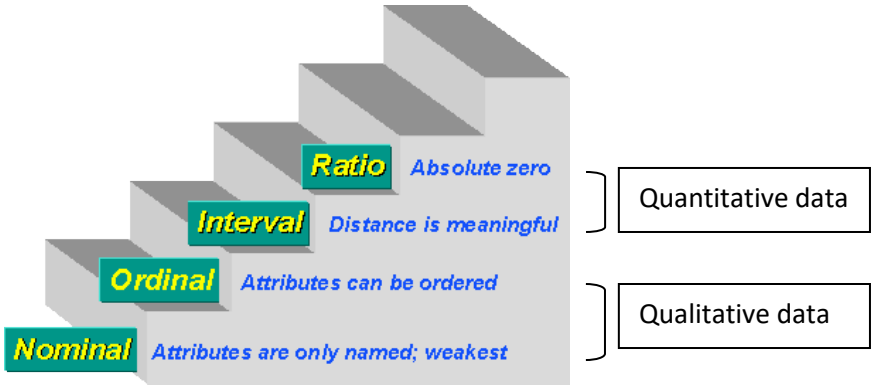
23

DESCRIPTIVE AND BASIC EXPLORATORY STATISTICS

Joshy C. G.
Fish processing division
Email ID:*cgjoshy@gmail.com*

Descriptive Statistics

Statistics is a set of procedures for gathering, measuring, classifying, computing, describing, synthesizing, analyzing, and interpreting systematically acquired data. The data can be collected either in qualitative or quantitative in nature.



Descriptive Statistics gives numerical and graphical procedures to summarize a collection of data in a clear and understandable way. Inferential statistics provides procedures to draw inferences about a population from a sample.

Types of Descriptive Statistics

1. Graphs & Frequency Distribution
   It summarize the distribution of individual observations or range of values in a given set of observations.
2. Measures of Central Tendency
   It computes the indices enabling the researcher to determine the average score of a given set of data
3. Measures of Variability
   It computes indices enabling the researcher to indicate how a given set of data spread out

Frequency Distribution

Frequency distribution organizes raw data or observations that have been collected. Frequency distribution can be computed for grouped as well ungrouped set of data.

Ungrouped Data

Listing all possible scores that occur in a distribution and then indicating how often each score occurs

Grouped Data

Combining all possible scores into classes and then indicating how often each score occurs within each class. It is easier to see patterns in the data, but lose information about individual scores.

For making a frequency table following Guidelines should be followed

- Intervals should not overlap, so no score can belong to more than one interval
- Make all intervals of the same width
- Make the intervals continuous throughout the distribution (even if an interval is empty)
- Use optimum class intervals
- Choose a convenient interval width

Graphical Display

Graphical display is used to depict certain characteristics and trends in a given set of data

Graphs for quantitative data

- Histogram
- Frequency Polygon
- Graphs for qualitative data
- Bar Chart
- Pie Chart

Histogram and Frequency Polygon

Histogram consists of a number of bars placed side by side

- The width of each bar indicates the interval size
- The height of each bar indicates the frequency of the interval
- There are no gaps between adjacent bars
- Continuous nature of quantitative data

A frequency polygon represents the shape of the data. It can be conceptualized by connecting the midpoints of the classes at the height specified by the frequency.

Bar Graph

- The qualitative data is summarized in a frequency, relative frequency, or percent frequency distribution
- On the horizontal axis, the labels used for each of the classes are specified
- On the vertical axis, frequency is specified

- The bars are separated to show that each class is a separate category

Pie Chart

- Commonly used graphical device for presenting relative frequency distributions for qualitative data
- Use the relative frequencies to subdivide a circle (360^0) into sectors that correspond to the relative frequency for each class
- A class with a relative frequency of 0.25 would take $0.25(360) = 90^0$ of the circle

Measures of Central Tendency

The central tendency of a distribution is an estimate of the 'centre' of a distribution of values of a given set of distribution. The major measures of central tendencies are

1) Mean
2) Median
3) Mode
4) Harmonic mean
5) Geometric mean

The mean is the arithmetic average of data values. It computes by adding up the observations and divide by total number of observations. It is the most commonly used measure of central tendency and it is affected by extreme values (outliers).

