Data preparation, basic statistical tools for analysis of experimental data

Dr Rejula K
Scientist, ICAR-CIFT, Cochin
rejula.iari09@gmail.com

Data preparation

The application of statistical methods is an important aspect that has a role in every step of research from planning till publication. Once data collection is over, the next logical step the researcher undertakes is to analyze them to draw meaningful conclusions. Before the actual statistical analysis is carried out, one has to look hard at the data and prepare it to facilitate further analysis. Data are entered into spread sheets from which it can be imported into many statistical software packages for analysis. Errors can creep into the master chart while entering the data. So a logical check should initially be carried out to spot the errors (Manikandan, 2010).

Example of Errors: blood group C, age - 260 years, pregnant male, etc

Data preparation-- Definition

Data Preparation involves checking or logging the data in; checking the data for accuracy; entering the data into the computer; transforming the data; and developing and documenting a database structure that integrates the various measures (**Web centre for social research methods**). Data Preparation is the process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis. (https://en.wikipedia.org/wiki/Data_preparation). **Data preparation** is the act of preparing (or pre-processing) raw data or disparate data sources into refined information assets that can be used effectively for various business purposes, such as analysis.

Data cleansing is one of the most common tasks in Data preparation. Common data cleansing activities involve ensuring the data is valid, complete consistent, uniform and accurate.

- **Valid** –data has correct data type, matches required patterns no cross-field issues
- **Complete** – ensuring all necessary data is available and where possibly, looking up needed data from external sources (e.g. finding the Zip/Postal code of an address via an external data source)
- **Consistent** – eliminating contradictions in the data
- **Uniform** –e.g. uniform data/time formats across fields, uniform units of measure for weights, lengths
- **Accurate** – where possible ensuring data is verifiable with an authoritative source
- **Outliers-** By looking at the measures of central tendency (mean, median) and range, one can get a general idea about the presence of outliers in a given data set. Outliers can also arise due to experimental / observational error or if the observation was made from a population other than from which the rest of data was collected.

Data preparation consist of three processes like, data cleaning, creating necessary variables, and formatting all variables (Karen Grace-Martin)

1. Data Cleaning

Data cleaning means finding and eliminating errors in the data, like

- Impossible or otherwise incorrect values for specific variables

- Cases in the data who met exclusion criteria and shouldn't be in the study
- Duplicate cases
- Missing data and outliers
- Skip-pattern or logic breakdowns

2. Creating New Variables

Once the data are free of errors, set up the variables that will directly answer research questions.

3. Formatting the Variables

Both original and newly created variables need to be formatted. Failing to format a missing value code or a dummy variable correctly will have major consequences for data analysis.

Data transformation

Many biological variables fit the normal probability distribution quite well, which is a result of the central limit theorem (when a large number of random numbers are taken, the means of those numbers are approximately normally distributed). Most tests for measurement variables assume that data are normally distributed (fit a bell-shaped curve). Parametric tests (eg ANOVA) assume that the data can be described by two parameters, the mean and standard deviation and data fit the normal distribution. Therefore if measurement variable is not normally distributed, there is more chance of a false positive result while analyzing the data with a test that assumes normality (John H. McDonald).

If a measurement variable does not fit a normal distribution or has greatly different standard deviations in different groups, data need to be transformed. In some cases, transforming the data will make it fit the assumptions better. To transform data, a mathematical operation on each observation need to be performed, and these transformed numbers can be used in the selected statistical test. There are many number of transformations possible, but it is better to use a transformation that other researchers commonly use in similar fields. Square-root transformation is commonly used for count data and the log transformation is commonly used for size data. If large numbers of observations are present, effects of different transformations on the normality need to be compared (John H.McDonald).

Common transformations

Transformations that are used occasionally in biology are Log transformation, Square-root transformation and Arcsine transformation.

Log transformation. This consists of taking the log of each observation; either base-10 logs or base-e logs, also known as natural logs. It makes no difference for a statistical test whether we use base-10 logs or natural logs, because they differ by a constant factor; the base-10 log of a number is just 2.303...× the natural log of the number. Log used should be mentioned while writing up the results, The back transformation is to raise 10 or e to the power of the number; If some of the counts are zero, the convention is to add 0.5 to each number.

