Short note on Biostatistics

Statistics is the numerically stated facts like the population of a country. It can also be referred to as the science dealing with data collection, tabulation and analysis and interpretation of data. In data analysis it can be classified into descriptive statistics and inferential statistics.

Descriptive Statistics – are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. It includes frequency counts, percentages, ranges (min-max), mean, median, mode, SD etc..

Inferential Statistics – are used to draw conclusions about a population by examining the sample. Accuracy of inference depends on representativeness of sample from population. Inferential statistics help researchers to test hypotheses and answer research questions, and derive meaning from the results. Researchers set the significance level for each statistical test they conduct and by using probability theory as a basis for their tests, researchers can assess how likely it is that the difference they find is real and not due to chance (*p-value*).

Biostatistics

Statistical methods applied in medicine, biology and public health are termed as ***biostatistics***. ***Biostatistics*** is the term used when tools of statistics are applied to data that is derived from biological sciences such as medicine. It may be stated that the application of statistical methods to the solution of biological problems. Biostatistics is known by many names such as medical statistics, health statistics and vital statistics.

Medical statistics : Statistics related to clinical and laboratory parameters, their relationship, efficacy of drug, diagnostic analysis etc.

Health statistics : Statistics related to health of people in a community, epidemiology of diseases, association of occurrence of various diseases with socioeconomic and demographic variables, control and prevention of diseases etc.

Vital statistics : Statistics related to vital events in life such as of birth, death, marriages, morbidity etc. These terms are overlapping and not exclusive of each other.

Uses of Biostatistics

Statistical methods are widely used in almost all fields. Most of the basic as well as advanced statistical methods are applied in fields such as medicine, biology, public health etc.

Statistical methods are useful in planning and conducting meaningful and valid research studies on medical, health and biological problems in the population for the prevention of diseases, for finding effective appropriate treatment modalities etc. Statistical methods needed in general are,

> Collection of medical and health data scientifically

> Summarizing the collected data to make it comprehensible

> Generalizing the result from the sample to the entire population with scientific validity

> Drawing conclusions from the summarized data and generalized results.

Example :

> To determine the normal limits of various laboratory and clinical parameters such as BP, pulse rate, Cholesterol level, Blood sugar level etc.

> To find difference between means and proportions of two different groups or places or periods.

> To find correlation between variables such as cholesterol and BMI, exercise and obesity etc..

> To find action of a drug or to compare between two drugs.

> To find relative potency of a new drug with respect to a standard drug.

> To find efficacy of a line of treatment or to compare between efficacies of two different line of treatments.

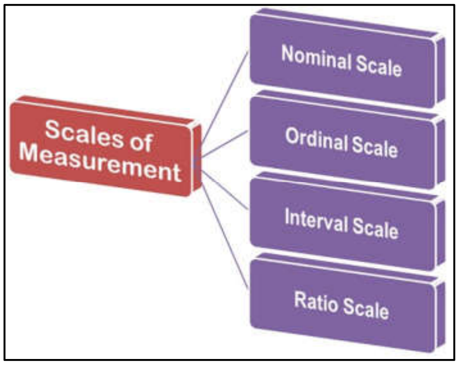
> In community medicine and public health to compare the prevalence of deaths among vaccinated and unvaccinated in a community etc..

COLLECTION AND ANALYSIS OF DATA

Levels of measurement / Scales of measurement

Level of measurement or scale of measure is a classification that describes the nature of information within the values assigned to variables. There are four scales of measures:

1. Nominal 2. Ordinal
3. Interval 4. Ratio



1. **Nominal:** A nominal variable is a categorical variable. Nominal variables have two or more categories without having any kind of natural order. They are variables with no numeric value, such as gender, occupation etc. It does not scale objects along any dimension.

   Example: Gender: (a) Male (b) Female

   Place of living: (a) Rural (b) Urban

2. **Ordinal:** The categories of the variable have an order. With ordinal scales, it is the order of the values is what's important and significant, but the numeric differences between each one is not really known. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.

Example: Stages of Cancer: (a) Stage I (b) Stage II (c) Stage III (d) Stage IV

Level of pain: (a) Mild (b) Moderate (c) Severe

3. **Interval:** Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values. Interval scales can be used for statistical analysis on the data sets like, mean, median, mode and standard deviation can be calculated.

