Review Article

SPER Publications and Solutions Pvt. Ltd.

Statistical analysis: A basic guide for pharmaceutical and bioscience researchers

Atul K. Goyal, Jyoti Saini

[1]Managing Director, Doc Navigator, New Chandigarh, Mohali, Punjab, India,
[2]Associate Editor, Journal of Pharmaceutical and Biosciences, India

Abstract

Background: Statistics is the science which summarizes the data of given population according to certain parameters. Pharmaceutical and bioscience research involves the generation of qualitative and qualitative data which is needed to be analyzed statistically. It is therefore essential for the researchers to have a basic understanding of statistical analysis. **Objective:** The present article is designed to enhance the basic understanding of statistical analysis among pharmaceutical and bioscience researchers without undergoing the technical and mathematical details. **Main Text:** All the sections are designed to improve the preliminary understanding of researcher by keeping the language and terminology at basic levels. The article highlights the type of data and scale, distribution patterns, measure of central tendency and variability, hypothesis testing, statistical comparison between two or more group, correlation and regression analysis. We also highlighted some new topics including the sensitivity, specificity, and prediction models. **Conclusion:** The article will help in developing the basic understanding of statistics among the pharmaceutical and bioscience researcher. The article will also help the researcher to identify suitable statistical methods to analyze the data of their research.

Keywords: Analysis of variance, correlation, data, hypothesis, *P*-value, statistics

Correspondence:
Atul K. Goyal,
Managing Director, Doc Navigator,
New Chandigarh, Mohali, Punjab, India.
Email: atlgyl@gmail.com

Introduction

Pharmaceutics and bioscience are the two major domains in which a large proportion of overall research is carried out. Every piece of research nowadays requires statistical validation. Researchers usually lack the basic understanding of the statistical analysis because they think that it is a mathematical thing and very difficult to understand. However, without the basic understanding of the statistics, researcher usually trapped with a weak study design, and large efforts usually need to made at the end of the study to organize the data for statistical analysis. Even in many cases, researcher conducted the whole research without any statistical framework and thought that statistics is the end part process. In such cases, the output of the research comes out to be very weak due to lack of statistics knowledge at initial stages of research.[1]

Being from biological background, the researchers in India left the mathematics subject after secondary school and statistics is usually teach in college and universities as an integral part of a full syllabus. Researcher finds it difficult to understand the basic of statistics because of lack of mathematical background and usually skip the statistics part of the curriculum. As a result, when students conducted the research at the later phase of academics, they faced difficulty in building the statistical knowledge. It is also not possible for them to learn the whole book on statistics to gain basic statistics understanding. The data available online is also scattered, confusing, and difficult to understand by the researcher because of technical terminology and mathematical language.[1]

Access this article online

Website: http://www.jpbs-online.com	**E-ISSN:** 2321-0125

How to cite this article: Goyal AK, Saini J. Statistical analysis: A basic guide for pharmaceutical and bioscience researchers. J Pharm BioSci 2023;11(4):1-6.
Source of Support: Nil, **Conflicts of Interest:** None declared

The present article is designed to enhance the basic understanding of a pharmaceutical and bioscience researcher about the basic of statistics so that researchers can plan their research at the beginning. This article is designed by omitting out the mathematics and technical terminology to make easy to understand by the researchers. The present article will help in guiding the researcher in developing a basic understanding of statistics both for academic and research purposes.

Qualitative and Quantitative Data

The data about the population could be the qualitative or quantitative [Figure 1]. It is important to know the difference between the two types of data as which statistical test we are going to apply on data is largely depends on the nature of data. The qualitative data are represented by the categorical variables which cannot be described in numeric form. For example, gender (male or female), color (red, green, or blue), and pain (mild, moderate, or severe) are example of qualitative data. Non-parametric test is exclusively analyzing the qualitative data. The quantitative data are the data which are presented in the numeric form. For example, height, age, weight, volume, and distance are the example of quantitative data which have a fixed numeric value. Parametric tests are used for the analysis of quantitative data.[2]

