtain summary statistics that describe the typical value and the spread of the observations. To compute summary statistics for each of several groups of cases, Means procedure or the Explore procedure can be used.

Descriptivesis used to calculate statistics that summarize the values of a variable like the measures of central tendency, measures of dispersion, skewness, kurtosis etc.

Explore produces and displays summary statistics for all cases or separately for groups of cases. Boxplots, stem-and leaf plots, histograms, tests of normality, robust estimates of location, frequency tables and other descriptive statistics and plots can also be obtained.

Crosstabs is used to count the number of cases that have different combinations of values of two or more variables, and to calculate summary statistics and tests. The variables you use to form the categories within which the counts are obtained should have a limited number of distinct values.

P-P plots provides the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.

Q-Q plots provide the quantiles of a variable's distribution against the quantiles of the normal distribution.

6.3 Tables:

Custom Tables submenu provides attractive, flexible displays of frequency counts, percentages and other statistics.

6.4 Compare Means:

This submenu provides techniques for testing differences among two or more means for both independent and related samples.

Means computes summary statistics for a variable when the cases are subdivided into groups based on their values for other variables.

One-Sample tTest procedure tests whether the mean of a single variable differs from a specified constant. For each test variable: mean, standard deviation, and standard error of the mean.

Independent Sample t test is used if two unrelated samples come from populations with the same mean. The observations should be from two unrelated groups, and for testing, the mean must be an appropriate summary measure for the variable to be compared in the two groups. For more than two independent groups, the *One-way ANOVA* option could be used.

Paired Sample t test is used to compare the means of the same subjects in two conditions or at two points in time i.e. to compare subjects who had been matched to be similar in certain respects and then to test if two related samples come from populations with the same mean. The related, or paired, samples often result from an experiment in which the same person is observed before and after an intervention. If the distribution of the differences of the values between the members of a pair is markedly non-normal you should consider one of the nonparametric tests.

One-Way ANOVA is used to test that several independent groups come from populations with the same mean. To see which groups are significantly different from each other, multiple comparison procedures can be used through *Post Hoc Multiple Comparison option* which consist of the options like *Least-significant difference, Duncan's multiple range test, Scheffe*etc. The contrast analysis can also be performed in order to compare the different groups or treatments by using the *Contrast* option. The data obtained using completely randomised design can be analysed through this option.

6.5 General Linear Model

This submenu provides techniques for testing univariateand multivariate Analysis-of-Variance models, including repeated measures.

Univariatesub-option could be used to analyse the experimental designs like Completely randomised design, Randomised block design, Latin square design, Designs for factorial experiments etc. The covariance analysis can also be performed and alternate methods for partitioning sums of squares can be selected. If only some of the interactions of a particular order are to be included, the *Custom* procedure should be used. If there is only one factor then One-Way ANOVA procedure should be used.

Multivariate analyses analysis-of-variance and analysis-of-covariance designs when you have two or more correlated dependent variables. Multivariate analysis of variance is used to test hypotheses about the relationship between a set of interrelated dependent variables and one or more factor or grouping variables. For example, you can test whether verbal and mathematical test scores are related to instructional method used, sex of the subject, and the interaction of method and sex. This procedure should be used only if there are several dependent variables which are related to each other. For a single dependent variable or unrelated dependent variables, the Univariate ANOVA procedures can be adopted. If the same dependent variable is measured on several occasions for each subject, the Repeated Measures procedure is to be used.

Repeated Measures is used to test hypotheses about the means of a dependent variable when the same dependent variable is measured on more than one occasion for each subject. Subjects can also be classified into mutually exclusive groups, such as males or females, or type of job held. Then you can test hypotheses about the effects of the between-subject variables and the within-subject variables, as well as their interactions.

6.6 Correlate

This submenu provides measures of association for two or more variables measured at the interval level.

Bivariate calculates matrices of Pearson product-moment correlations, and of Kendall and Spearman nonparametric correlations, with significance levels and optional univariate statistics. The correlation coefficient is used to quantify the strength of the linear relationship between two variables. The *Pearson correlation coefficient* should be used only for data measured at the interval or ratio level. Spearman and Kendall correlation coefficients are nonparametric measures which are particularly useful when the data contain outliers or when the distribution of the variables is markedly non-normal. Both the Spearman and Kendall coefficients are based on assigning ranks to the variables.

Partial calculates *partial correlation coefficients* that describe the relationship between two variables, while adjusting for the effects of one or more additional variables. If the value of a dependent variable from a set of independent variables is to be predicted then the Linear Regression procedure may be used. If there are no control variables then the Bivariate Correlations procedure can be adopted. Nominal variables should not be used in the partial correlation procedure.

Distances calculates statistics measuring either similarities or dissimilarities (distances), either between pairs of variables or between pairs of cases. These similarity or distance measures can then be used with other procedures, such as factor analysis, cluster analysis, or multidimensional scaling, to help analyze complex datasets. Dissimilarity (distance) measures for interval data are Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, or customized; for count data, chi-square or phi-square; for binary data, Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, shape, or Lance and Williams. Similarity measures for interval data are Pearson correlation or cosine; for binary data, Russel and Rao, simple matching, Jaccard, etc.

6.7 Regression

This submenu provides a variety of regression techniques, including linear, logistic, nonlinear, weighted, and two-stage least-squares regression.

Linear is used to examine the relationship between a dependent variable and a set of independent variables. If the dependent variable is dichotomous, then the logistic regression procedure should be used. If the dependent variable is censored, such as survival time after surgery, use the Life Tables, Kaplan-Meier, or proportional hazards procedure.

Curve Estimation produces curve estimation regression statistics and related plots for 11 different curve estimation regression models. A separate model is produced for each dependent variable. One can also save predicted values, residuals, and prediction intervals as new variables.

Logistic estimates regression models in which the dependent variable is dichotomous. If the dependent variable has more than two categories, use the Discriminant procedure to identify variables which are useful for assigning the cases to the various groups. If the dependent variable is continuous, use the Linear Regression procedure to predict the values of the dependent variable from a set of independent variables. In recent versions there are two options **Binary Logistic** as well as **Multinomial Logistic.**

Probit performs probit analysis which is used to measure the relationship between a response proportion and the strength of a stimulus. For example, the probit procedure can be used to examine the relationship between the proportion of plants dying and the strength of the pesticide applied or to examine the relationship between the proportion of people buying a product and the magnitude of the incentive offered. The Probit procedure should be used only if the response is dichotomousbuy/not buy, alive/dead--and several groups of subjects are exposed to different levels of some stimulus. For each stimulus level, the data must contain counts of the totals exposed and the totals responding. If the response variable is dichotomous but you do not have groups of subjects with the same values for the independent variables you should use the Logistic Regression procedure.