Square-root transformation

This consists of taking the square root of each observation. The back transformation is to square the number. If there is negative numbers, add a constant to each number to make them all positive. People

often use the square-root transformation when the variable is a count of something, such as bacterial colonies per petri dish, blood cells going through a capillary per minute, mutations per generation, etc.

Arcsine transformation. This consists of taking the arcsine of the square root of a number. (The result is given in radians, not degrees, and can range from $-\pi/2$ to $\pi/2$.) The numbers to be arcsine transformed must be in the range 0 to 1. This is commonly used for proportions, which range from 0 to 1.The back-transformation is to square the sine of the number.

Data Preparation in SPSS

Assignment of Variable properties

Variable properties can be assigned to the data entered in SPSS, within Variable View of the Data Editor. The Variable view contains descriptions of the attributes of each variable in the data file. In the Variable view rows are variables and columns are variable attributes. Variables can be added or deleted and the following attributes (Variable name, Data type Descriptive variable and value labels, User-defined missing values and Measurement level) of variables can be modified

Variable name The following rules apply to variable names:

- The name must begin with a letter. The remaining characters can be any letter, any digit, a period, or the symbols @, #, _, or $.
- Variable names cannot end with a period.
- Variable names that end with an underscore should be avoided (to avoid conflict with variables automatically created by some procedures).
- The length of the name cannot exceed eight characters.
- Blanks and special characters (for example, !, ?, ', and *) cannot be used.
- Each variable name must be unique; duplication is not allowed. Variable names are not case sensitive

Data type

Numeric. A variable whose values are numbers.

String. Values of a string variable are not numeric, and hence not used in calculations. They can contain any characters up to the defined length. Uppercase and lowercase letters are considered distinct. Also known as an alphanumeric variable

Descriptive variable and value labels descriptive value labels can be assigned for each value of a variable. This is particularly useful if data file uses numeric codes to represent non-numeric categories (for example, codes of 1 and 2 for male and female). Value labels can be up to 60 characters long. Value labels are not available for long string variables (string variables longer than 8 characters).

User-defined missing values

Data values specified as user-missing are flagged for special treatment and are excluded from most calculations. Up to three values can be entered; discrete (individual) missing values, a range of missing values, or a range plus one discrete value. Ranges can be specified only for numeric variables. All string values, including null or blank values, are considered valid values unless they are explicitly defined as missing. To define null or blank values as missing for a string variable, enter a single space in one of the fields for Discrete missing values.

Measurement level

One of three measurement levels can be selected.

Scale. Data values are numeric values on an interval or ratio scale--for example, age or income. Scale variables must be numeric.

Ordinal Data values represent categories with some intrinsic order (for example, low, medium, high; strongly agree, agree, disagree, strongly disagree). Ordinal variables can be either string (alphanumeric) or numeric values that represent distinct categories (for example, 1 = low, 2 = medium, 3 = high).

Nominal. Data values represent categories with no intrinsic order--for example, job category or company division. Nominal variables can be either string (alphanumeric) or numeric values that represent distinct categories--for example, 1 = Male, 2 = Female.

Defining Missing Values

All string values, including null or blank values, are considered valid values it is explicitly defined as missing. Missing or invalid data are very common and cannot be ignored. Survey respondents may refuse to answer certain questions, may not know the answer, or answer in a format not expected. If Missing or invalid data are not filtered or identified data analysis may not provide accurate results.

Basic statistical tools in SPSS

Transform menu

Transform menu is used to calculate new values and variables, and to recode data,

Recode into Same Variables

Recode into Same Variables reassigns the values of existing variables or collapses ranges of existing values into new values.

Count Occurrences of Values within Cases

This dialog box creates a variable that counts the occurrences of the same value(s) in a list of variables for each case.

Categorize Variables

Categorize Variables convert's continuous numeric data to a discrete number of categories. The procedure creates new variables containing the categorical data. Data are categorized based on percentile groups, with each group containing approximately the same number of cases. For example, a specification of 4 groups would assign a value of 1 to cases below the 25th percentile, 2 to cases between the 25th and 50th percentile, 3 to cases between the 50th and 75th percentile, and 4 to cases above the 75th percentile.