Quantitative data when presented for a single individual are called as discrete data. For example, age of patient in 34 years, here the 34 is a discrete value. When we talk about a population, the quantitative data take the form of continuous data. For example, age of cancer patients in the present study ranges from 20 to 55 years, here the 20–25 years is a continuous data and age of any patient from the whole patient population can belong to any number between 20 and 25.[2]

Nominal, Ordinal, Interval, and Ratio Scale

There are four scales in the statistics on which the data are dealt with [Figure 2]. The nominal scale deals with the qualitative data or categorical data; for example, gender, color, and pain describe previously are the categorical variable to be dealt with nominal scale. Mode is only measure of central tendency which is available on the nominal scale.[3]

The ordinal scale deals with the discrete data and represents a meaningful order in data; for example, rank of the students in a class. Here, there is no fixed interval between the ranks, for example, rank 1 student may score 98 marks, rank 2 student my score 90 marks, and rank 3 student may score 88 marks. Hence, there is no fixed interval between the scores obtained by various rank holders. Mode and median are the measure of central tendency which is available on the ordinal scale.[3]

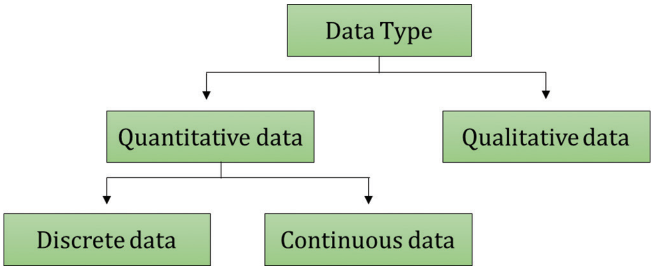

Figure 1: Quantitative and qualitative type of data

In case when the interval is fixed between the two levels, the scale is called an interval scale. Although the interval is fixed, there is no true zero in the interval scale so the value can go in the negative. For example, in temperature value in thermometer has fixed interval and temperature may go in negative. Mode, median, and arithmetic mean are the measure of central tendency which is available on the ordinal scale. Only addition and subtraction calculations are possible in the variables that existed on interval scale; multiplication and division calculations are not possible due to a lack of true zero.[3]

The last scale is the ratio scale on which most of the quantitative data is dealt with. The ratio scale has a fixed interval and also has a true zero point so the values cannot go in negatives. For example, weight, volume, distance, and height all have fixed interval, true zero point and cannot go in negative. Mode, median arithmetic, and geometric mean are the measure of central tendency which is available on the ordinal scale. Addition, subtraction, multiplication, and division, all calculations are possible on ratio scale due to the existence of true zero.[3]

Binomial, Poisson, and Normal Distribution

Similar to scale, there also occurs three data distribution patterns in statistical analysis. The binomial distribution is followed when the data can have only one possible outcome out of two possible outcomes, for example, the gender of an individuals could either be male or female, never both. In binomial distribution, an experiment is repeated under identical conditions. The trials are repeated for a fixed number of times and every trial gives only two possible outcomes. The probability of outcome remains unchanged from trial to trial and all the trials are independent, and the outcome of the previous trial did not affect the further trials.[4]

Poisson distribution deals with discrete data and is said to be followed when the chances of happening of an adverse event are calculated over a fixed period of time, for example, number of road accidents in a

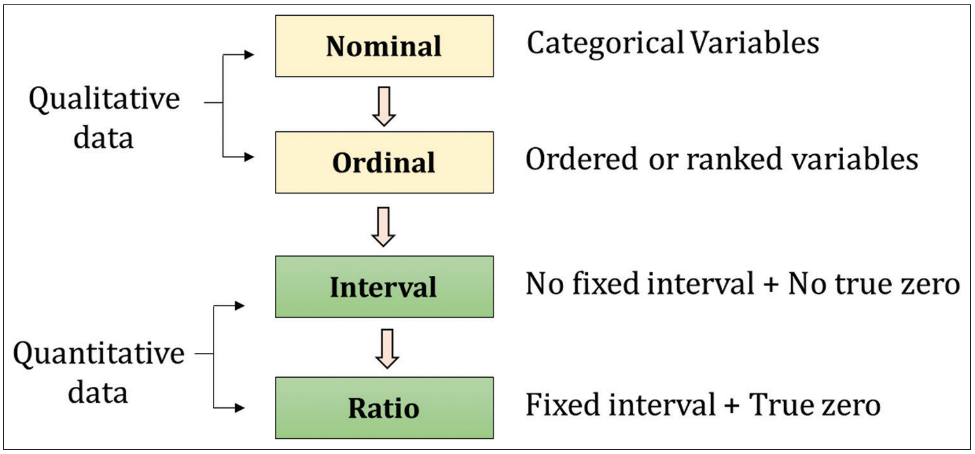
Qualitative data
Quantitative data
Nominal
Ordinal
Interval
Ratio
Categorical Variables
Ordered or ranked variables
No fixed interval + No true zero
Fixed interval + True zero