Nonlinear estimates nonlinear regression models, including models in which parameters are constrained. The nonlinear regression procedure can be used if one knows the equation whose parameters are to be estimated, and the equation cannot be written as the sum of parameters times some function of the independent variables. In nonlinear regression the parameter estimates are obtained iteratively. If the function is linear, or can be transformed to a linear function, then the Linear Regression procedure should be used.

Weight Estimation estimates a linear regression model with differential weights representing the precision of observations. This command is in the Professional Statistics option. If the variance of the dependent variable is not constant for all of the values of the independent variable, weights which are inversely proportional to the variance of the dependent variable can be incorporated into the analysis. This results in a better solution. The Weight Estimation procedure can also be used to estimate the weights when the variance of the dependent variable is related to the values of an independent variable. If you know the weights for each case you can use the linear regression procedure to obtain a weighted least squares solution. The linear regression procedure provides a large number of diagnostic statistics which help you evaluate how well the model fits your data.

2-Stage Least Squares performs two-stage least squares regression for models in which the error term is related to the predictors. This command is in the Professional Statistics option. For example, if you want to model the demand for a product as a function of price, advertising expenses, cost of the materials, and some economic indicators, you may find that the error term of the model is correlated with one or more of the independent variables. Two-stage least squares allows you to estimate such a model.

The **Loglinear** submenu provides general and hierarchical log-linear analysis and logit analysis.

6.8 Classify

This submenu provides cluster and discriminant analysis.

Two Step Cluster performs Two Step Cluster Analysis procedure which is an exploratory data analysis tool designed to reveal natural clustering within a dataset that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques. The Log-likelihood and Euclidean Distance Measures are used as the similarity measure between two clusters.

K-means Cluster performs cluster analysis using an algorithm that can handle large numbers of cases, but that requires you to specify the number of clusters. The goal of cluster analysis is to identify relatively homogeneous groups of cases based on selected characteristics. If the number of clusters to be formed is not known, then Hierarchical Cluster procedure can be used. If the observations are in known groups and one wants to predict group membership based on a set of independent variables, then the Discriminant procedure can be used.

Hierarchical Cluster combines cases into clusters hierarchically, using a memory-intensive algorithm that allows you to examine many different solutions easily.

Discriminant is used to classify cases into one of several known groups on the basis of various characteristics. To use the Discriminant procedure the dependent variable must have a limited number of distinct categories. Independent variables that are nominal must be recoded to dummy or contrast variables. If the dependent variable has two categories, Logistic Regression can be used. If the dependent variable is continuous one may use Linear Regression.

Nearest Neighbor performs Nearest Neighbor Analysis for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

6.9 Dimension Reduction

This submenu provides factor analysis, correspondence analysis, and optimal scaling.

Factor is used to identify factors that explain the correlations among a set of variables. Factor analysis is often used to summarize a large number of variables with a smaller number of derived variables, called factors.

Correspondence Analysis analyzes correspondence tables (such as cross-tabulations) to best measure the distances between categories or between variables. This command is in the Categories option.

Distances computes many different measures of similarity, dissimilarity or distance. Many different measures can be used to quantify how much alike or how different two cases or variables are. Similarity measures are constructed so that large values indicate much similarity and small values indicate little similarity. Dissimilarity measures estimate the distance or unlikeness of two cases. A large dissimilarity value tells that two cases or variables are far apart. In order to decide which similarity or dissimilarity measure to use, one must consider characteristics of the data. Special measures are available for interval data, frequency counts, and binary data. If the cases are to be classified into groups based on similarity or dissimilarity measures, one of the Cluster procedures should be used.

6.10 Scale

This submenu provides reliability analysis and multidimensional scaling.

Reliability analysis allows to study the properties of measurement scales and the items that compose the scales. The Reliability Analysis procedure calculates a number of commonly used measures of scale reliability and also provides information about the relationships between individual items in the scale. This provides several statistics like descriptives for each variable and for the scale, summary statistics across items, inter-item correlations and covariances, reliability estimates, ANOVA table, intraclass correlation coefficients, Hotelling's T2, and Tukey's test of additivity.

6.11 Nonparametric Tests:

This submenu provides nonparametric tests for one sample, or for two and more paired or independent samples. Legacy dialogs sub-submenu consists following tests

Chi-Square is used to test hypotheses about the relative proportion of cases falling into several mutually exclusive groups. For example, if one wants to test the hypotheses that people are equally likely to buy six different brands of cereals, one can count the number buying each of the six brands. Based on the six observed counts Chi-Square procedure could be used to test the hypothesis that all six cereals are equally likely to be bought. The expected proportions in each of the categories don't have to be equal. The hypothetical proportions to be tested should be specified.

Binomial is used to test the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values. For example, to test that the probability of an item on the assembly line is defective is one out of ten (p=0.1), take a sample of 300 items and record whether each is defective or not. Then use the binomial procedure to test the hypothesis of interest.

Runs is used to test whether the two values of a dichotomous variable occur in a random sequence. The runs test is appropriate only when the order of cases in the data file is meaningful.

1-Sample K-S is used to compare the observed frequencies of the values of an ordinal variable, such as rated quality of work, against some specified theoretical distribution. It determines the statistical significance of the largest difference between them. In SPSS, the theoretical distribution can be **Normal, Uniform or Poisson**. Alternative tests for normality are available in the Explore procedure, in the Summarize submenu. The P-P and Q-Q plots in the Graphs menu can also be used to examine the assumption of normality.

2-Independent Samples is used to compare the distribution of a variable between two non-related groups. Only limited assumptions are needed about the distributions from which the sample are selected. The Mann-Whitney U test is an alternative to the two sample t-test. The actual values of the data are replaced by ranks. The Kolmogorov-Smirnov test is based on the differences between the observed cumulative distributions of the two groups. The Wald-Woflowitz runs tests sorts the data values from smallest to largest and then performs a runs test on the group’s numbers. The Moses Test of Extreme Reaction is used to test for differences in range between two groups.

K-Independent Samples is used to compare the distribution of a variable between two or more groups. Only limited assumptions are needed about the distributions from which the samples are selected. The Kruskal-Wallis test is an alternative to one-way analysis of variance, with the actual values of the data replaced by ranks. The Median tests counts the number of cases in each group that are above and below the combined median, and then performs a chi-square test.