   Example: The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

4. **Ratio:** Ratio scales are the ultimate level, when it comes to measurement scales because they tell us about the order, they tell us the exact value between units and they also have an absolute zero – which allows for a wide range of both descriptive and inferential statistics to be applied. We can express one measurement with other in a ratio like one is double the other and so on. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

   Example : Age :____, BP : ________, Weight : ________

Qualitative Data

Qualitative data arise when the observations fall into separate distinct categories.

Examples: Colour of eyes : blue, green, brown etc , Exam result : pass or fail

Socio-economic status : low, middle or high.

Such data are inherently discrete, in that there are a finite number of possible categories into which each observation may fall.

Data are classified as:

Nominal if there is no natural order between the categories (eg eye colour), or

Ordinal if an ordering exists (eg exam results, socio-economic status).

Quantitative Data

Quantitative or numerical data arise when the observations are counts or measurements. The data are said to be discrete if the measurements are integers (eg number of people in a household, number of cigarettes smoked per day) and continuous if the measurements can take on any value, usually within some range (eg weight).

Data are classified as **Interval** and **Ratio**.

Data Collection Techniques

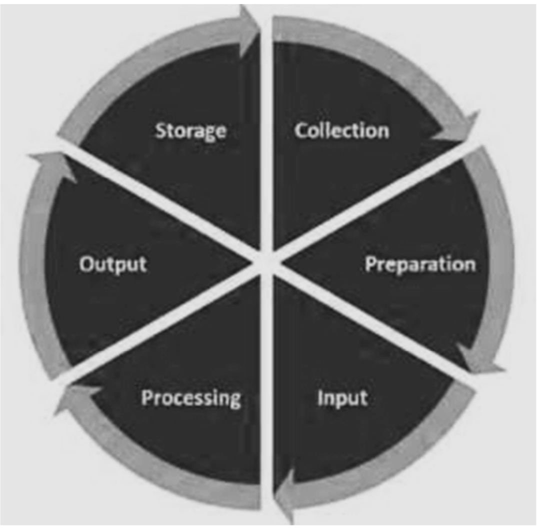
Information you gather can come from a range of sources. Likewise, there are a variety of techniques to use when gathering primary data. Listed below are some of the most common data collection techniques.

1. Interviews
2. Questionnaires and Surveys
3. Observations
4. Focus Groups
5. Case Studies
6. Documents and Records

Technique	Key Facts
Interviews	➢ Interviews can be conducted in person or over the telephone ➢ Questions should be focused, clear, and encourage open-ended responses ➢ Interviews are mainly qualitative in nature
Questionnaires and Surveys	➢ Responses can be analysed with quantitative methods by assigning numerical values to Likert-type scales ➢ Results are generally easier (than qualitative techniques) to analyse ➢ Pre-test/Post-test can be compared and analysed
Observations	➢ Allows for the study of the dynamics of a situation, frequency counts of target behaviours. ➢ Can produce qualitative (e.g., narrative data) and quantitative data (e.g., frequency counts, mean length of interactions, and instructional time)
Focus Groups	➢ A facilitated group interview with individuals that have something in common ➢ Gathers information about combined perspectives and opinions ➢ Responses are often coded into categories and analysed thematically
Case Studies	➢ Involves studying a single phenomenon ➢ Examines people in their natural settings ➢ Uses a combination of techniques such as observation, interviews, and surveys ➢ Researcher can become a confounding variable
Documents and Records	➢ Consists of examining existing data in the form of databases, meeting minutes, reports, attendance logs, financial records, newsletters, etc. ➢ This can be an inexpensive way to gather information but may be an incomplete data source

Data Processing

Data processing is the collection and manipulation of electronic data to produce meaningful information. Data processing is a form of information processing, which is the modification (processing) of information in any manner detectable by an observer. Data in its raw form is not useful to any organization. Data processing is the method of collecting raw data and translating it into usable information. It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization. The raw data is collected, filtered, sorted, processed, analyzed, stored, and then presented in a readable format.



Data Processing Cycle

Data processing may involve various processes, including:

- Validation – Ensuring that supplied data is correct and relevant.
- Sorting – Arranging items in some sequence and/or in different sets.
- Summarization (statistical) or (automatic) – Reducing detailed data to its main points.
- Aggregation – Combining multiple pieces of data.
- Analysis – The collection, organization, analysis, interpretation and presentation of data.
- Reporting – List detail or summary data or computed information.
- Classification – Separation of data into various categories.