Figure 2: Nominal, ordinal, interval, and ratio scale

month or number of deaths in a hospital in a year. Poisson distribution does not have a fixed number of trials, but the number of successes is recorded for a fixed time interval.[4]

As name suggests, continuous distribution deals with the continuous data, for example, age or weight distribution in a population. The normal distribution is an example of an unimodal distribution. An unimodal distribution is a distribution with one clear peak. The normal distribution has two parameters, the mean and standard deviation. The normal distribution has a bell-shaped curve with normal curve symmetrical maximum at mean. The mean, median, and mode coincide in the normal distribution. Skewness in the normal curve is zero and curve is mesokurtic [Figure 3].[4]

Measure of Central Tendency and Variability

While analyzing the data about a population, researchers are interested in measuring of central tendency and variability in the data. The central tendency is defined by the value around which most of the values are clustered whereas variability describes how much values tend to fall far away from the central value. The central tendency for the quantitative data is measured by the mean, median, and mode while the variability is measured by the range and standard deviation. Mean is most common method to measure the central tendency, and median and mode are used when there is high number of outliers exists in the data. Outliers are the value which fall very far away from the central values. In case of quantitative data, mode is the only method available to measure the central tendency.[5]

Skewness

When in a continuous quantitative data, the distribution is not forming a uniform bell-shaped curve but gets distorted due to high number of outliers, the data are said to be skewed. The skewedness in the data measures the degree of distortion in the symmetrical normal distribution bell curve. The data are said to be positively skewed when the right side tail is much longer and mean is greater than median and mode. The data are said to be negatively skewed when the left side tail is much longer and mode is greater than median and mean. Skewness in the data is said to be absent when a normal bell shape curve is formed and mean is equal to the median and mode [Figure 4].[6]

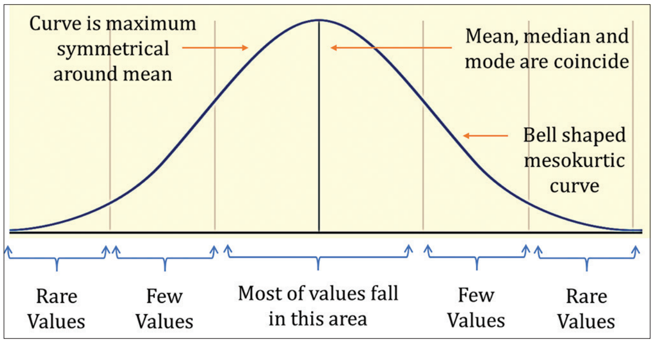

Figure 3: A normal distribution curve

Kurtosis

Kurtosis is the measure of outliers present in a normal continuous distribution. A high degree of kurtosis is called as negative kurtosis. A platykurtic curve is made in such case which indicates large number of outliers and the tail of distribution is said to as heavy tail. A low degree of kurtosis is called as positive kurtosis. A leptokurtic curve is made in such case that indicates lack of outliers and the tail of distribution is said to as light tail. In case of perfect normal distribution, the kurtosis is zero and distribution curve is called as mesokurtic [Figure 5].[6]