2 Related Samples is used to compare the distribution of two related variables. Only limited assumptions are needed about the distributions from which the samples are selected. The Wilcoxon and Sign tests are nonparametric alternative to the paired samples t-test. The Wilcoxon test is more powerful than the Sign test. *McNemar's test*is used to determine changes in proportions for related samples. It is often used for "before and after" experimental designs when the dependent variable is dichotomous. For example, the effect of a campaign speech can be tested by analysing the number of people whose preference for a candidate changed based on the speech. Using McNemar's test you analyse the changes to see if change in both directions is equally likely.

K Related Samples is used to compare the distribution of two or more related variables. Only limited assumptions are needed about the distributions from which the samples are selected. *The Friedman test* is a nonparametric alternative to a single-factor repeated measures analysis of variance. You can use it when the same measurement is obtained on several occasions for a subject. For example, the Friedman test can be used to compare consumer satisfaction of 5 products when each person is asked to rate each of the products on a scale. *Cochran's Q test*can be used to test whether several dichotomous variables have the same mean. For example, if instead of asking each subject to rate their satisfaction with five products, you asked them for a yes/no response about each, you could use Cochran's test to test the hypothesis that all five products have the same proportion of satisfied users. *Kendall's W measures*the agreement among raters. Each of your cases corresponds to a rater, each of the selected variables is an item being rated. For example, if you ask a sample of customers to rank 7 ice-cream flavours from least to most liked, you can use Kendall's W to see how closely the customers agree in their ratings.

6.12 Forecasting

This submenu provides create models, seasonal decomposition, spectral analysis, autocorrelations, cross-correlations etc.

Autocorrelations calculates and plots the autocorrelation function (ACF) and partial autocorrelation function of one or more series to any specified number of lags, displaying the Box-Ljung statistic at each lag to test the overall hypothesis that the ACF is zero at all lags.

Cross-correlations calculates and plots the cross-correlation function of two or more series for positive, negative, and zero lags.

Spectral analysis calculates and plots univariate or bivariate periodograms and spectral density functions, which express variation in a time series (or covariation in two time series) as the sum of a series of sinusoidal components. It can optionally save various components of the frequency analysis as new series.

6.13 Survival:

The submenu provides techniques for analyzing the time for some terminal event to occur, including Kaplan-Meier analysis and Cox regression.

6.14Multiple Response:

This submenu provides facilities to define and analyze multiple-response or multiple-dichotomy sets.

Quality Control submenu provides facilities to for obtaining control charts and Pareto charts.

Complex Samples submenu provides procedures for Sampling from Complex Designs. The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

Other than this **Analyze** menu there are several other important menus available in SPSS.

6.15 Transform

Compute calculates the values for either a new or an existing variable, for all cases or for cases satisfying a logical criterion.

Random Number Seed sets the seed used by the pseudo-random number generator to a specific value, so that you can reproduce a sequence of pseudo-random numbers.

Count creates a variable that counts the occurrences of the same value(s) in a list of variables for each case.

Recode into Same Variables reassigns the values of existing variables or collapses ranges of existing values into new values.

Recode into Different Variables reassigns the values of existing variables to new variables or collapses ranges of existing values into new variables.

Rank Cases creates new variables containing ranks, normal scores, or similar ranking scores for numeric variables.

Automatic Recode reassigns the values of existing variables to consecutive integers in new variables.

Create Time Series creates a time-series variable as a function of an existing series, for example, lagged or leading values, differences, cumulative sums. This command is in the Trends option.

Replace Missing Values substitutes non-missing values for missing values, using the series mean or one of several time-series functions. This command is in the Trends option.

Run Pending Transforms executes transformation commands that are pending due to the Transformation Options setting in the Preferences dialog.

6.16 Utilities

Command Index take you to the dialog box for a command if you know its name in the SPSS command language.

Fonts lets you choose a font, style, and size for SPSS Data Editor, output, and syntax windows.

Variable Information displays the Variables window, which shows information about the variables in your working data file, and allows you to scroll the data editor to a specific variable, or copy variable names to the designated syntax window.

File Information displays information about the working data file in the output window.

Output Page Titles lets you specify a title and subtitle for output from SPSS. They appear in the page header, if it is displayed. (Preferences in the Edit menu controls the page header.)

Define Sets defines sets of variables for use in other dialog boxes.

Use Sets lets you select which defined sets of variables should appear in the source-variable lists of other dialog boxes.

Grid Lines turns grid lines on and off in the Data Editor window. This command is available when the Data Editor is active.

Value Labels turns on and off the display of Value Labels (instead of actual values) in the Data Editor window. When Value Labels are displayed you can edit data with a pop-up menu of labels. This command is available when the Data Editor is active.

Auto New Case turns on and off the automatic creation of new cases by cursor movement below the last case in the Data Editor window. This command is available when the Data Editor is active.

Designate Window designates the active window to receive output from SPSS commands (if it is an output window); or to receive commands pasted from dialog boxes (if it is a syntax window). You can also designate a window by clicking the !button on its icon bar. This command is available when an output or syntax window is active.

6.17 Graphs

The Chart Builder available in Graph menu allows to build charts from predefined gallery charts or from the individual parts (for example, axes and bars). You build a chart by dragging and dropping the gallery charts or basic elements onto the canvas, which is the large area to the right of the Variables list in the Chart Builder dialog box.

Legacy Dialogs submenu provides following graph submenus

Bar generates a simple, clustered, or stacked bar chart of the data.

3-D Bar Charts allows to generate bar graph in 3-dimensional axis.

Line generates a simple or multiple line chart of the data.

Area generates a simple or stacked area chart of the data.

Pie generates a simple pie chart or a composite bar chart from the data.

High-Low plots pairs or triples of values, for example high, low, and closing prices.

Boxplot generates boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.

Error Bar Charts plot the confidence intervals, standard errors, or standard deviations of individual variables.

Scatter/dot generates a simple or overlay scatter plot, a scatter plot matrix, or a 3-D scatter plot from the data.

Histogram generates a histogram showing the distribution of an individual variable.

Practical exercise using SPSS.