Processing of data: Coding and Tabulation

After the data have been collected, the researcher turns to the task of analyzing them. The data, after collection, has to be processed and analyzed in accordance with the outline laid down for the purpose at the time of developing the research plan. This is essential for a scientific study and for ensuring that we have all relevant data for making contemplated comparisons and analysis.

Coding: Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness (i.e., there must be a class for every data item) and also that of mutual exclusively which means that a specific answer can be placed in one and only one cell in a given category set. Coding is necessary for efficient analysis and through it the several replies may be reduced to a small number of classes which contain the critical information required for analysis.

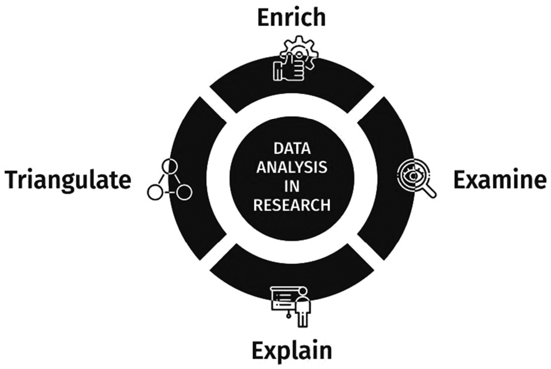
Tabulation: It is a systematic & logical presentation of numeric data in rows and columns to facilitate comparison and statistical analysis. It facilitates comparison by bringing related information close to each other and helps in further statistical analysis and interpretation. In other words, the method of placing organised data into a tabular form is called as tabulation. It may be complex, double or simple depending upon the nature of categorisation.

(Give examples of tables: Frequency tables, Cross tables)

(Graphical Presentation of data: Bar diagrams, Pie charts, Histograms, Frequency polygons, Line graphs, Scatter diagrams etc.)

Data Analysis

Data analysis in research is an illustrative method of applying the right statistical or logical technique so that the raw research data makes sense. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis. A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision.



Methods used for data analysis

After the data is prepared for analysis, researchers are open to using different research and data analysis methods to derive meaningful insights. For sure, statistical techniques are the most favoured to analyze numerical data. The method is again classified into two groups. First, ‘Descriptive Statistics’ used to describe data. Second, ‘Inferential statistics’ that helps in comparing the data.

Descriptive statistics

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables.

(Measures of central tendency & Measures of dispersion: Refer class notes)

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. In statistics, a central tendency is a central or typical value for a probability distribution. It may also be called a centre or location of the distribution. These values have the property that most of the observations in the data set accumulate around these values. The common measures of central tendency are the arithmetic mean, median and mode.

Arithmetic mean: It is the average of all the given observations

Arithmetic mean, $\overline{X} = \Sigma Xi /n$

Given scores: 6, 4, 5, 8, 2. Find the arithmetic mean.

$\Sigma Xi = 6 + 4 + 5 + 8 + 2 = 25$

Arithmetic mean = $\Sigma Xi /n = 25/5 =$ **5**

Median: It is the middlemost observation. ie, the observation in the (n+1)/2th position.

Xi : 6, 4, 5, 8, 2

Arrange: 2, 4, 5, 6, 8

Median = Middle observation = 5

Mode: It is the most frequent observation. ie, the observation which is repeated maximum number of times.

Xi : 6, 4, 5, 6, 2

Mode = Most repeated observation = 6

Measures of Dispersion / Measures of Variation

(also called variability, scatter, or spread)

It is the numeric value which gives the amount of variation in a distribution. That is, how the observations are spread among themselves. A measure of dispersion indicates the scattering of data. Measures of dispersion describe the spread of data around a central value. A high value indicates high variation and a low value indicates low variation. Zero indicates no variation, ie, all observations are same.

There are four measures of variation

- ***(i) Range***
- ***(ii) Mean Deviation/Average Deviation***
- ***(iii) Standard Deviation (SD)***
- ***(iv) Quartile Deviation (QD)***

(i) Range: It is the simplest method of measurement of dispersion and defines the difference between the largest and the smallest item in a given distribution.

Given the scores: 6, 4, 5, 8, 2. Find the range.

Smallest observation (minimum) = 2

Largest observation (maximum) = 8

Range = 8 – 2 = 6

(ii) Mean Deviation: It is the average of absolute values (positive values) of deviations from the arithmetic mean.

(Deviation: It is the difference of each observation from arithmetic mean).

For finding the mean deviation, first we have to find the mean $\overline{x}$.

Then find the deviation of each observation from the arithmetic mean. ie, difference of each observation from arithmetic mean.