Null Hypothesis and Alternate Hypothesis

While conducting basic research, the first step is to form a hypothesis. The hypothesis could be of null hypothesis and alternate hypothesis. The aim of the researcher is always to reject the null hypothesis and accept the alternate hypothesis. For example, if a researcher wants to prove that "the drug is effective against disease," researcher will make a null hypothesis "drug is not effective against the disease." In this case, by rejecting the null hypothesis, the researcher is actually proving that the drug is effective against the disease. Although it looks tricky to made efforts to reject own hypothesis, this is how things work in statistics. The alternate hypothesis is the same thing which a researcher wants to prove. For example, if a researcher wants to prove that "the drug is effective against disease," researcher will make an alternate hypothesis "drug is effective against the disease" and simply tried to accept it. The question is why a null hypothesis is required when we can simply follow an alternate hypothesis. The thing is that the null hypothesis gives more statistical power and confidence level as compared to the alternate hypothesis, therefore, researcher usually made a null hypothesis and made efforts to reject it.[7]

P-value

A null hypothesis is rejected based on the level of significance which indirectly shows the level of confidence to reject a null hypothesis. The famous *P*-value is the measure of the level of significance. A null

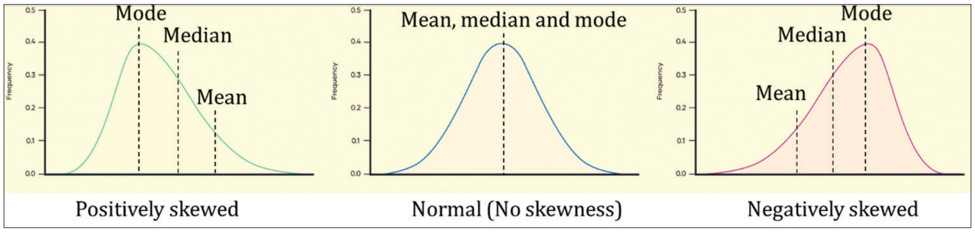

Figure 4: Skewness in the normal distributed curve

hypothesis is usually rejected when $P < 0.05$ which implies that out of 100% outcome, only 5% of outcomes may be due to the chance and rest of 95% of the outcomes are actually due to the intervention. It means that researcher is 95% confident that outcomes are due to the intervention, not by chance.[8]

The rejection regions can exist on one side or both side of distribution tail [Figure 6]. In one tailed critical region, the rejection region occurs only one side, for example, students who get marks more than 33 will be pass, here the rejection region is only below 33. In case of two-tailed critical regions, the rejection region occurs on both sides, for example, human baby having very high or very low weight faced greater mortality.[8]

Type I and Type II Errors

The aim of the researcher is to reject a false null hypothesis and accept a true null hypothesis. Errors come in hypothesis when a false null hypothesis is accepted which actually should have been rejected, or when a true null hypothesis is rejected which actually should have been accepted. The type I error (α) called false positive occurred due to the rejection of true null hypothesis, and type II errors (β) called true negative occurs due to the acceptance of false null hypothesis [Table 1].[7]

Type I error is considered more serious than type II error. Let's compare the Type 1 error and Type II error with a police case. If the punishment is death, a Type I error is extremely serious because in that case, an innocent person goes to jail and guilty person goes free whereas the Type II errors are less serious as in that case, only guilty person goes free but innocent person does not punished [Table 2].[8]

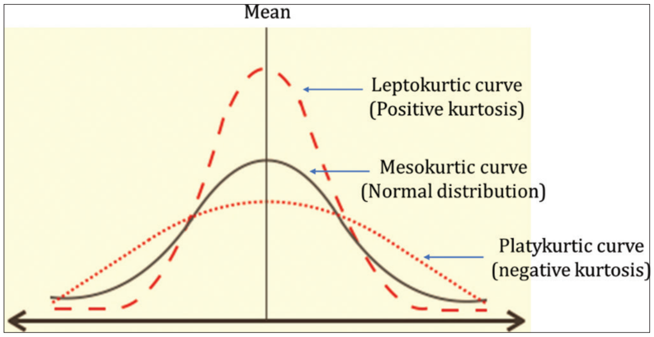

Figure 5: Kurtosis in the normal distribution

Compare one Group to a Specific Value

While conducting the research, a researcher may want to compare the data obtained from the population study with a single value. For example, a clinician wants to compare how many patients have their blood pressure, hemoglobin, or electrolytes within the normal range. In this case, the student t-test is being used in case when data are quantitative and Wilcoxon rank sum test is used if the data are qualitative [Table 3].[9]