Rank Cases

Rank Cases creates new variables containing ranks, normal and Savage scores, and percentile values for numeric variables. New variable names and descriptive variable labels are automatically generated, based on the original variable name and the selected measure(s). A summary table lists the original variables, the new variables, and the variable labels.

Functions

Many types of functions are supported in SPSS, including:

- Arithmetic functions
- Statistical functions
- String functions
- Date and time functions
- Distribution functions
- Random variable functions
- Missing value functions

Arithmetic Functions

- ABS(numexpr) Numeric. Returns the absolute value of numexpr, which must be numeric.
- ARSIN(numexpr) Numeric. Returns the inverse sine, in radians, of numexpr, which must evaluate to a numeric value between -1 and +1.
- EXP(numexpr) Numeric. Returns e raised to the power numexpr, where e is the base of the natural logarithms and numexpr is numeric. Large values of numexpr may produce results that exceed the capacity of the machine.
- LN(numexpr) Numeric. Returns the base-e logarithm of numexpr, which must be numeric and greater than 0.
- LG10(numexpr) Numeric. Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.
- SQRT(numexpr) Numeric. Returns the positive square root of numexpr, which must be numeric and not negative.
- TRUNC(numexpr) Numeric. Returns the value of numexpr truncated to an integer (toward 0).

Statistical Functions

- CFVAR(numexpr,numexpr[,...]) Numeric. Returns the coefficient of variation (the standard deviation divided by the mean) of its arguments that have valid values. This function requires two or more arguments, which must be numeric. You can specify a minimum number of valid arguments for this function to be evaluated.
- MAX(value,value[,...]) Numeric or string. Returns the maximum value of its arguments that have valid values. This function requires two or more arguments. You can specify a minimum number of valid arguments for this function to be evaluated.
- MEAN(numexpr,numexpr[,...]) Numeric. Returns the arithmetic mean of its arguments that have valid values. This function requires two or more arguments, which must be numeric. You can specify a minimum number of valid arguments for this function to be evaluated.
- MIN(value,value[,...]) Numeric or string. Returns the minimum value of its arguments that have valid values. This function requires two or more arguments. You can specify a minimum number of valid arguments for this function to be evaluated.
- SD(numexpr,numexpr[,...]) Numeric. Returns the standard deviation of its arguments that have valid values. This function requires two or more arguments, which must be numeric. You can specify a minimum number of valid arguments for this function to be evaluated.
- SUM(numexpr,numexpr[,...]) Numeric. Returns the sum of its arguments that have valid values. This function requires two or more arguments, which must be numeric. You can specify a minimum number of valid arguments for this function to be evaluated.

- VARIANCE(numexpr,numexpr[,...]) Numeric. Returns the variance of its arguments that have valid values. This function requires two or more arguments, which must be numeric. You can specify a minimum number of valid arguments for this function to be evaluated

Simple summary measures

Different summary measures are appropriate for different types of data, depending on the level of measurement.

For categorical data, the most typical summary measure is the number or pecentage of cases in each category. The mode is the category with the greatest number of cases. For ordinal data, the median (the value above and below which half the cases fall) may also be a useful summary measure if there is a large number of categories.

The Frequencies procedure produces frequency tables that display both the number and percentage of cases for each observed value of a variable.

Crosstabulation Tables

Crosstabulation tables (contingency tables) display the relationship between two or more categorical (nominal or ordinal) variables. The size of the table is determined by the number of distinct values for each variable, with each cell in the table representing a unique combination of values.

Significance Testing for Crosstabulations

Pearson chi-square tests the hypothesis that the row and column variables are independent. The lower the significance value, the less likely it is that the two variables are independent (unrelated).

Adding a Layer Variable

Layer Variable is sometimes referred to as the control variable because it may reveal how the relationship between the row and column variables changes when we "control" for the effects of the third variable.

Creating a Categorical Variable from a Scale Variable

When a new variable is created based on the recoded values of another variable, the recoded categories should be:

Mutually exclusive. There shouldn't be any overlap in category definitions. For example, the categories 25-50 and 50-75 are not mutually exclusive, since both categories contain the value 50.

Exhaustive: There should be appropriate categories for all values of the original variable. For example, the categories 25-49 and 50-74 would not include values that might fall between 49 and 50.

Computing New Variables

New variables can be computed based on simple to highly complex equations. Example: compute a new variable that is the difference between the values of two existing variables. The new variable is displayed in the Data Editor.

Creating and Editing Charts

A wide variety of chart types are available and many of those charts are available in two different formats:

- Standard charts. Charts created from the main Graphs menu and charts created by statistical procedures.
- Interactive charts. Charts created from the Interactive sub-menu of the Graphs menu and charts created from pivot tables.

Editing Standard Charts

Charts can be edited in a variety of ways. Example, Add a title, Remove the small category of "missing" data, Display percentages

Pivot Tables

Many of the results in the Viewer are presented in tables that can be pivoted interactively. That is, we can rearrange the rows, columns, and layers.

Manipulating a Pivot Table

Options for manipulating a pivot table include:

- Transposing rows and columns
- Moving rows and columns
- Creating multidimensional layers
- Grouping and ungrouping rows and columns
- Showing and hiding cells
- Rotating row and column labels
- Finding definitions of terms

Independent-Samples T Test

The Independent-Samples T Test procedure compares means for two groups of cases. Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors.

Paired-Samples T Test

The Paired-Samples T Test procedure compares the means of two variables for a single group. It computes the differences between values of the two variables for each case and tests whether the average differs from 0.

One-Way (or Single Factor) ANOVA

One-Way ANOVA is used to Compare the means of the samples or groups in order to make inferences about the population means; Used for one independent variable with three or more levels. If more than one independent variableis involved, factorial ANOVA or higher versions are appropriate. This technique is an extension of the two-sample t test. Factor variable values should be integers, and the dependent variable should be quantitative.

Assumptions in ANOVA

- Observations are independent; Each group is an independent random sample from a normal population
- Variances on the dependent variable are equal across groups; The groups should come from populations with equal variances (To test this assumption, Levene's homogeneity-of-variance test is used)
- The dependent variable is normally distributed for each group

Bivariate Correlations

Correlations measure how variables or rank orders are related. Before calculating a correlation coefficient, data need to be checked for outliers (which can cause misleading results) and evidence of a linear relationship. Pearson's correlation coefficient is a measure of linear association. Two variables can be perfectly related, but if the relationship is not linear, Pearson's correlation coefficient is not an appropriate statistic for measuring their association.

- **Correlation Coefficients**. For quantitative, normally distributed variables,Pearson correlation coefficient is used.
- If data are not normally distributed or have ordered categories, Kendall's tau-b or Spearman, which measure the association between rank orders. Correlation coefficients range in value from -1 (a perfect negative relationship) and +1 (a perfect positive relationship). A value of 0 indicates no linear relationship. When interpreting your results, be careful not to draw any cause-and-effect conclusions due to a significant correlation.
- Test of Significance. If the direction of association is known in advance, select One-tailed. Otherwise, select Two-tailed.
- Flag significant correlations. Correlation coefficients significant at the 0.05 level are identified with a single asterisk, and those significant at the 0.01 level are identified with two asterisks

Partial Correlations

The Partial Correlations procedure computes partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables. Correlations are measures of linear association. Two variables can be perfectly related, but if the relationship is not linear, a correlation coefficient is not an appropriate statistic for measuring their association. Two or more numeric variables and one or more numeric control variables has to be selected to Obtain Partial Correlations. If the direction of association is known in advance, select One-tailed. Otherwise, select Two-tailed.

Linear Regression

Linear Regression estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. The dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.

Assumptions. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear, and all observations should be independent

Logistic Regression

Logistic regression is used to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model.

References

S Manikandan (2010). Preparing to analyse data. J Pharmacol Pharmacother. 2010 Jan-Jun; 1(1): 64–65. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3142762/

Web centre for social research methods https://www.socialresearchmethods.net/kb/statprep.php

Data preparation. From Wikipedia, the free encyclopedia. Available at. https://en.wikipedia.org/wiki/Data_preparation

Karen Grace-Martin. Preparing Data for Analysis is (more than) Half the Battle. Available at. https://www.theanalysisfactor.com/preparing-data-analysis/

John H.McDonald. Handbook of Biological Statistics. Available at http://www.biostathandbook.com/transformation.html