Exercise 1: The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

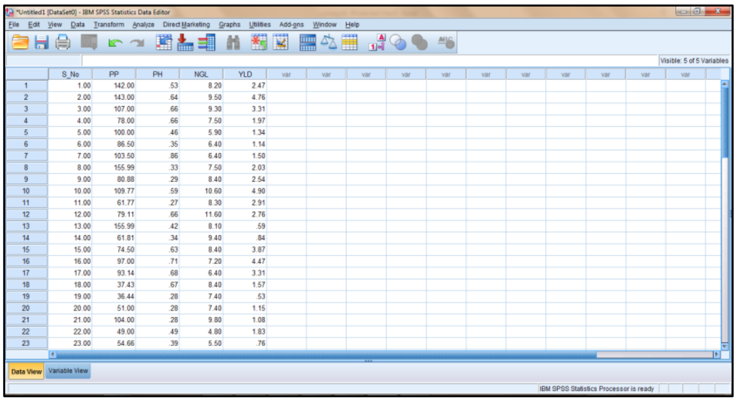
S.No.	PP	PH	NGL	Yield	S.No.	PP	PH	NGL	Yield
1	142.00	0.525	8.2	2.470	24	55.55	0.265	5.0	0.430
2	143.00	0.640	9.5	4.760	25	88.44	0.980	5.0	4.080
3	107.00	0.660	9.3	3.310	26	99.55	0.645	9.6	2.830
4	78.00	0.660	7.5	1.970	27	63.99	0.635	5.6	2.570

5	100.00	0.460	5.9	1.340	28	101.77	0.290	8.2	7.420
6	86.50	0.345	6.4	1.140	29	138.66	0.720	9.9	2.620
7	103.50	0.860	6.4	1.500	30	90.22	0.630	8.4	2.000
8	155.99	0.330	7.5	2.030	31	76.92	1.250	7.3	1.990
9	80.88	0.285	8.4	2.540	32	126.22	0.580	6.9	1.360
10	109.77	0.590	10.6	4.900	33	80.36	0.605	6.8	0.680
11	61.77	0.265	8.3	2.910	34	150.23	1.190	8.8	5.360
12	79.11	0.660	11.6	2.760	35	56.50	0.355	9.7	2.120
13	155.99	0.420	8.1	0.590	36	136.00	0.590	10.2	4.160
14	61.81	0.340	9.4	0.840	37	144.50	0.610	9.8	3.120
15	74.50	0.630	8.4	3.870	38	157.33	0.605	8.8	2.070
16	97.00	0.705	7.2	4.470	39	91.99	0.380	7.7	1.170
17	93.14	0.680	6.4	3.310	40	121.50	0.550	7.7	3.620
18	37.43	0.665	8.4	1.570	41	64.50	0.320	5.7	0.670
19	36.44	0.275	7.4	0.530	42	116.00	0.455	6.8	3.050
20	51.00	0.280	7.4	1.150	43	77.50	0.720	11.8	1.700
21	104.00	0.280	9.8	1.080	44	70.43	0.625	10.0	1.550
22	49.00	0.490	4.8	1.830	45	133.77	0.535	9.3	3.280
23	54.66	0.385	5.5	0.760	46	89.99	0.490	9.8	2.690

Source: Design Resources Server. Indian Agricultural Statistics Research Institute(ICAR), New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on <05-05-2015>).

1. Find mean, standard deviation, minimum and maximum values of all the characters.
2. Find correlation coefficient between each pair of the variables.
3. Give a scatter plot of the variable PP with dependent variable yield.
4. Fit a multiple linear regression equation where yield is dependent variable whereas all other characters as independent variables.

At first enter the entire data in the data editor as given below,

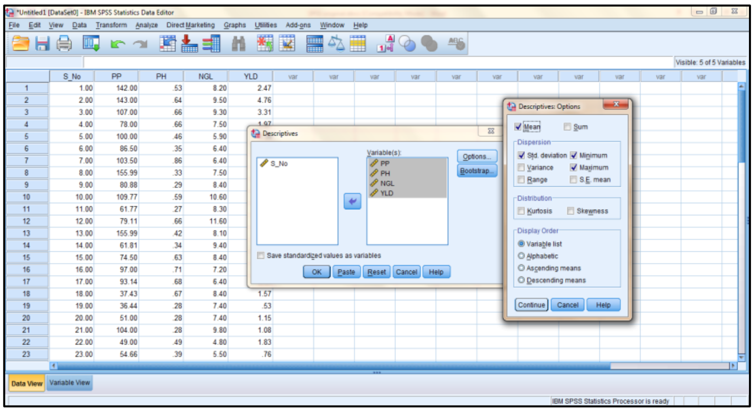


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	4.76
3	3.00	107.00	.66	9.30	3.31
4	4.00	78.00	.66	7.50	1.97
5	5.00	100.00	.46	5.90	1.34
6	6.00	86.50	.35	6.40	1.14
7	7.00	103.50	.86	6.40	1.50
8	8.00	155.99	.33	7.50	2.03
9	9.00	80.88	.29	8.40	2.54
10	10.00	109.77	.59	10.60	4.90
11	11.00	61.77	.27	8.30	2.91
12	12.00	79.11	.66	11.60	2.76
13	13.00	155.99	.42	8.10	.59
14	14.00	61.81	.34	9.40	.84
15	15.00	74.50	.63	8.40	3.87
16	16.00	97.00	.71	7.20	4.47
17	17.00	93.14	.68	6.40	3.31
18	18.00	37.43	.67	8.40	1.57
19	19.00	36.44	.28	7.40	.53
20	20.00	51.00	.28	7.40	1.15
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

There are several ways to answer Q no. 1 in SPSS. Commands following first way is as follows,

Analyze → Descriptive Statistics → Descriptives…→ Put PP, PH, NGL, YLD in the variables list→ Choose appropriate options from Options tab→PressContinue→Ok