Compare Two Groups

In the pre-post studies where the intervention is given to the patients at baseline and effect of intervention is measured in the same group after 3–6 months of follow-up period, paired tests are used. For example, a researcher administrated a drug for 3 months in a group of patients and wanted to study its effect on blood pressure after 3 months of treatment. For quantitative data, paired t-test is used to compare the means of two groups. For qualitative data, Wilcoxon rank sum test is used [Table 3].[9]

In case–control studies, or randomize control trials where two independent groups need to be compared, the unpaired test is used. For example, a researcher administrated a drug in case group and take another group of patients as control in which no drug or placebo is administrated, and after few weeks of treatment, he wanted to compare blood pressure between the two groups. Here, the unpaired t-test will be used to compare two means if data are quantitative, and Mann–Whitney U-test or Kendall's s-test could be used if data are qualitative [Table 3].[9]

Compare More Than two Groups

In some studies, more than two groups could be needed to compare. For example, a researcher wants to compare blood pressure in three groups, the first group doing yoga, the second group is doing other physical exercises and the third group does not perform any physical activity. In such cases, when three groups are needed to be compared and data are quantitative, the one-way analysis of variance (ANOVA) is used. In case of qualitative data, the Kruskal–Wallis test is used [Table 3].[9]

In case the researcher wants to compare two variables between the three groups, for example, a researcher wants to compare blood

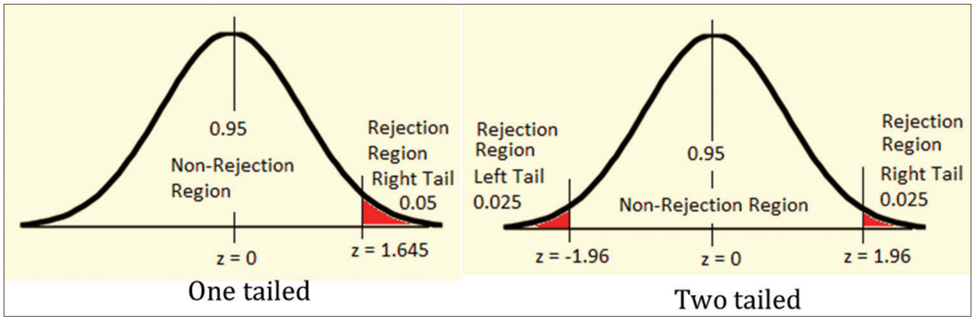

Figure 6: One-tailed and two-tailed rejection regions

Table 1: Type I and Type II errors

	Truth	
	Null hypothesis true	**Null hypothesis false**
Decision		
Reject null hypothesis	Type I error	Correct decision
Accept null hypothesis	Correct decision	Type II error

Table 2: Seriousness of Type I and Type II errors

	Truth	
	No guilty	**Guilty**
Verdict		
Guilty	Type I error: Innocent person goes to jail and guilty person goes free	Correct decision
No guilty	Correct decision	Type II error: Guilty person goes free

Table 3: Parametric and non-parametric tests with their function

Parametric test	Functions	Non-parametric test
Student t-test	Compare one group to a specific value	Wilcoxon rank Sum test
Paired t-test	Compare two paired group	
Unpaired t-test	Compare two independent groups	Mann-Whitney U-test Kendall's s-test
One way ANOVA	Compare of three or more group with one variable	Kruskal Wallis test
Two way ANOVA	Compare three or more groups with two variables	Friedman test
Pearson's correlation	Measure association between two variables	Spearman rank correlation Kendall's rank correlation
Z-test	Test of independence, or test of goodness of fit or test of homogeneity	Chi-square

ANOVA: Analysis of variance

pressure and heart rate in the above-mentioned three groups. In such cases, two-way ANOVA is used if the data are quantitative. For qualitative data, Friedman test, also called as F-test, is used [Table 3].[9]