Take the absolute values (positive) of the deviations $|x_i - \overline{x}|$.

Find the average of the deviations,

Mean Deviation= $\Sigma|x_i - \overline{x}|/n$.

(iii) Standard Deviation: It is the square root of average of squares of deviations from the arithmetic mean.

(Deviation: It is the difference of each observation from arithmetic mean).

Standard Deviation (SD) = $\sqrt{\dfrac{\Sigma(x_i - \overline{x})^2}{n}}$

For finding the standard deviation, first we have to find the mean $\overline{x}$.

Then find the deviation of each observation from the arithmetic mean. ie, difference of each observation from arithmetic mean.

Take the squares of the deviations $(x_i - \overline{x})^2$.

Find the average of the squares of deviations: $\Sigma(x_i - \overline{x})^2/n$.

Finding the square root of the above result will give the Standard Deviation.

Example: Given scores: 6, 4, 5, 8, 2. Find the Standard deviation.

Σ Xi = 6 + 4 + 5 + 8 + 2 = 25

Arithmetic mean = Σ Xi /n = 25/5 = **5**

$\Sigma(x_i - \overline{x})^2/n = 1^2+(-1)^2+0^2+3^2+(-3)^2/5 = 4$

$$\textbf{Standard deviation} = \sqrt{\frac{\Sigma(x_i - \overline{x})^2}{n}}$$

$$= \sqrt{4} = \mathbf{2}$$

(iv) Quartile Deviation: It is the deviation of the quartiles. Quartiles divide a distribution into four. There are three quartiles. The observation in 1/4th position is the first quartile Q_1, observation in 1/2th position is the second quartile Q_2 and observation in 3/4th position is the third quartile Q_3.

Quartile deviation = $(Q_3 - Q_1)/2$

Inferential statistics

With inferential statistics you take that sample data from a small number of people and try to determine if the data can predict whether the drug will work for everyone (i.e. the population).

There are two main areas of inferential statistics:

Estimating parameters: This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).

Hypothesis tests: This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Tests of Significance / Hypothesis Testing

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favour or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance.

Every test of significance begins with a ***Null Hypothesis H_0***. ***H_0*** represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the ***null hypothesis*** might be that the new drug is no better, on average, than the current drug. We would write $\boldsymbol{H_0}$: there is no difference between the two drugs on average.

The ***Research Hypothesis/alternative hypothesis***, ***Ha***, is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write Ha: the two drugs have different effects, on average. The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write Ha: the new drug is better than the current drug, on average.

The final conclusion once the test has been carried out is always given in terms of the ***null hypothesis***. We either ***"reject*** $\boldsymbol{H_0}$ ***in favour of Ha"*** or ***"do not reject*** $\boldsymbol{H_0}$***"***.

p-value :

When you perform a hypothesis test in statistics, a ***p-value*** helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the null hypothesis.

The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a ***p-value*** to weigh the strength of the evidence (what the data are telling you about the population). The ***p-value*** is a number between 0 and 1 and interpreted in the following way:

The significance level for a given hypothesis test is a value for which a ***P-value*** less than or equal to is considered statistically significant. Typical values for are *0.05, and 0.01*. These values correspond to the probability of observing such an extreme value by chance.

A small ***p-value*** (typically ≤ 0.05) indicates strong evidence against the ***null hypothesis***, so you reject the ***null hypothesis***. A large ***p-value*** (> 0.05) indicates weak evidence against the ***null hypothesis***, so you fail to reject the ***null hypothesis***.

Population:

A research population is generally a large collection of individuals or objects that is the main focus of interest of the researcher. It is for the benefit of the population that researches are done. Due to the large sizes of populations, researchers often cannot test every individual in the population because it is too expensive and time-consuming. This is the reason why researchers rely on samples.

Sample:

A sample is simply a subset of the population. The sample must be representative of the population from which it was drawn and it must have good size for further statistical analysis. The main function of the sample is to allow the researchers to conduct the study to individuals from the population so that the results of their study can be used to derive conclusions that will apply to the entire population. The population “gives” the sample, and then it “takes” conclusions from the results obtained from the sample.

Factors influencing Sample size calculation:

- Type I error
- Type II error
- Effect size
- Standard deviation of population

(i) Type I error, α - error (Level of significance) : cut-off level at which we say a p-value is significant. Probability of concluding that there is a statistically significant difference. Typically 5%.