Descriptives

Variable(s): PP, PH, NGL, YLD

□ Save standardized values as variables

Descriptives: Options

☑ Mean □ Sum

Dispersion

☑ Std. deviation ☑ Minimum

□ Variance ☑ Maximum

□ Range □ S.E. mean

Distribution

□ Kurtosis □ Skewness

Display Order

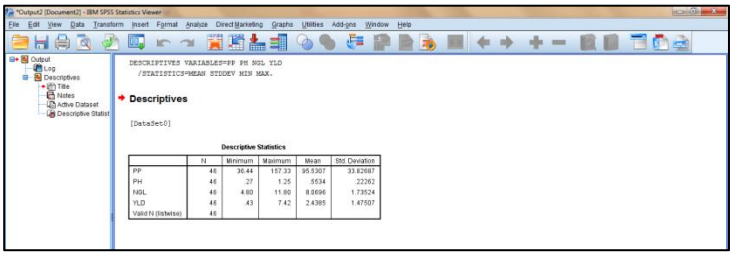
● Variable list

○ Alphabetic

○ Ascending means

○ Descending means

Output:



```
DESCRIPTIVES VARIABLES=PP PH NGL YLD
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

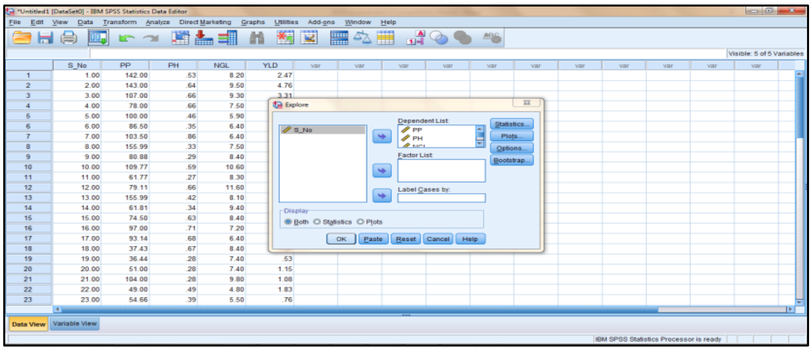
[DataSet0]

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PP	46	36.44	157.33	95.5307	33.82687
PH	46	.27	1.25	.5534	.22262
NGL	46	4.80	11.80	8.0696	1.73524
YLD	46	.43	7.42	2.4385	1.47507
Valid N (listwise)	46				

Another way:

Analyze → Descriptive Statistics → Explore…→ Put PP, PH, NGL, YLD in the Dependent list→ Choose both Statistics and plot→Press Ok



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 5 of 5 Variables
Explore
Statistics...
Plots...
Options...
Bootstrap...
OK Paste Reset Cancel Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	4.76
3	3.00	107.00	.66	9.30	3.31
4	4.00	78.00	.66	7.50	
5	5.00	100.00	.46	5.90	
6	6.00	86.50	.35	6.40	
7	7.00	103.50	.86	6.40	
8	8.00	155.99	.33	7.50	
9	9.00	80.88	.29	8.40	
10	10.00	109.77	.59	10.60	
11	11.00	61.77	.27	8.30	
12	12.00	79.11	.66	11.60	
13	13.00	155.99	.42	8.10	
14	14.00	61.81	.34	9.40	
15	15.00	74.50	.63	8.40	
16	16.00	97.00	.71	7.20	
17	17.00	93.14	.68	6.40	
18	18.00	37.43	.67	8.40	
19	19.00	36.44	.28	7.40	.53
20	20.00	51.00	.28	7.40	1.15
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

S_No

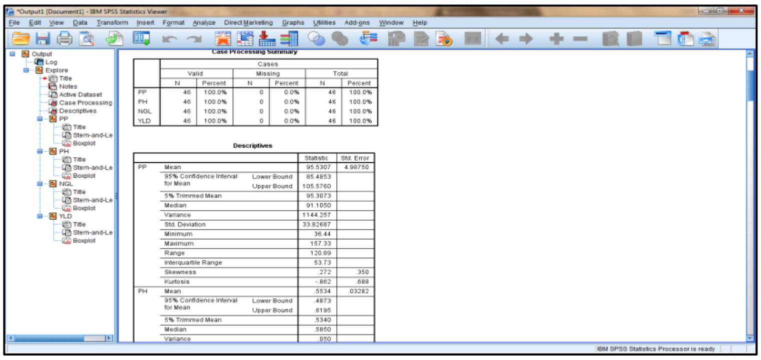
Dependent List: PP, PH, NGL

Factor List:

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Output:



*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Explore
Title
Notes
Active Dataset
Case Processing
Descriptives
PP
Title
Stem-and-Le
Boxplot
PH
Title
Stem-and-Le
Boxplot
NGL
Title
Stem-and-Le
Boxplot
YLD
Title
Stem-and-Le
Boxplot
IBM SPSS Statistics Processor is ready

Case Processing Summary

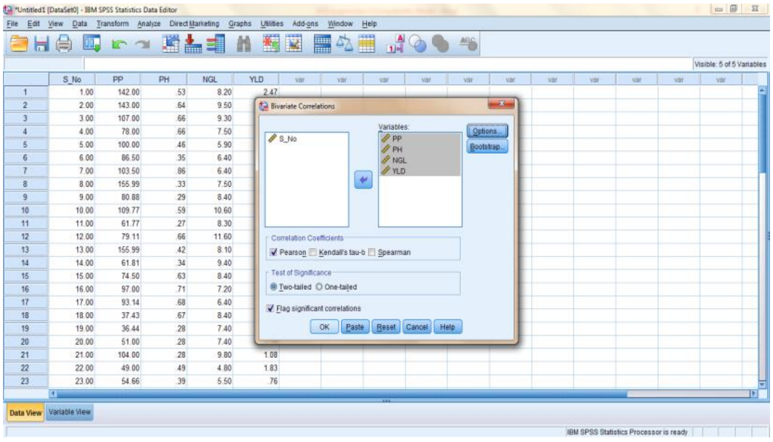
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
PP	46	100.0%	0	0.0%	46	100.0%
PH	46	100.0%	0	0.0%	46	100.0%
NGL	46	100.0%	0	0.0%	46	100.0%
YLD	46	100.0%	0	0.0%	46	100.0%

Descriptives

			Statistic	Std. Error
PP	Mean		95.5307	4.98750
	95% Confidence Interval for Mean	Lower Bound	85.4853	
		Upper Bound	105.5760	
	5% Trimmed Mean		95.3073	
	Median		91.1050	
	Variance		1144.257	
	Std. Deviation		33.82687	
	Minimum		36.44	
	Maximum		157.33	
	Range		120.89	
	Interquartile Range		53.73	
	Skewness		.272	.350
	Kurtosis		-.862	.688
PH	Mean		.5534	.03282
	95% Confidence Interval for Mean	Lower Bound	.4873	
		Upper Bound	.6195	
	5% Trimmed Mean		.5340	
	Median		.5850	
	Variance		.050	

To answer Q no 2 follow the following steps

Analyze → Correlate → Bivariate→ Put PP, PH, NGL, YLD in the Valiables list→ Choose Pearson's correlation coefficient→Press Ok



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 5 of 5 Variables
Bivariate Correlations
Options...
Bootstrap...
OK Paste Reset Cancel Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	
3	3.00	107.00	.66	9.30	
4	4.00	78.00	.66	7.50	
5	5.00	100.00	.46	5.90	
6	6.00	86.50	.35	6.40	
7	7.00	103.50	.86	6.40	
8	8.00	155.99	.33	7.50	
9	9.00	80.88	.29	8.40	
10	10.00	109.77	.59	10.60	
11	11.00	61.77	.27	8.30	
12	12.00	79.11	.66	11.60	
13	13.00	155.99	.42	8.10	
14	14.00	61.81	.34	9.40	
15	15.00	74.50	.63	8.40	
16	16.00	97.00	.71	7.20	
17	17.00	93.14	.68	6.40	
18	18.00	37.43	.67	8.40	
19	19.00	36.44	.28	7.40	
20	20.00	51.00	.28	7.40	
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

S_No

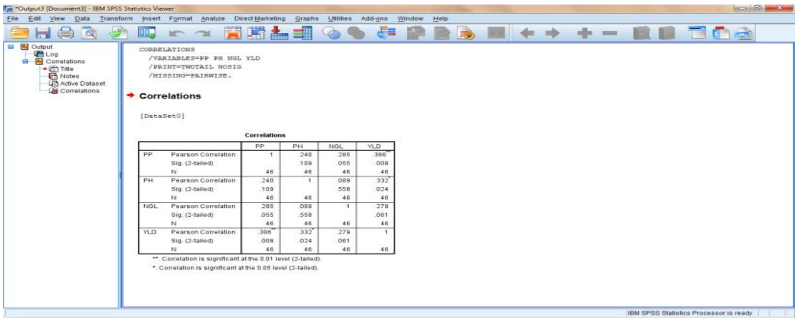
Variables: PP, PH, NGL, YLD

Correlation Coefficients: ☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

Output:



*Output3 [Document3] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Correlations
Title
Notes
Active Dataset
Correlations
IBM SPSS Statistics Processor is ready