Correlation

Correlation is the degree of relatedness between the two variables. Correlation could be the positive correlation and negative correlation. In positive correlation, if the value of one variable is increased, a corresponding increase in the value of the second variable is observed. For example, a positive correlation occurs between height and weight of an individual. A negative correlation between the two variables is exists when there is an increase in the value of one variable, and a corresponding decrease in the value of second variable is observed. For example, negative correlation occurs between the height and temperature in the atmosphere [Figure 7].[10]

Correlation could also be linear or non-linear. In case of linear correlation, a change in one variable leads to change in another variable at a constant ratio. In case of non-linear correlation, a change in one variable does not leads to change in another variable at a constant ratio. The correlation analysis is represented by the scatter diagrams. Scatter diagram is a type of graph in which values of two variable (X and Y) are plotted in forms of dots against two axis and then a straight line is drawn with the aims to cross maximum dots. This line will represent the degree and magnitude of correlation.[10]

Correlation is measured by the correlation coefficient. The value of correlation coefficient range between 0 and 1. Higher the value of correlation coefficient, higher will be the correlation between the two variables. For example, if the value of correlation coefficient for two variables i.e., height and weight is +0.84, then it means there is 84% correlation between height and weight, or there are 84% chances that increase in height leads to corresponding increase in weight or vice versa. The correlation coefficient comes with the +ve or -ve sign. The +ve sign indicate the positive correlation between the two variables and -ve sign indicate the negative correlation between the two variables.[10]

In case when data is quantitative, Karl Pearson's coefficient of correlation (r) is used to calculate the correlation between the two variables. In case of qualitative data, Spearman's Rank coefficient of correlation (p) is used to analyze correlation between two variables.[10]

Regression

Regression analysis used to predict the value of an unknown variable from the value of known variable. Regression lines used to estimate value of unknown variable from the given value of known variable. For example, if plot the data of height and weight about population, we will get a regression line; now if we know the value of weight, we can predict the height using the regression line sloop or vice versa [Figure 8].[10]

Sensitivity and Specificity

Sensitivity describes the accuracy or ability of a technique to detect a true positive state in an assay whereas specificity is described the precision or ability of a technique to detect a false positive state. Fisher's exact test and chi-square test are used to analyze the sensitivity and specificity of an assay. A 2 × 2 matrix is must to be constructed to calculate the sensitivity and specificity of an assay [Table 4].[11]

For example, let's compare the two assays one is gold standard and another is a newly developed assay to detect the HIV antigen in patients. A total of 100 samples were screened for HIV antigen using both assay and results are entered in table as: (a) one column contain samples positive in both assays, (b) one column contain sample negative in both assays, (c) sample which are positive in new assay but negative in standard assay, (d) samples which are positive in standard assay but negative in new assay. Now if we analyze the data, the sensitivity of new assay comes to be 0.87 and specificity of assay 0.83 which means that newly developed assay can detect the HIV antigen with 87% sensitivity and 83% specificity as compared to the gold standard.[11]