(ii) Type II error (β - error) and Power (1- β): Power is the ability of a statistical test to show if a significant difference truly exist symbolized as 1- β. In hypothesis testing, it is important to have a sizable sample to allow statistical tests to show significant differences where they exist. Typically 80%, 90%.

(iii) Effect Size: It is the difference the researcher expects to see. What has been seen previously from reviews. What is a clinically important difference?

(iv) Standard deviation of population: It is the standard deviation of the outcome variable, in most of the cases obtained from previous studies.

1. Estimating the sample size for a descriptive study based on a proportion

To calculate the sample size based on the sample required to estimate a proportion, the following formula is used:

$$n \geq \frac{(z)^2 pq}{m^2}$$

n is the required sample size, z is the normal distribution value corresponds to 95% limits (1.96) or 99% limits (2.58), p = proportion of population having that characteristic, which can be known from previous studies or other sources, q = 1 – p (or 100 – p if p and q are expressed in percentages), m is the allowable error.

Example : A study on anaemic children in schools. Proportion of anaemic children in a similar study is found to be 30%. Find the minimum sample size required at a confidence limit of 95% and accepting an error of 10% of the population.

$$n \geq \frac{(1.96)^2 \text{ x } 30 \text{ x } 70}{10^2} = 80.67 \approx 81 \text{ or more samples}$$

2. Estimating the sample size for difference in means (comparison of two means)

The sample size is obtained by

$$n \geq \frac{\mathbf{2(SD)^2 (Z_\beta + Z_{\alpha/2})^2}}{\mathbf{Difference^{\,2}}}$$

n is the required sample size, $Z_{\alpha/2}$ is the desired level of significance - type I error, normal distribution value corresponds to 95% limits (1.96) or 99% limits (2.58), Z_β represents the desired power, typically 80% power (0.84), SD the standard deviation of the outcome variable, obtained from previous studies, difference is the effect size, ie, the difference in means.

Example : An interventional study on anaemic children in schools to improve their Hb level. From a previous study the mean difference is found to be 1.25 and the standard deviation is 2, calculate the required sample size to determine the significant difference with 5% level of significance and 80% power.

$$n \geq \frac{\mathbf{2(SD)^2 (Z_\beta + Z_{\alpha/2})^2}}{\mathbf{Difference^{\,2}}}$$

$$= \frac{\mathbf{2(2)^2 (0.84 + 1.96)^2}}{\mathbf{Difference^{\,2}}}$$

$= 40.14$

$\approx$ 40 or more samples

CORRELATION AND REGRESSION ANALYSIS

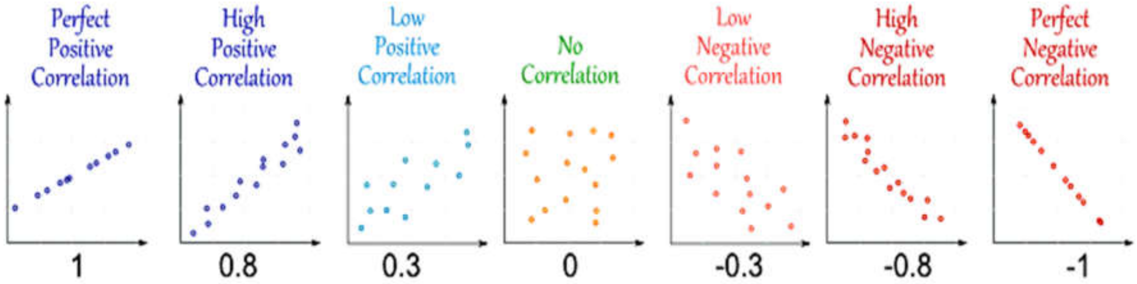
Correlation

It is a statistical measure which shows the relationship between two or more variables moving in the same direction or in opposite direction. With correlation, two or more variables may be compared to determine if there is a relationship and to measure the strength of that relationship. The correlation coefficient gives the strength of relationship between the variables.

- Correlation gives degree and direction of relationship
- Correlation does not require an independent (predictor) variable
- Correlation results do not explain why the relation occurs

The correlation may be either positive, negative or zero. The first role of correlation is to determine the strength of relationship between the two variables represented on the x-axis and y- axis. The measure of this magnitude is called the ***correlation co-efficient***. The data required to compute this coefficient are two continuous measurements (x, y) obtained on the same entity.

If there is a perfect relationship, a straight line can be drawn through all the data points. The greater the change in y for a constant change in x, the steeper the slope of the line. In a less than perfect relationship between two variables, the closer the data points are located on a straight line, the stronger the relationship and greater the correlation coefficient. In contrast, a zero correlation would indicate absolutely no linear relationship between the two variables.



Positive Correlation

One variable increases with increase of the other or decreases with decrease of the other. Eg: Body temperature and pulse.

Negative Correlation

One variable increases with decrease of the other or decreases with increase of the other. Eg: Insulin and blood sugar.

Zero Correlation

There is no relation between the variables.

The Coefficient of Correlation

A measure of the strength of linear relationship between two variables that is defined in terms of the covariance of the variables divided by their standard deviations.

Correlation coefficient, r = $\dfrac{\textbf{Covariance (x, y)}}{\textbf{(SDx) x (SDy)}}$

The following formulas gives the result of correlation coefficient.

Karl Pearson Correlation coefficient, r = $\dfrac{\Sigma\,(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma\,(x_i - \overline{x})^2}\;\sqrt{\Sigma\,(y_i - \overline{y})^2}}$

Spearman Rank correlation = $1 - \dfrac{6\,\Sigma\,d_i^2}{n\,(n^2 - 1)}$

Regression Analysis

In regression analysis, researchers control the values of at least one of the variables and assign objects at random to different levels of these variables. Where correlation simply described the strength and direction of the relationship, regression analysis provides a method for describing the nature of the relationship between two or more continuous variables. Correlation coefficient can support the interpretation associated with regression. If a linear relationship is established, the magnitude of the effect of the independent variable can be used to predict the corresponding magnitude of the effect on the dependent variable.

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (response) and independent variable(s) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables.

Regression analysis is a statistical method to estimate or predict the values of one variable (dependent variable) for the given values of independent variable.

> Dependent variable is to be estimated or predicted (response)

> Independent variable is the given variable (predictor)

Example: weight of a baby depends on age.

So age is the independent variable whereas weight is dependent variable.

Uses of Regression Analysis

- Describe one variable with level of other
- Understanding association eg: birth wt. & gestation
- Identify the variable which influence a particular one
- Prediction of dependent variable for given values of independent variable
- To identify the abnormal values or outliers

Types

- Simple Linear Regression (1 response – 1 predictor)
- Multiple Regression (1 response – Many predictors)
- Logistic Regression (Any response or predictors – Nominal / Ordinal)

1. Simple Linear Regression (1 response – 1 predictor)

The dependent variable is continuous, independent variable can be continuous or discrete, and nature of regression line is linear. Linear is used to denote that the relationship between two variables can be described by a straight line. With linear regression, a relationship is established between the two variables and a response for the dependent variable can be made based on a given value for the independent variable. For example

Injury Severity Score can be used to predict length of hospital stay.

The Regression equations are calculated using the following formula.

Regression equation of x on y : $(x - \overline{x}) = \dfrac{\Sigma\,(x_i - \overline{x})\,(y_i - \overline{y})}{\Sigma(y_i - \overline{y})^2}\,(y - \overline{y})$

Regression equation of y on x : $(y - \overline{y}) = \dfrac{\Sigma\,(x_i - \overline{x})\,(y_i - \overline{y})}{\Sigma(x_i - \overline{x})^2}\,(x - \overline{x})$

Predict x (response) given y (predictor), it is the regression line of x on y

Predict y (response) given x (predictor), then use the regression line of y on x

2. Multiple Regression (1 response – Many predictors)

The dependent variable (response) is predicted by using several independent variables (predictors) You could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety and lecture attendance.

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

3. Logistic Regression (Any response or predictors – Nominal / Ordinal)

This is the regression model in which the dependent variable is not continuous, ie, it is categorical. Independent variables can be continuous or discrete, and nature of regression line is linear. For example Smoking habit (Yes/No) can be used to predict COPD (Yes/No).

Binomial Logistic Regression

When the dependent (variable to predict) is binary (only two levels), eg : Yes/No

Multinomial Logistic Regression

When the dependent (variable to predict) is have more than two
levels eg : Opinion : Agree/Disagree/Neutral

ANOVA

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures such as the "variation" among and between groups used to analyze the differences among more than two group means in a sample. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.

One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test. One-way has one independent variable (with 2 levels) and two-way has two independent variables (can have multiple levels). For example, a one-way Analysis of Variance could have one independent variable (brand of cereal) and a two-way Analysis of Variance has two independent variables (brand of cereal, calories).