```
CORRELATIONS
  /VARIABLES=PP PH NGL YLD
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

```
[DataSet0]
```

Correlations

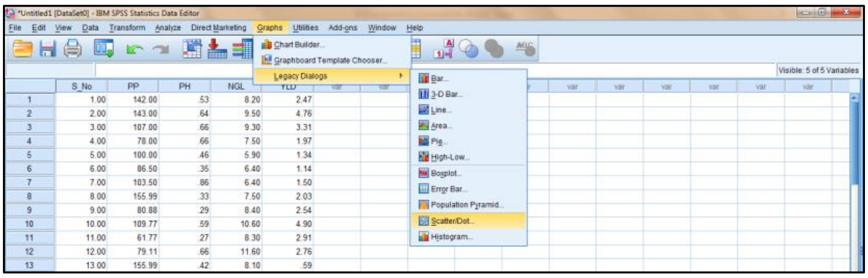
		PP	PH	NGL	YLD
PP	Pearson Correlation	1	.240	.285	.386**
	Sig. (2-tailed)		.109	.055	.008
	N	46	46	46	46
PH	Pearson Correlation	.240	1	.089	.332*
	Sig. (2-tailed)	.109		.558	.024
	N	46	46	46	46
NGL	Pearson Correlation	.285	.089	1	.279
	Sig. (2-tailed)	.055	.558		.061
	N	46	46	46	46
YLD	Pearson Correlation	.386**	.332*	.279	1
	Sig. (2-tailed)	.008	.024	.061	
	N	46	46	46	46

**. Correlation is significant at the 0.01 level (2-tailed).

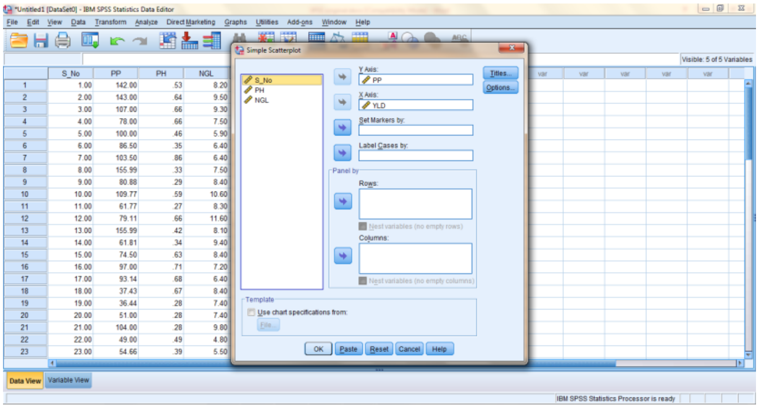
*. Correlation is significant at the 0.05 level (2-tailed).

To give the scatter plot of the variable PP with dependent variable yield use following steps:

Graphs → Legacy dialogs→ Scatterplot→ Put PP at Y axis and YLD at X axis→ Press Ok


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Chart Builder...
Graphboard Template Chooser...
Legacy Dialogs
Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...
Visible: 5 of 5 Variables

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	2.47
2	2.00	143.00	.64	9.50	4.76
3	3.00	107.00	.66	9.30	3.31
4	4.00	78.00	.66	7.50	1.97
5	5.00	100.00	.46	5.90	1.34
6	6.00	86.50	.35	6.40	1.14
7	7.00	103.50	.86	6.40	1.50
8	8.00	155.99	.33	7.50	2.03
9	9.00	80.88	.29	8.40	2.54
10	10.00	109.77	.59	10.60	4.90
11	11.00	61.77	.27	8.30	2.91
12	12.00	79.11	.66	11.60	2.76
13	13.00	155.99	.42	8.10	.59


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Simple Scatterplot
S_No
PH
NGL
Titles...
Options...
Panel by
Nest variables (no empty rows)
Nest variables (no empty columns)
Template
File...
OK
Paste
Reset
Cancel
Help
Visible: 5 of 5 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

Y Axis: PP

X Axis: YLD

Set Markers by:

Label Cases by:

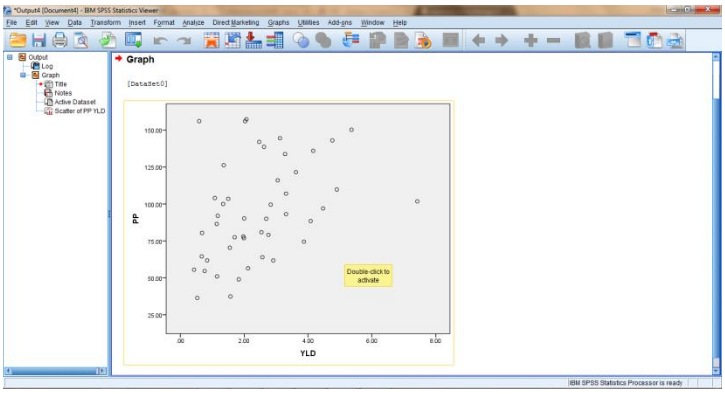
Rows:

Columns:

□ Use chart specifications from:

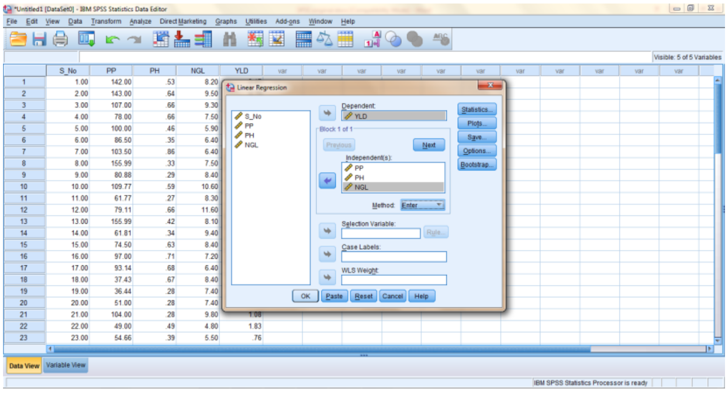
	S_No	PP	PH	NGL
1	1.00	142.00	.53	8.20
2	2.00	143.00	.64	9.50
3	3.00	107.00	.66	9.30
4	4.00	78.00	.66	7.50
5	5.00	100.00	.46	5.90
6	6.00	86.50	.35	6.40
7	7.00	103.50	.86	6.40
8	8.00	155.99	.33	7.50
9	9.00	80.88	.29	8.40
10	10.00	109.77	.59	10.60
11	11.00	61.77	.27	8.30
12	12.00	79.11	.66	11.60
13	13.00	155.99	.42	8.10
14	14.00	61.81	.34	9.40
15	15.00	74.50	.63	8.40
16	16.00	97.00	.71	7.20
17	17.00	93.14	.68	6.40
18	18.00	37.43	.67	8.40
19	19.00	36.44	.28	7.40
20	20.00	51.00	.28	7.40
21	21.00	104.00	.28	9.80
22	22.00	49.00	.49	4.80
23	23.00	54.66	.39	5.50

Output:


*Output4 [Document4] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Graph
Title
Notes
Active Dataset
Scatter of PP YLD
Graph
[DataSet0]
150.00
125.00
100.00
75.00
50.00
25.00
PP
.00
2.00
4.00
6.00
8.00
YLD
Double-click to activate
IBM SPSS Statistics Processor is ready

To fit a multiple linear regression equation taking yield as dependent variable and all other characters as independent variables perform following steps

Analyze → Regression → Linear → Put Yld in Dependent variable and PP, PH, NGL in independent variable list → Press Ok


File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

Visible: 5 of 5 Variables

	S_No	PP	PH	NGL	YLD
1	1.00	142.00	.53	8.20	
2	2.00	143.00	.64	9.50	
3	3.00	107.00	.66	9.30	
4	4.00	78.00	.66	7.50	
5	5.00	100.00	.46	5.90	
6	6.00	86.50	.35	6.40	
7	7.00	103.50	.86	6.40	
8	8.00	155.99	.33	7.50	
9	9.00	80.88	.29	8.40	
10	10.00	109.77	.59	10.60	
11	11.00	61.77	.27	8.30	
12	12.00	79.11	.66	11.60	
13	13.00	155.99	.42	8.10	
14	14.00	61.81	.34	9.40	
15	15.00	74.50	.63	8.40	
16	16.00	97.00	.71	7.20	
17	17.00	93.14	.68	6.40	
18	18.00	37.43	.67	8.40	
19	19.00	36.44	.28	7.40	
20	20.00	51.00	.28	7.40	
21	21.00	104.00	.28	9.80	1.08
22	22.00	49.00	.49	4.80	1.83
23	23.00	54.66	.39	5.50	.76

Linear Regression

Dependent: YLD

Block 1 of 1

Independent(s): PP, PH, NGL

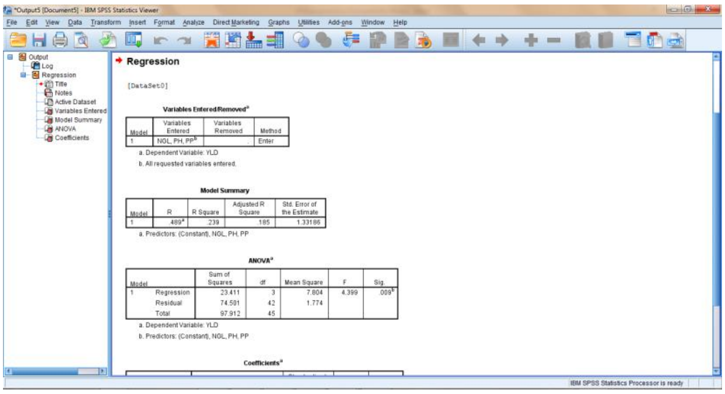
Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Output:


File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Regression
Title
Notes
Active Dataset
Variables Entered
Model Summary
ANOVA
Coefficients
IBM SPSS Statistics Processor is ready

*Output5 [Document5] - IBM SPSS Statistics Viewer

Regression

[DataSet0]

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	NGL, PH, PP[b]	.	Enter

a. Dependent Variable: YLD

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.489[a]	.239	.185	1.33186

a. Predictors: (Constant), NGL, PH, PP

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.411	3	7.804	4.399	.009[b]
	Residual	74.501	42	1.774		
	Total	97.912	45			

a. Dependent Variable: YLD

b. Predictors: (Constant), NGL, PH, PP

Coefficients[a]

Exercise 2. An experiment was conducted to study the hybrid seed production of ottle gourd under open field conditions. The main aim of the investigation was to compare natural pollination. The pollination is performed at noon (1-3pm)} under field conditions. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination. The data were collected on number of fruit set for the period of 45 days, fruit weight (kg), seed yield per plant (g) and seedling length (cm). The data obtained is as given below:

Group	No. of fruit Set(45days)	Fruit weight (kg)	Seed yield/plant (g)	Seedling length (cm)
1	8	2.0	148.6	17.0
1	7	1.9	137.7	16.9
1	6	1.8	150.9	16.4
1	8	1.9	173.4	18.4
1	7	1.8	145.3	18.0
1	8	1.9	139.1	17.1
1	7	1.9	151.5	18.3
1	7	1.8	141.8	19.0
1	6	1.9	141.4	18.5
1	7	1.9	139.2	18.7
2	6.3	2.6	225.6	18.3
2	6.7	2.8	198.7	18.2