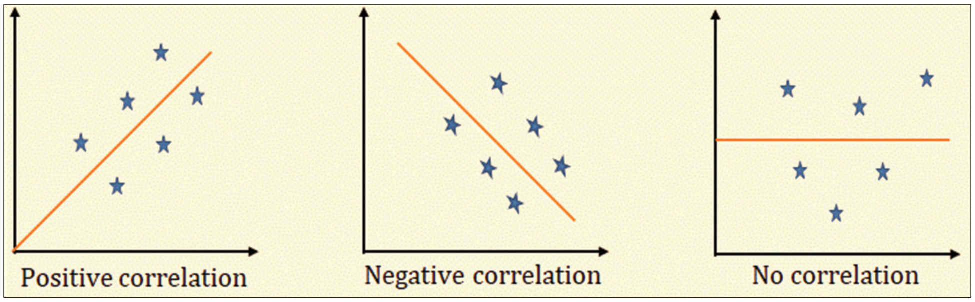

Figure 7: Positive, negative, and absence of a correlation

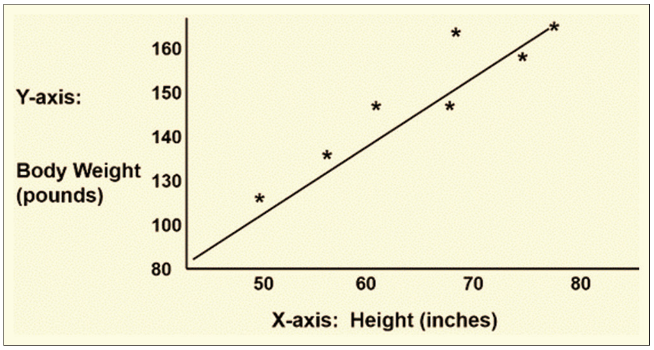

Figure 8: Regression analysis

Table 4: Analysis of sensitivity and specificity of newly developed assay

	Gold standard assay	
	HIV +ve	HIV -ve
New assay		
HIV +ve	(a) 35	(b) 10
HIV -ve	(c) 5	(d) 50

Survivorship Curve

Survivorship curves give the probability of happening a particular event like death, over a period of survival. The Kaplan–Meier curve is used to create the survivorship curves. In this method, the data contained two columns, in the first column, a quantitative number of month or years is entered and in the second column, survival or death of the individual is entered. For example, the survival curve can be used to calculate the probability of mortality in cancer patients according to different stages of cancer.[12]

Predicting Outcomes

Sometimes a researcher wants to analyze the accuracy of a quantitative variable in predicting a particular outcome. For example, a researcher wants to analyze the accuracy of molecular marker in predicting the survival of a cancer patient. In such case, the molecular expression data are quantitative and the outcome is qualitative, i.e., survival or death. Receiver operating characteristic curve analysis is carried out in such cases which give the outcome as AUC (area under the curve). For example, if the value of AUC comes to be 0.84 for survival and 0.56 for death that it means that the molecular marker can predict the survival rate in cancer patient with 84% accuracy and death rate with 56% accuracy.[13]

Conclusion

In the present article, we describe the type of data and statistical scales on which data are analyzed. We also explained the distribution patterns of data. The methods to measure the central tendency and variability are also mentioned. We also explain the outliers in data and their measurement using skewness and kurtosis. Testing of a hypothesis is also described along with the explanation of *P*-value and types of errors. Suitable statistical methods to compare one group with a value and comparison of two or more group is also put forward. Correlation and regression analysis were also explained. At last new methods of calculation sensitivity, specificity and predicting outcome are explained. The present article is designed by keeping the language and terminology as simple as possible so that basic understanding about the statistical analysis of data could be developed in the pharmaceutical and bioscience researchers.

References

1. Goyal A, Saini J. Statistical medicine as a distinct medical specialty. J Pharm Biosci 2023;11:1.
2. Ali Z, Bhaskar SB. Basic statistical tools in research and data analysis. Indian J Anaesth 2016;60:662-9.
3. Marateb HR, Mansourian M, Adibi P, Farina D. Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies. J Res Med Sci 2014;19:47-56.
4. Bonamente M. Three fundamental distributions: Binomial, gaussian, and poisson. In: Statistics and Analysis of Scientific Data. New York: Springer; 2017. p. 35-54.
5. Christopher A. Making sense of data: Measures of central tendency and variability. In: Interpreting and Using Statistics in Psychological Research. Vol. 1. United States: SAGE Publications, Inc; 2017. p. 93-128.
6. Tyagi H. RxPG Series: Biostatistics Buster. New Delhi: Jaypee Brothers Publication; 2003.
7. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, Type I and Type II errors. Ind Psychiatry J 2009;18:127-31.
8. Andrade C. The P value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. Indian J Psychol Med 2019;41:210-5.
9. Kaur A, Kumar R. Comparative analysis of parametric and non-parametric tests. J Comput Math Sci 2015;6:336-42.
10. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. Perspect Clin Res 2017;8:148-51.
11. Shreffler J, Huecker MR. Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. In: StatPearls. Treasure Island, FL: StatPearls; 2020.
12. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res 2010;1:274-8.
13. Konar S, Auluck N, Ganesan R, Goyal AK, Kaur T, Sahi M, *et al*. A non-linear time series based artificial intelligence model to predict outcome in cardiac surgery. Health Technol 2022;12:1169-81.