The median is the "middle most observation" in a given set of observations. If n is odd, the median is the middle number and if n is even, the median is the average of the 2 middle numbers. Median is not affected by extreme values.

The mode is the most frequently observation in a given set of observations. Mode is not affected by extreme values.

The harmonic mean is the average of the reciprocal of the observations

The geometric mean is the n^{th} root of the products of the observations

Averages or measure of central tendency are representatives of a frequency distribution, but they fail to give a complete picture of the distribution. Measures of central tendency do not tell anything about the scatterness of observations within the distribution.

Measures of Dispersion

Measures of Dispersion quantify the scatterness or variation of observations from their average or measures of central tendencies. It describes the spread, or dispersion, of scores in a distribution. The three most commonly used measures are

1) Range
2) Variance
3) Standard Deviation

Range is the simplest measure of variability and it is the difference between the highest and the lowest observation in a given set of data. It is very unstable and unreliable indicator.

Range= H-L

Variance measures the variability of observations from its mean. It computes the sum of squared diference between observations and mean. Standard Deviation is the square root of variance.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Measures of Relative Dispersion

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal, then their variability can be compared directly by using their S.Ds. However, if their means are widely different or if they are expressed in different units of measurement, S.Ds cannot be used as such for comparing their variability. In such situations, the relative measures of dispersions can be used.

The coefficient of variation (C.V) is a commonly used measure of relative dispersion and it is ratio of SD to the Mean multiplied by 100.

C.V. = (S.D / Mean) x 100

The C.V. is a unit-free measure and it is always expressed as percentage. The C.V. will be small if the variation is small. Of the two groups, the one with less C.V. is said to be more consistent.

Graphical Representation of the data

In a graphical representation the data is represented by symbols, such as bars in a bar chart, lines in a line chart, or slices in a pie chart. A chart can represent tabular numeric data, functions or some kinds of qualitative structures. Graphs make it easier to see certain characteristics and trends in a set of data

The Graphs for quantitative data are

- Histogram
- Frequency Polygon

The Graphs for qualitative data are

- Bar chart
- Pie chart

Histogram and Frequency Polygon

A histogram is a graphical representation showing a visual impression of the distribution of data. It is an estimate of the probability distribution of a continuous variable. A Histogram consists of a number of bars placed side by side

- The width of each bar indicates the interval size
- The height of each bar indicates the frequency of the interval
- There are no gaps between adjacent bars
- Continuous nature of quantitative data

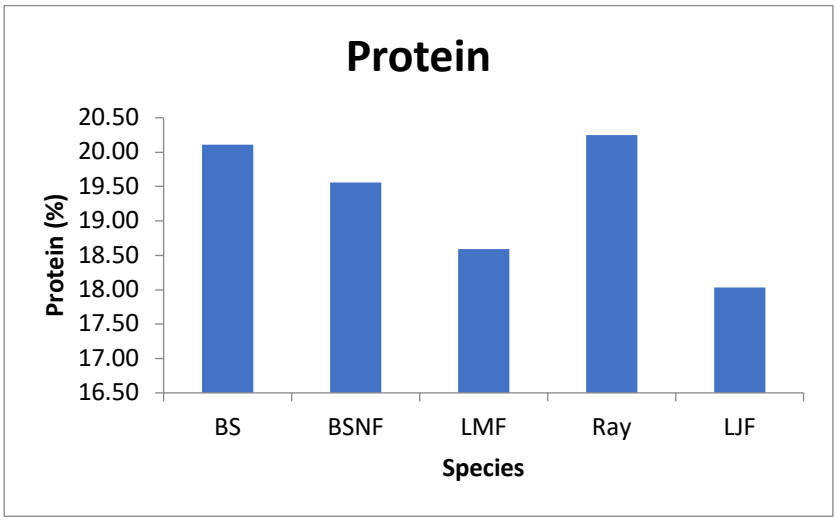
A frequency polygon represents the shape of the data. It can be conceptualized by connecting the midpoints of the classes at the height specified by the frequency.

Example of histogram

Bar Graph

A bar graph is a chart with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column bar chart.

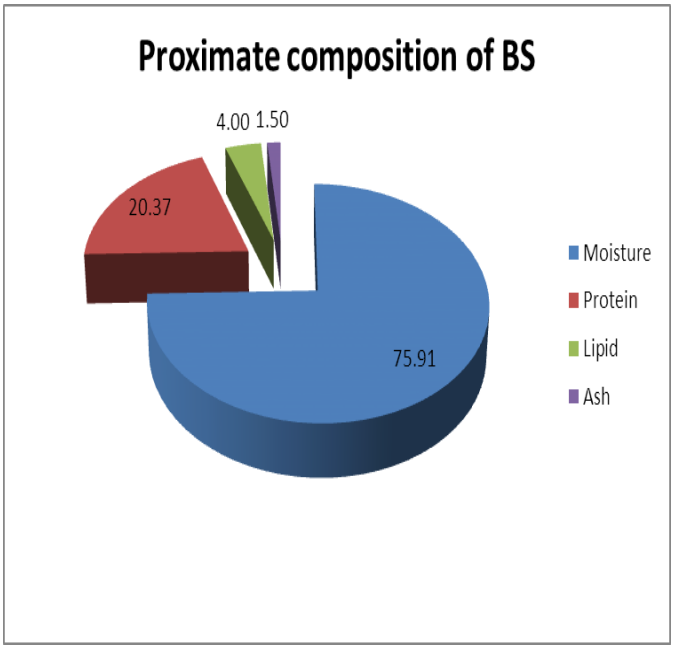
- The qualitative data is summarized in a frequency, relative frequency, or percent frequency distribution
- On the horizontal axis, the labels used for each of the classes are specified
- On the vertical axis, frequency is specified
- The bars are separated to show that each class is a separate category



Pie Chart

A pie chart (or a circle graph) is a circular chart divided into sectors, illustrating proportion. In a pie chart, the arc length of each sector (and consequently its central angle and area), is proportional to the quantity it represents.

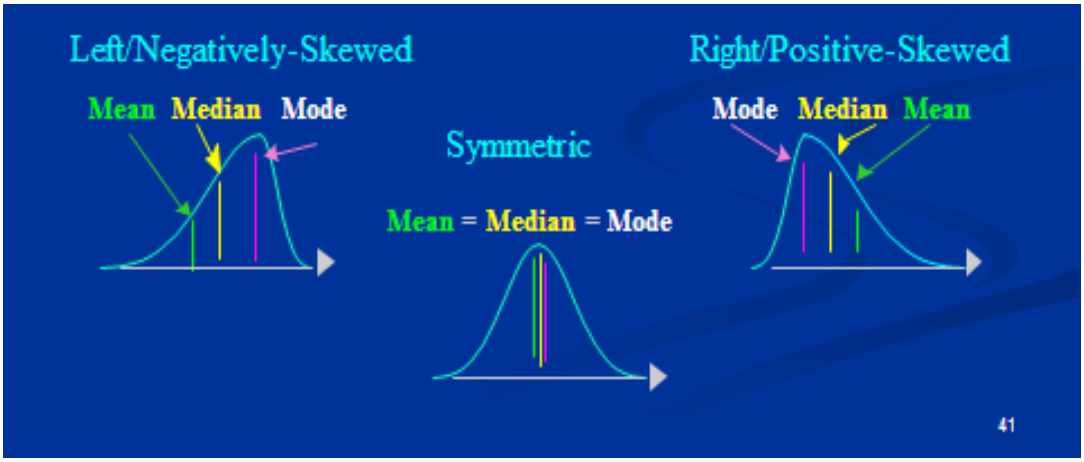
- Commonly used graphical device for presenting relative frequency distributions for qualitative data
- Use the relative frequencies to subdivide a circle (360^0) into sectors that correspond to the relative frequency for each class
- A class with a relative frequency of 0.25 would take $0.25(360) = 90^0$ of the circle

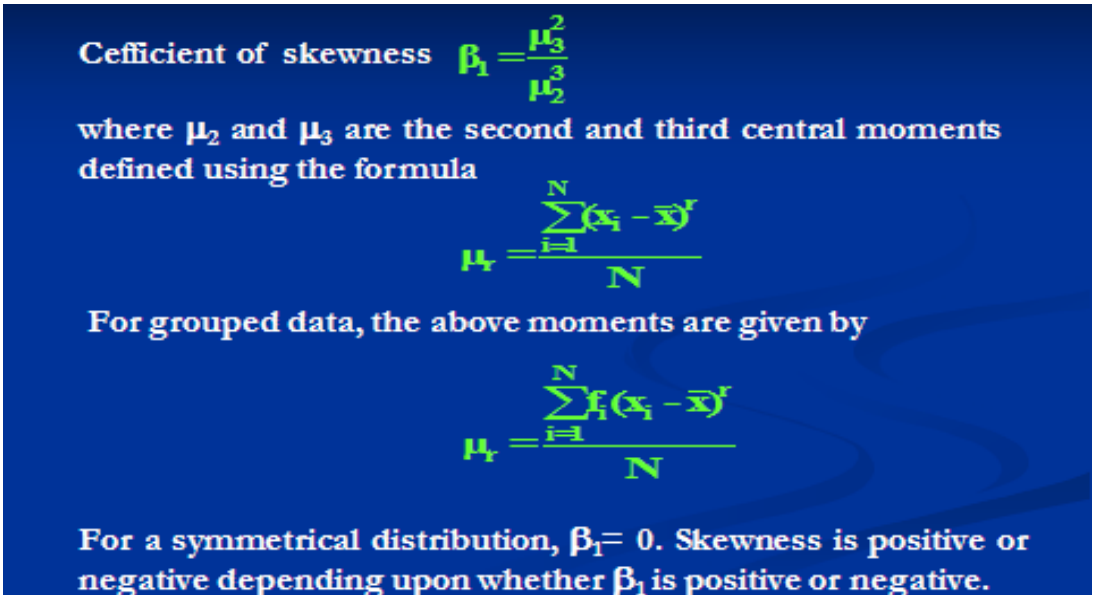


Distribution of a given data

Skewness and Kurtosis are the main statistics used to measure the shape or distribution of a given set of data.

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the *tail* on the left side of the probability density function is *longer* than the right side and the bulk of the values (possibly including the median) lie to the right of the mean. A positive skew indicates that the *tail* on the right side is *longer* than the left side and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically implying a symmetric distribution. Kurtosis measures the peakedness of shape distribution of a given set of data. The distribution is called normal if $\beta_2 = 3$; β_2 is more than 3, the distribution is said to be leptokurtic β_2 is less than 3, the distribution is said to be platykurtic (where $\boldsymbol{\beta_2 = \frac{\mu_4}{\mu_2^2}}$)





Exploratory Data Analysis

Exploratory data analysis employs a variety of techniques (mostly graphical)

- Scatter Plot
- Stem and Leaf
- Boxplot

Five Number System gives a good identification of center and spread of the data

- Maximum
- Minimum
- Median = 50th percentile
- Lower quartile Q_1 = 25th percentile
- Upper quartile Q_3 = 75th percentile

Scatter Diagram

- A graphical presentation of the relationship between two quantitative variables.
- One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.
- The general pattern of the plotted points suggests the overall relationship between the variables.

Stem-and-Leaf Display

- Shows both the rank order and shape of the distribution of the data.
- It is similar to a histogram on its side, but it has the advantage of showing the actual data values.
- The first digits of each data item are arranged to the left of a vertical line.
- To the right of the vertical line we record the last digit for each item in rank order.
- Each line in the display is referred to as a stem.

- Each digit on a stem is a leaf.
- Box Plot
- A boxplot is a graph of the five – number summary
- A central box spans the quartiles
- A line in the box marks the median
- Lines extend from the box out to the smallest and largest observations
- Boxplots can be drawn either horizontally or vertically

References

Kothari,C.R. 1990. Research Methodology: Methods and Techniques. Second edition. Wishwa prakashan. New Delhi.P 468

Snecdecor. G.W. and Cochran. W.G. 1989. Statistical Methods. John Wiley & Sons. New York. P 503.

****