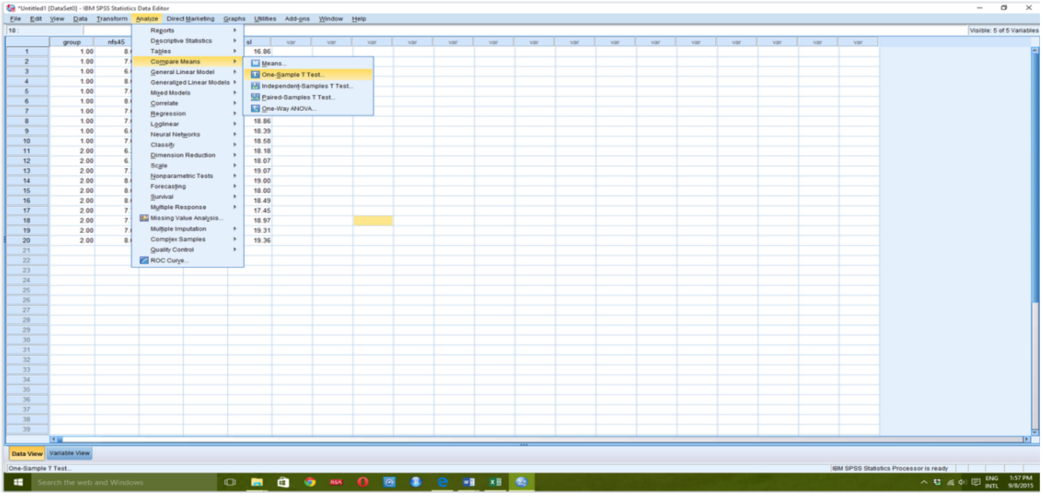
2	**7.3**	**2.6**	**231.7**	**19.2**
2	**8**	**2.6**	**218.4**	**19.1**
2	**8**	**2.7**	**235.2**	**18.1**
2	**8**	**2.6**	**217.8**	**18.6**
2	**7.7**	**2.4**	**213.2**	**17.6**
2	**7.7**	**2.7**	**211.6**	**19.1**
2	**7**	**2.5**	**201.1**	**19.4**
2	**8**	**2.5**	**215.6**	**19.5**

1. Test whether the mean of the population of Seed yield/plant (g) is 200 or not.

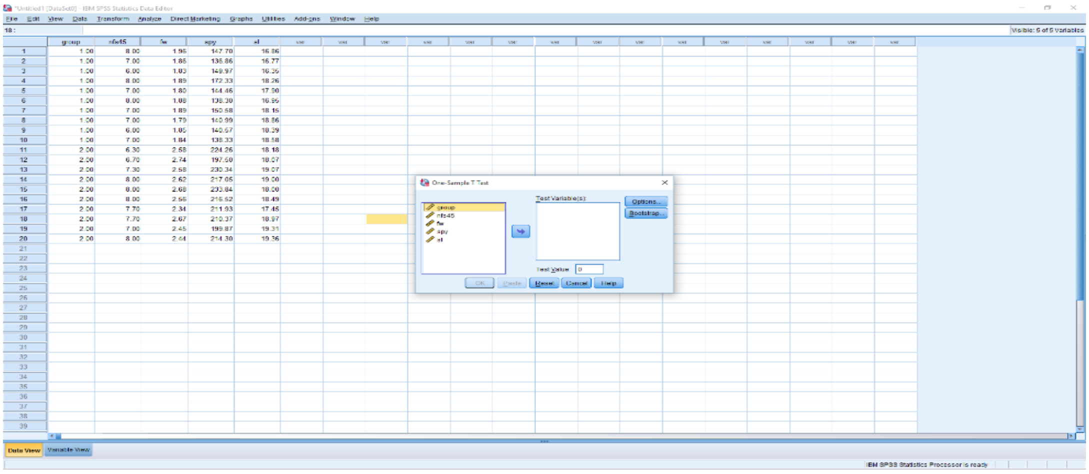
2. Test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different.

Test Procedure in SPSS

1. To test whether the mean of the population of Seed yield/plant (g) is 200 or not use the following steps. Select **Analyze → Compare Means → One-Sample T Test**



This selection displays the following screen


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 5 of 5 Variables
One-Sample T Test
group
nfs45
fw
spy
sl
Test Variable(s):
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

Test Value: 0

	group	nfs45	fw	spy	sl
1	1.00	8.00	1.96	147.70	16.86
2	1.00	7.00	1.86	136.86	16.77
3	1.00	6.00	1.83	149.97	16.35
4	1.00	8.00	1.89	172.33	18.26
5	1.00	7.00	1.80	144.46	17.90
6	1.00	8.00	1.88	138.30	16.95
7	1.00	7.00	1.89	150.58	18.15
8	1.00	7.00	1.79	140.99	18.86
9	1.00	6.00	1.85	140.57	18.39
10	1.00	7.00	1.84	138.33	18.58
11	2.00	6.30	2.58	224.26	18.18
12	2.00	6.70	2.74	197.50	18.07
13	2.00	7.30	2.58	230.34	19.07
14	2.00	8.00	2.62	217.05	19.00
15	2.00	8.00	2.68	233.84	18.00
16	2.00	8.00	2.56	216.52	18.49
17	2.00	7.70	2.34	211.93	17.45
18	2.00	7.70	2.67	210.37	18.97
19	2.00	7.00	2.45	199.87	19.31
20	2.00	8.00	2.44	214.30	19.36

Select syp and send it to the test variable(s): box and define the Test Value as 200. Click ok.

2. To Test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different.

Steps:

1. select**Analyze → Compare Means → Independent-Samples T Test.**

2. Select group and send it to the Grouping Variables box.

3. nfs45, fw, syp, sl under Test Variables(s) box.

4. Select Define Groups in the Independent-Samples T Test dialog box.

5. **Use Specified values**→ Define Groups as 1 and 2.

6. Click **OK.**

REFERENCES:

1. Design Resources Server. Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India. https://drs.icar.gov.in/

2. Morgan, G.A., Barrett, K. C. Leech, N.L.andGloecknerG.W. (2019).IBM SPSS for Introductory Statistics: Use and Interpretation. Sixth Edition, Routledge.

3. Nie, N. H., Bent, D. H. and Hull, C. H.(1970). SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill.