Understanding Medical Biostatistics

10

R. L. Sapra

Variation is a law of nature that makes this universe beautiful. In healthcare, two human beings, though genetically similar, may not respond equally to the same drug. The same drug may also have a varying response and become ineffective in an individual over a period of time. When we look at a population of individuals, variations are so prominent that no two individuals are ever exactly alike. There may be several factors for these variations among individuals which include those which are due to biological, genetic, environmental, or ecological effects [1]. Variations also occur when we sample these individuals. Moreover, variability in the observers themselves may also contribute to variations in assessment. These variations are bound to occur whatever may be the reasons and consequently lead to uncertainties in clinical practice, identification of risk factors, and policy planning.

Statistics is the science that manages these uncertainties and helps in minimizing their impact on decisions. Medical biostatistics manages uncertainties in fields related to medicine and health. It comprises a large number of statistical tools/methods such as descriptive and inferential statistics, associations and relationships, diagnosis, prediction and forecasting, assessment of risk and survival analysis, studies and experimental designs and sampling and survey methodology. We discuss here some of the basic concepts of medical biostatistics in the form of questions and answers to make this subject interesting and easy to implement in day-to-day clinical research.

The original version of the book was revised: The authors and co-authors have been cited within the chapters. The correction to this book is available at https://doi.org/10.1007/978-981-16-5248-6_49.

"If your experiment needs a statistician, you need a better experiment."Ernest Rutherford British Physicist (1871–1937).

S. Nundy et al., *How to Practice Academic Medicine and Publish from Developing Countries?*, https://doi.org/10.1007/978-981-16-5248-6_10

10.1 Understanding the Basics of Biostatistics

We begin by explaining what are data, variables, the types of variables and then go on to statistics.

10.1.1 What Are Data?

Data are a set of related raw information. For example, the data may be the demographic information about patients such as their age, gender, medical condition, and diagnosis. The raw information cannot be used as such, and require statistical methods for better understanding and interpretation. The word 'statistics' is generally used together with data and refers to the results of data analysis.

10.1.2 What Is a Variable?

A variable is a characteristic whose values vary. For example, the serum cholesterol, creatinine, height, weight, age, and the colour of eyes and skin of subjects vary from individual to individual and hence they are variables. Variables are measured on different scales of measurement depending on their nature.

10.1.3 What Are the Different Scales of Variable Measurement?

There are four scales that are used for measuring variables.

Nominal/Qualitative/Categorical Nominal variables, also called qualitative/categorical, consist of categories/classes, e.g., colour of eyes (blue, grey, black), status of health (healthy or unhealthy), and type of cancer (benign or malignant). A nomi-

nal variable with two categories is called a binary variable and polytomous if it has more than two categories. Nominal data can also assign numeric codes to categories, e.g., healthy =1, unhealthy =2. However, no mathematical manipulations (addition, subtraction, multiplication, and division) can be done on these numeric codes.

Ordinal Ordinal data are ordered or ranked data. For example, pain score (1—low, 2—moderate, 3—severe), intelligence level (1—poor, 2—average, 3—sharp), satisfaction level (1—highly dissatisfied, 2—dissatisfied, 3—neutral, 4—satisfied, 5—highly satisfied) are scored on an ordinal scale. In an ordinal scale we can say $1 > 2 > 3 > 4 > 5$. But we cannot add or subtract these numbers. In other words, we cannot say that the difference between 2 and 1 is equal to 4 and 3. Ordinal variables cannot be used for calculating mean and standard deviations. However, we can use the median or mode for ordinal data to measure the central tendency.

Interval Interval scales are numeric which show both order and direction. We can add or subtract these numbers but cannot multiply or divide them. Interval scales do not have a true zero point. We can measure mean, median and mode, etc., and calculate the dispersion. Examples include age in years (1, 2, 3, 4…), time of each day, or temperature in Celsius or Fahrenheit. In interval scales, we cannot calculate ratios. We cannot say 60 degrees C is twice as hot as 30 degrees. However, we can calculate the exact difference between the two values.

Ratio The Ratio scale is a quantitative scale and has a true zero, unlike the interval scale. Here the ratio of two measurements has a meaning interpretation. For example, a weight of 60 kg is twice as heavy as a weight of 30 kg. Examples of ratio scale variables include body weight, height, and creatinine levels. Ratio scales do not have a negative number. All mathematical manipulations can be done on ratio scale variables.

10.1.4 What Is the Likert Scale?

The Likert scale was developed by Rensis Likert for psychometric assessment such as opinions, beliefs, and attitudes on a variety of items. Likert scales include 4-, 5-, 7-, and 9-point scales to assess the response of the subjects through a questionnaire containing a number of items. The first scale, i.e., a 4-point scale is also called a 'forced-scale' as it does not give an option to the respondent to stay neutral in his opinion. The others are odd-numbered point scales where the respondent can exercise this option.

An example of a 4-point scale can be the feedback of trainees on items such as course material and quality of contents. Responses on a 4-point scale can be—excellent, good, average, and poor. The difficulty with the 4-point scale is that the respondent is forced to answer when he has no opinion. This drawback distorts the results. On many occasions, a respondent might not answer the item.

Examples of a 5-point scale for assessing the satisfaction level regarding the nursing services provided by a hospital can be as follows:

10.1.4.1 Example 1

1. Highly dissatisfied.
2. Dissatisfied.
3. Cannot say (Neutral).
4. Satisfied.
5. Highly satisfied.

10.1.4.2 Example 2

1. Very poor.
2. Poor.
3. Average.
4. Good.
5. Excellent.

The other example could be an opinion about the implementation of some medical policy.

10.1.4.3 Example 3

1. Strongly disagree.
2. Disagree.
3. Neutral.
4. Agree.
5. Strongly agree.

A 7-point or 9-point scale includes more options so that options from strongly disagree to strongly agree are equally spaced and convey meaning to the options. The following is the 7-point scale example:

10.1.4.4 Example 4

1. Strongly disagree.
2. Disagree.
3. Somewhat disagree.
4. Neutral.
5. Somewhat agree.
6. Agree.
7. Strongly agree.

A 9-point scale has 9 options and the middle point 5 is the neutral point where the respondent neither agrees nor disagrees. However, the 5- and 7-point scales are commonly used as the 9-point scale becomes somewhat complicated due to the higher number of classes and maybe confusing to a respondent.

10.1.5 What Are Descriptive and Inferential Statistics?

The study of statistics can be broadly classified into two categories - descriptive and inferential statistics. Descriptive statistics organizes and summarizes data so that it gives some meaning to it. Descriptive statistics comprises the most basic components of statistics such as *proportions*, measures of *central tendency*, measures of *variability* or spread of the data and *graphical representation* (bar diagrams, pie charts, scatter plots, etc.). Measures of central tendency provide the central or average value of the data, whereas measures of variability or spread focus on the dispersion of values in the data set.

Inferential statistics arrive at conclusions about a population based on sample data drawn from the population. Inferential statistics includes a large number of statistical tools such as probability distributions (Normal, Poisson, Student-t, F, Chi-square distributions), testing of hypotheses, correlation and regression analyses, analyses of variance, etc.

10.1.6 What Is a Measure of Central Tendency and What Are the Various Measures of Central Tendency, Their Merits and Demerits?

When we have data containing a large number of values, the first question that comes to mind is 'Can we find a single value which can best describe our data rather than handling lots of numbers?' The single value or the representative value is called the central tendency or central location of the data. There are a large number of measures of central value in statistics. However, mean, median, and mode are common measures. Below we discuss these measures in detail.

10.1.6.1 Mean

The mean, also known as the arithmetic mean (AM) or average, is the most commonly used statistic. It is the average of all values, is simple to compute, and has several good statistical properties. However, the mean may not be an appropriate choice, particularly when the data has extreme values/outliers or has a skewed distribution. Also, the mean cannot be used for data scored on an ordinal scale such as intelligence, honesty, and pain score.

10.1.6.2 Median

The median is the central value in the data arranged either in an ascending or descending order. It is the appropriate choice for skewed (asymmetric) data. It can also be used when data is scored on an ordinal scale such as intelligence score and satisfaction level. Unlike the mean which is severely affected by extreme values, the median is not affected by the extreme values and is preferred over the mean. In the medical field, hospital stay (number of days) is generally skewed data as most patients stay for a few days and only a few have a longer stay. Here the median seems to be a more appropriate choice compared to the mean.

10.1.6.3 Mode

The mode is not as commonly used as the mean or median and is the value in the data which occurs most frequently. Systolic and diastolic blood pressure, i.e., 120/80 is perhaps the most common example of a mode as most normal human beings have this level of blood pressure. It is simple to understand, easy to locate, and like the median unaffected by extreme values, but has several demerits. The mode is not based on all observations as it ignores all values except the one which has the maximum frequency. It is inappropriate for bimodal or multimodal data and is not considered to be a good measure of a central value.

10.1.6.4 Example

The following is the age (years) distribution of 15 subjects. Let us calculate the mean, median, and mode of the data.

Data	10, 1, 2, 15, 10, 14, 13, 9, 9, 10, 11, 10, 5, 6, 6
Arranged data	1, 2, 5, 6, 6, 9, 9, 10, 10, 10, 10, 11, 13, 14, 15
Mean (divide the sum of 15 observations by 15)	8.73
Median (eighth observation in the arranged data)	10
Mode (observation with maximum frequency)	10

Let us see how the extreme value affects the mean and not the median. In the above data set if the data value 15 is replaced by 95 by mistake or otherwise, the mean increases to 14.07 whereas the median remains the same.

10.1.6.5 Less Frequently Used Measures

In addition to the mean, median, and mode the statistical literature also includes several less frequently used measures such as the harmonic, geometric, truncated, interquartile, midrange, midhinge, and trim mean which are rarely used except in specific conditions.

10.1.7 What Is Dispersion or Variability in a Data Set and How Is It Measured?

Measures of central tendency such as the mean, median, etc. focus on a central value that can describe the data but they do not give any idea how the values in the data set are scattered or dispersed. For example, consider two data sets having four values in each - data set-I: 0,25,75,100 and data set-II: 48,51,52,49. Surprisingly both the data sets have the same value for the mean, i.e., 50. But if we look at the values in the data set-I, the values are highly dispersed and away from the mean, whereas they are much closer in data set-II and to the mean value as well. Thus, with measures of central tendency alone, we will not be able to understand the distribution of the data accurately and we have to have a measure that understands the

spread or variability among the data values. Below, we discuss the various measures of dispersion.

10.1.7.1 Range

The Range (R) is the simplest measure and provides a broader idea of the dispersion. It is the difference between the maximum and minimum values of a data set. For example, the range for the data set-I mentioned above is 100 and for the set-II is 4. Thus, there is a high level of dispersion in set-I compared to set-II. The major drawback of this dispersion measure is that it is based on only two values (minimum and maximum) and ignores all the other values which lie between these two.

10.1.7.2 Interquartile Range

The Interquartile range (IQR) also known as the midspread, middle 50% or H-spread describes the middle 50% of the values when placed in order from the lowest to the highest. It is the difference between the third and first quartile. The first quartile is the value in the data set that has 25% of the values below it. The median is the middle value and is also known as the second quartile. The third quartile holds 25% of the data values above it. The IQR is preferred over the range (R) and is used for constructing boxplots to visualize the spread and identify outliers.

10.1.7.3 Standard Deviation

The Standard deviation ($\boldsymbol{\sigma}$), also called the root mean square deviation, is the most popular and commonly used measure for describing the variation in a data set. Unlike the range and interquartile range, the standard deviation is based on all the values and measures the variation from the mean value. The formula for calculating the standard deviation is:

$$\sigma = \sqrt{\frac{\Sigma\left(x_i - \mu\right)^2}{N}}$$

where $\boldsymbol{x}_i$ is each value of the population, $\boldsymbol{\mu}$ is the population mean, and $\mathbf{N}$ is the population size.

The above formula is used for calculating the population standard deviation when the data is considered a population. However, while calculating the sample standard deviation, the divisor is $n-1$, where n is the sample size.

The standard deviation measures squared deviations of all the data values from the mean and is the most reliable measure of dispersion. It is rigidly defined and difficult to compute. However, it is the least affected by the sampling fluctuations. The square of the standard deviation is called the variance.

10.1.7.4 Median Absolute Deviation

When our data is skewed, the median is the appropriate measure of average and the median absolute deviation (MAD) should be preferred over the standard deviation for measuring the dispersion. MAD is the median of the absolute deviations from the median.

$$\text{MAD} = \text{median}\left(|\, x - \text{median}\left(x\right)|\right).$$

Table 10.1 Calculation of MAD

x	x-median	|x-median|
40	−4	4
40	−4	4
44	0	0
46	2	2
46	2	2
45	1	1
39	−5	5
80	36	36
11	−33	33

Note: When our data is symmetrical, we should prefer the mean as the measure of central tendency and standard deviation (SD) as the measure of dispersion. However, when it is skewed or has extreme values, we should prefer the median as a measure of central tendency and the IQR or median absolute deviation (MAD) as the dispersion measure

Let us calculate MAD for the data (40,40,44,46,46,45,39,80,11) (Table 10.1). Follow these steps:

- Step 1: Arrange the data in an ascending order: 11,39,40,40,44,45,46,46,80.
- Step 2: Choose the middle value (5th) as the median: 44.
- Step 3: Subtract the median from each value and calculate its absolute value as shown in the following table.
- Step 4: Arrange |x-median| in an ascending order and find the middle value (5th): 0,1,2,2,4,4,5,33,36. Here the MAD is 4. However, the mean and standard deviations for this data are 43.44 and 16.49, respectively. The standard deviation is very high as it is affected by the extreme values (11,80). However, the mean (43.44) is closer to the median (44).

10.1.8 What Is Meant by the Distribution of Data?

The distribution of data shows how often each value or values in the intervals occur in the data. The following graph shows how the age (years) of the patients is distributed. The height of the bar gives the count (frequency) of the patients lying in an interval (Fig. 10.1). The distribution shows that the highest number of patients (14) have ages between 41 and 51 followed by 9 who are between 31 and 41 years.

The subject of statistics includes a large number of theoretical distributions such as Binomial, Beta, Cauchy, '*F*', Geometric, '*t*', Inverse-Normal, Normal, Negative-Binomial, Poisson, and Uniform. However, the Normal distribution is the most common. These are also called probability distributions.

10.1.9 What Is a Normal Distribution Curve?

A normal distribution curve is bell-shaped and symmetrical where the mean, median, and mode are equal (Fig. 10.2). It is also called a Gaussian distribution and most of the large populations in nature follow this pattern. A standard normal

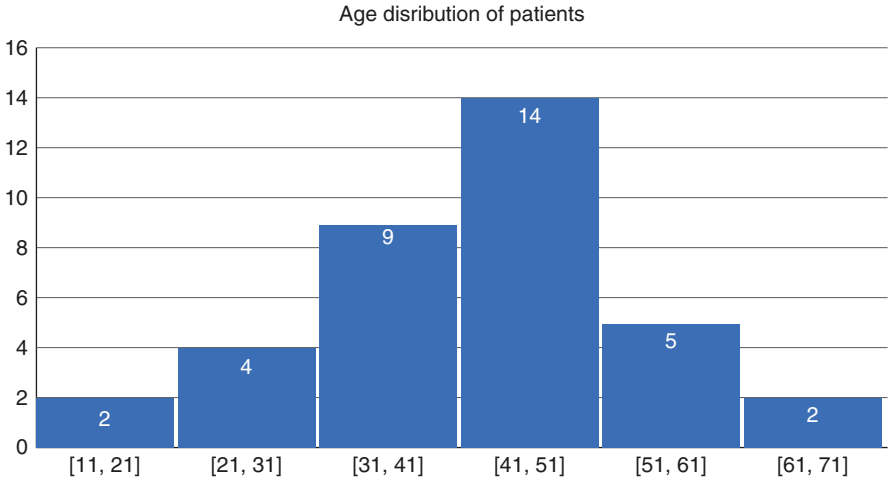


Fig. 10.1 Age distribution of Patients

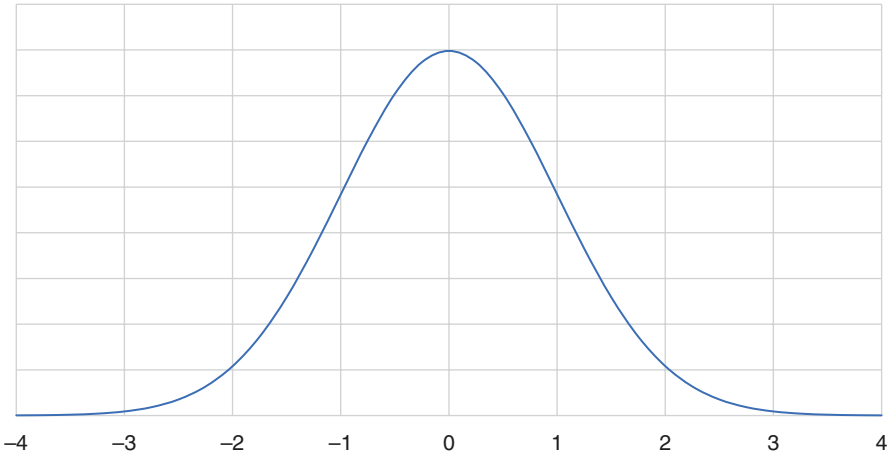


Fig. 10.2 Normal distribution curve

distribution curve has a mean of 0 and a standard deviation of 1. A normal distribution curve has a skewness of 0 and kurtosis 3.

10.1.10 What Is Skewness and What Are Its Implications?

When we plot the data, it may not give a perfect symmetric curve. The lack of symmetry about the mean is generally known as skewness. The curve can be negatively skewed with a longer tail on the left or positively skewed on the right side as shown in the following figures. Skewness indicates a direction and deviation from normality. A normal curve is symmetrical and has a skewness of zero. However, the converse is not necessarily true, i.e., a curve having a skewness of zero may not

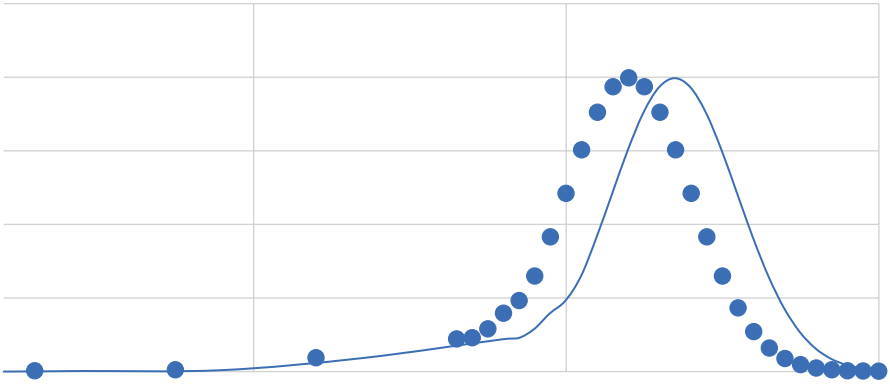


Fig. 10.3 Positively skewed curve

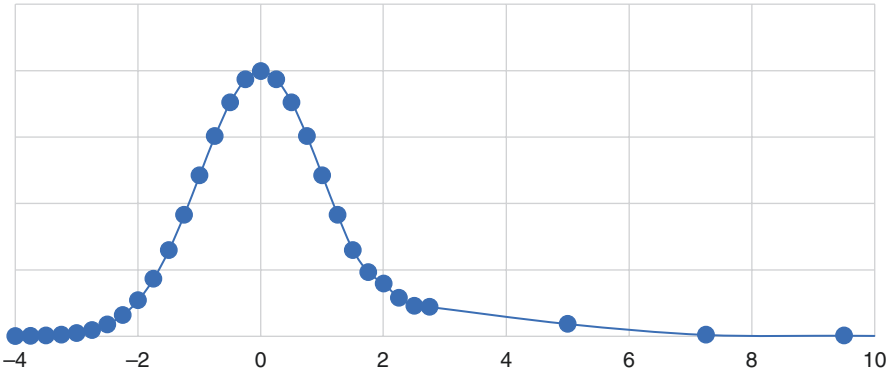


Fig. 10.4 Negatively skewed curve

necessarily be symmetrical. Figures 10.3 and 10.4 give positively (longer tail on the right) and negatively (longer tail on the left) skewed curves.

10.1.11 What Is a Kurtosis and What Are Implications?

Kurtosis is a measure of the peakedness in a curve. A normal curve is known as a mesokurtic curve, which has a moderate peak and a kurtosis of 3. A Leptokurtic curve is a high peaked curve with a low kurtosis (<3), has a lower spread of the data indicated by the lighter tails and lacks outliers (Fig. 10.5). A Platykurtic curve has a high kurtosis (>3), has a wider spread of the data with heavier tails, and may include outliers. From the kurtosis number, one can get a clear visualization of the spread of the data.

Curves are classified as follow:

- Platykurtic (Kurtosis <3.0).
- Mesokurtic (Kurtosis =3.0).
- Leptokuric (Kurtosis >3.0).

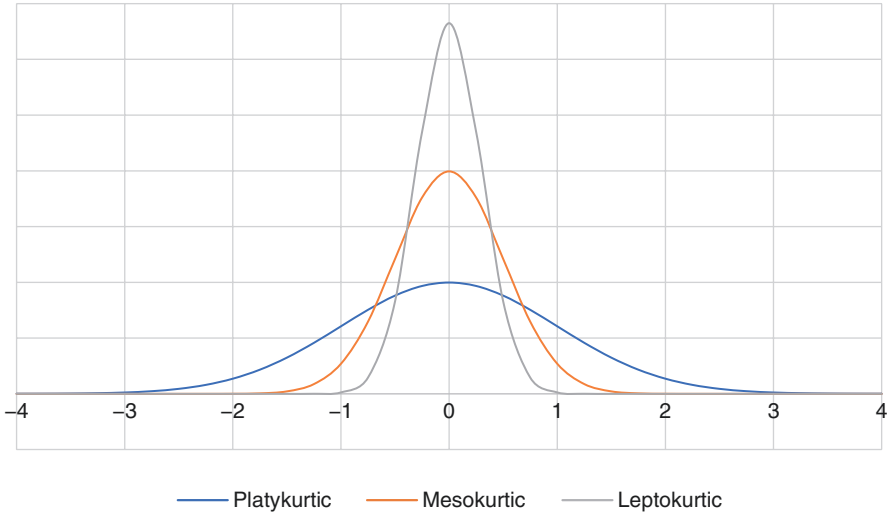


Fig. 10.5 Curves with varying kurtosis levels

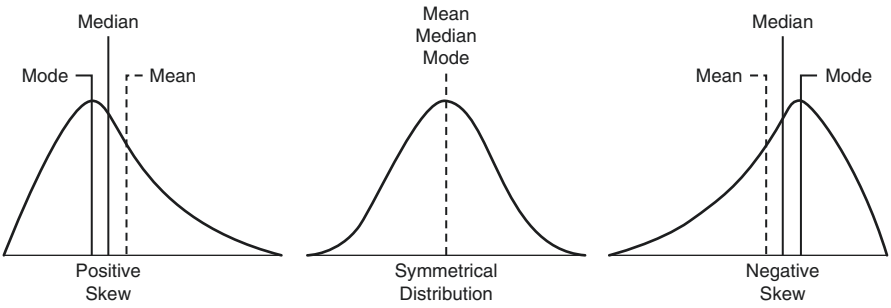


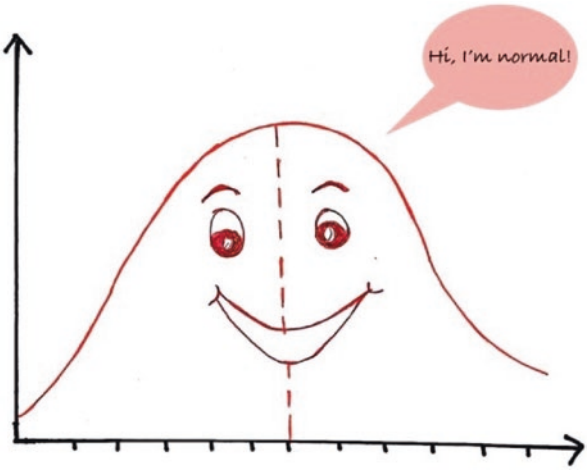
Fig. 10.6 Showing Mean, Median, and Mode with respect to skewness

10.1.12 Is There Any Relationship Between Skewness and the Three Measures of Central Tendency (Mean, Median, and Mode)?

Most textbooks show a general relationship between the skewness and three measures as shown below (Fig. 10.6):

- *Normally distributed data: Mean = Median = Mode*
- *Positively skewed data: Mean > Median > Mode.*
- *Negatively skewed data: Mean < Median < Mode.*

However, the latest findings contradict the above relation/ship in the case of positively skewed data of adult residents across US households where the mean was found to be less than the median and mode [2].



10.1.13 What Is an Outlier in a Data Set?

Before data are subjected to analysis, they are examined and cleaned. We may unexpectedly get very high or low or both values in the data which severely affect results and interpretation. An outlier in a data set is a value that is located at an abnormal distance from most of its values. The minimum or maximum or both values are often outliers but not always. One can think of a value in normally distributed data that is three or more standard deviations away from either side of the mean. According to Maddala, 'An outlier is an observation that is far removed from the rest of the observations'. [3] Outliers are extreme observations and may impact the analysis and lead to misleading interpretation and inferences, if proper precautions are not taken. There are a number of methods available in the literature for detecting and handling outliers.

10.1.14 Why Do Outliers Occur?

Some of the common causes for outliers are:

- Variability in the population and sampling fluctuations.
- Malfunctioning of the instruments, measurements, and human errors.
- Sampling from a highly asymmetric distribution.
- Error in data transmission or transcription.
- Mixing of two distributions.

10.1.15 How Can We Detect Outliers?

The statistical literature supports a number of methods (graphical and model based) for the detection of outliers in a dataset. However, the Box and Whiskers plot is the

most commonly used in descriptive statistics and is based on quartiles. In a boxplot an outlier is a value that is less than the Q_1–1.5(IQR) or greater than Q_3 + 1.5(IQR), where Q_1 and Q_3 are the first and third quartile; IQR is the interquartile range and is equal to Q_3-Q_1. Figure 10.7 shows the Box and Whisker plot of data showing 25, 20, 19.5, and 1 as outliers.

10.1.16 How Should We Handle Outliers?

There are a number of methods available in the literature for identifying and handling outliers. Aguinis reviews and discusses best-practice recommendations for defining, identifying, and handling outliers [4]. The subject of handling outliers is not as simple as it looks. The simplest method is to correct the wrong value or to drop the records containing the outliers. Dropping the records introduces the bias in the results and we suggest the readers to refer to the recommendations of the said paper before deleting the outlier.

10.1.17 What Is Data Heaping?

Data heaping or rounding is a measurement error that arises from an intentional coarsening of the data. Heaped data includes exact and rounded-off values. If the heaping occurs at random, then it does not pose any problem. However, this is not usually so and results in misleading inferences about the parameters of the population if it is ignored. The most common example of data heaping is self-reported income and age from the retrospective data. Statistical literature includes several methods to overcome and handle the heaping of data.

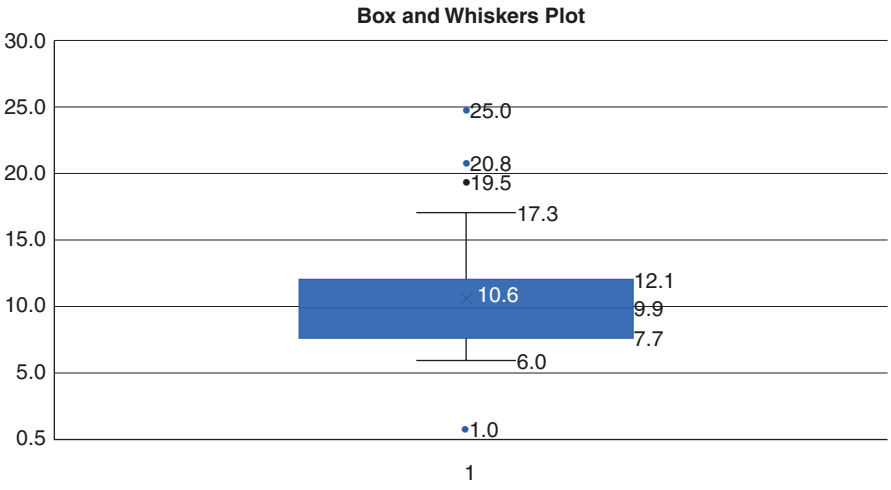


Fig. 10.7 Box and Whiskers plot

10.1.18 What Is the Difference Between a Parameter and an Estimate?

A parameter is a characteristic, say mean, proportion, etc. of a population. It is also called the true value of the population, whereas an estimate is the characteristic of a sample that predicts the true value on the basis of sample data. For example, if we measure the height of all Indian adult males and calculate its mean, it will be a true value and is called the parameter. But measuring the height of an entire population is not feasible and is also an expensive proposition. Thus, we try to estimate the true population value on the basis of a small sample. However, if we calculate the mean on the basis of a sample, it will be an estimate which may or may not be equal to the parameter (true mean) as different samples give rise to different estimates. We try to estimate the true value precisely with some degree of confidence (normally 95%) and with lesser sampling fluctuations or narrow confidence intervals. When we estimate a parameter, we report a confidence interval along with the estimate.

10.1.19 What Are Inferential Methods?

The most important component of biostatistics is inferential statistics which includes a large number of statistical methods which help in testing hypotheses and drawing inferences about the population parameters based on the sample data. Inferential methods are often used to compare differences between a treatment group and a prespecified value or between different treatment groups. Inferential statistics can tell us with a certain degree of confidence, whether the difference between the two groups is a true difference or it is likely due to chance outcomes as a result of sampling fluctuations (different samples resulting in different estimates). Based on a single sample, inferential statistics can also suggest the probable value of a population parameter (mean or proportion) lying in the range popularly known as the confidence interval. These methods are different for qualitative variables than for quantitative ones.

They also vary further depending upon the sample size or sampling distribution. It has been theoretically established that when the sample size is large, the distribution pattern of the mean/proportion tends to be Gaussian (normal having a bell-shaped curve). Thus, when the sample size is large and the distribution is Gaussian (normal) we use parametric inferential methods, otherwise we use non-parametric methods.

10.1.20 What Are Inferential Methods for Proportions?

Inferential methods for proportions are commonly used for calculating confidence intervals as well as for testing differences between statistics which are estimated on the basis of proportions such as the prevalence of a disease, odds ratio (OR), relative risk (RR), sensitivity, and specificity. These methods also include equivalence tests

(superiority, equivalence, noninferiority) for testing whether the two groups are essentially equivalent and whether the difference, if any is medically relevant. These methods generally use the Z-test, Chi-square test, likelihood ratio test, Fisher Exact test, and McNemar test. Users have to be cautious about these tests because of their limitations. The chi-square test is suited for large sample sizes whereas the Fisher Exact test can be used with small sample sizes for a 2×2 classification. A sample is considered large enough for using the chi-square test if the expected counts in every cell are 5 or more. The McNemar test is used for matched-pair proportions and is mainly used in pre-post designs. Except for the Z-test, all the above-said tests are non-parametric and do not require the assumption of normality.

10.1.21 What Are Inferential Methods for Means?

10.1.21.1 Parametric Methods

A large number of parametric inferential methods are available for testing the difference of means between groups when the sample size is large and the underlying distribution of the study variable follows a Gaussian pattern. Thus, the normality assumption is tested before going for the analysis. These methods include Student unpaired and paired t-tests for comparing two means or a mean with a specified value one way and two-way ANOVA, Repeated measure ANOVA, etc. The Analysis of variance (ANOVA) is a technique for comparing three or more means and uses the F-test for testing the null hypothesis of the equality of means. Repeated measures ANOVA is the equivalent of the one-way ANOVA, but for related groups, and is the extension of the Student paired t-test.

10.1.22 What Are Post Hoc Tests?

ANOVA is used when we want to test differences in the means among three or more groups. ANOVA uses Fisher's F-test to do this. If the p-value corresponding to the F-test is less than the threshold (0.05) then we reject the null hypothesis of equality of all means and look for which pairs of means differ significantly. The F-test does not tell which pairs of means differ. Post hoc tests are applied to test individual pairs. The tests are –Bonferroni, Tuckey, Dunnett, Scheffe, Fisher's LSD, Newman-Keul, and Dunkan. They are an integral part of ANOVA for making multiple comparisons. The Bonferroni and Tuckey tests are the most used in the medical field.

10.1.23 What Are the Assumptions of the Student *t*-test?

The Student t-test is the most commonly used for comparing a mean with a prespecified value (constant) or two unpaired means. The assumption for a t-test is that the measurement applied to the data collected follows a continuous numeric scale (e.g., body weight, cholesterol level) or an ordinal scale, such as the scores

for an IQ test. The Student t-test assumes that the values within each group should be independent and the means normally distributed. The test also requires homogeneity of variance (i.e., the standard deviations of the samples are approximately equal).

10.1.24 What Are the Tests Used for Testing the Normality of the Data?

Many statistical procedures such as t-tests, linear regression analysis, discriminant analysis, and Analysis of Variance (ANOVA) require assumptions of normality. If this is not present the inferences drawn may not be reliable and valid. There are three common methods for assessing normality. These are graphical methods (histograms, boxplots, Q-Q plots), numerical methods (skewness and kurtosis) and formal normality tests – the Shapiro–Wilk test, Kolmogorov–Smirnov test, Lilliefors test, and the Anderson–Darling test. A standard statistical software package has all three tests to assess normality. One can have an idea of the normal curve from the histogram of the data. If the curve is bell shaped, the data seems to follow a normal distribution. Skewness measures the lack of symmetry and a normal curve has a value of zero. A larger negative value means the curve is negatively skewed and a larger positive value, positively skewed. Kurtosis measures the peakedness of the curve. A normal curve has a kurtosis of three. If the kurtosis is greater than three then the curve will have a high peak and flatter tails. If it is less than three, the curve will have a low peak. However, to be more specific, data can be tested for normality using formal tests. Results of a simulation study [5] show that the Shapiro–Wilk test is the most powerful normality test, followed by the Anderson–Darling, Lilliefors and Kolmogorov–Smirnov tests. However, the power of all four tests is still low for a small sample size. If the p-value of the test is less than 0.05, the data does not follow the Gaussian pattern or does not have a normal distribution.

10.1.25 What Should We Do If the Normality Assumption Fails?

If the data fails for normality, we should first apply the appropriate transformations (log, square root, reciprocal, arcsine, etc.) for transforming the data to normal and then apply the test. If the transformation does not help, we should go for the corresponding non-parametric test.

10.1.25.1 Non-Parametric Methods

Non-parametric methods are distribution-free methods and particularly used when the normality assumption fails. These methods can be applied when the data is ordinal or the sample size is small. They may not be as efficient as parametric tests when the data is normally distributed. Table 10.2 shows the parametric and their corresponding non-parametric tests for comparing two or more groups.

Questions and answers related to testing of hypotheses:

Table 10.2 Parametric and non-parametric methods

Number of groups	Parametric	Non-parametric
Two	Unpaired student t-test	Mann–Whitney U test
Two	Paired student t-test	Wilcoxon signed-rank test
More than two	ANOVA	Kruskal–Wallis

10.1.26 What Are the Null and Alternative Hypotheses?

A null hypothesis is a statement about the population(s) saying such as there is 'no effect' of a factor or 'no association' between two attributes or 'no difference' in characteristics (mean, efficacy, accuracy) in the population. The null hypothesis eliminates any prejudice or presumption with respect to a population. Whereas an alternative hypothesis is a contrasting statement that says, 'there is an effect' or 'there is an association' or 'there is a difference' among the populations. The alternative hypothesis is what we might believe to be true or hope to be true.

For example, we want to compare the efficacies of two drugs using the Student t-test we can set the hypothesis as follow:

- Null Hypothesis (H_0): There is no difference in the efficacy of drug A and B.
- Alternative Hypothesis (H_1): The drugs differ in their efficacy.
- Alternative Hypothesis (H_1): Drug A has a higher efficacy than B.
- Alternative Hypothesis (H_1): Drug A has a lower efficacy than B.

In case when we want to test the association, we set the hypothesis as follow:

- Null Hypothesis (H_0): There is no association (correlation coefficient = 0).
- Alternative Hypothesis (H_1): There is an association (correlation coefficient $\neq 0$).

We can also set the two hypotheses as follows:

- Null Hypothesis (H_0): Prevalence of a disease (p) ≤ 0.20.
- Alternative Hypothesis (H_1): Prevalence of a disease (p) > 0.20.

When there are several means to be tested using ANOVA, the hypotheses are written thus:

- Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$

10.1.27 What Errors Do We Make While Testing a Hypothesis?

Testing a hypothesis is commonly known as the null hypothesis significance testing (NHST) where we find the evidence against the null hypothesis using various statistical tools such as the Student *t*-test, F-test, and Z test. If the evidence is strong enough ($p < 0.05$) we reject the null hypothesis in favour of an alternate hypothesis. Since all statistical significance testing is based on a fairly small single sample that is associated with fluctuations, errors are bound to creep into the testing process.

Two kinds of errors known as type-I and II occur while testing a hypothesis. For example, if in reality, two drugs have the same efficacy but on the basis of the *p*-value, we prove that they are different this is called the type-I error or alpha error. Let us consider the other situation where in reality the drugs differ but we are showing that they are not. This is called the type-II error or beta error. We define these errors more precisely as follows:

- Type-I error: Probability of rejecting a true null hypothesis is also known as an Alpha(α) error or false-positive. This probability is generally kept at 5% or sometimes 1% and this level is called the significance level.
- Type-II error: The probability of not rejecting the false null hypothesis is also known as the Beta(β) error or false negative. This probability is generally kept at 20%.

10.1.28 What Is the Relationship Between Type-I and Type-II Errors and Power?

The Type-I error is the probability of a false-positive result and is conventionally kept below 0.05 (5%). If we reduce this error, the Type-II error increases and consequently, the power of the test decreases. Thus, the lower the level of significance, the lower is the power. The main disadvantage with a low-powered statistical test is that it is difficult to reject the false null hypothesis. Such a test may not permit us to recommend a drug even if it is better than the control.

10.1.29 What Is the Power of a Statistical Test?

The power of a statistical test is the probability of rejecting the null hypothesis when the alternative hypothesis is true and is equal to 1-β. It is the probability of detecting significant differences. For example, an investigator sets a power of 80%. This will enable him or her, to observe an effect of that size or larger in his study 80 times out of 100. Increasing the sample size enhances the power. With high power, even the smaller differences in populations can be detected.

10.1.30 What Is a Confidence Level and Confidence Interval?

The confidence level is a measure of uncertainty associated with a *sampling method*. The higher the confidence level, the lower is the uncertainty. Suppose we want to estimate the population mean. We draw a random sample and calculate the mean. If we repeat this process 100 times, we may get different values of the mean because of sampling fluctuations. Suppose we also compute the interval estimate for each sample along with the sample means. Some interval estimates would include the true population mean and some would not. A 95% confidence level means that we

would expect 95% of the interval estimates to include the population mean. This does not mean there is a 95% probability that the population mean is in the interval estimate.

The confidence interval is a range of plausible values in which we are fairly sure our unknown true value (for example, mean, proportion) lies. The confidence interval is always associated with the confidence level. The width of the confidence interval is indicative of the reliability of the estimate. The narrower the interval, the higher is the reliability. For a given confidence level, the confidence interval depends upon variability (standard deviation) as well as sample size. If the standard deviation is high or sample size is low, the confidence interval will be wider. We can use the following formula for calculating the lower and upper limits of the confidence interval of a mean:

$$\text{Lower limit of confidence interval is } \bar{X} - Z\alpha/2 \times \left[\sigma/\sqrt{n}\right].$$

$$\text{Upper limit of confidence interval is } \bar{X} + Z\alpha/2 \times \left[\sigma/\sqrt{n}\right].$$

where

$\bar{X}$ = Mean, σ = Standard deviation, Z = Standard score, α = Confidence level.

For a 90% confidence level(α) the value of Z is 1.645, for 95%, 1.96 and for 99%, 2.576.

Example: The sample data (10, 1, 2, 15, 10, 14, 13, 9, 9, 10, 11, 10, 5, 6, 6) has a mean of 8.73 and a sample standard deviation is 4.06. Using the above confidence interval formula, the 95% confidence interval of the mean is 6.68–10.79. Half of this interval width (2.05) is called the margin of error or precision.

10.1.31 What Are One- and Two-Tailed Tests?

While calculating the sample size to test a hypothesis, the researcher must specify whether the researcher is using a one- or two-tailed test. A one-tail alternative hypothesis test is a one-sided test. For example, drug 'A' is better than 'B' or drug 'A' is inferior to 'B'. A two-tail alternative hypothesis test is a two-sided test. For example, drugs 'A' and 'B' differ. If we do not have any idea about the performance of the drugs, we should choose the two-tail test; otherwise, we can use a one-tail test. The advantage of a one-tail test is that it requires a smaller sample size for testing the significance compared to a two-tail test.

10.1.32 What Is the *p*-value?

The *p*-value is the most frequently used inferential statistics for testing a null hypothesis. The *p*-value developed by Sir Ronald Fisher almost a century ago is highly controversial and is generally misunderstood by researchers. It is the probability of an observed or more extreme result assuming that the null

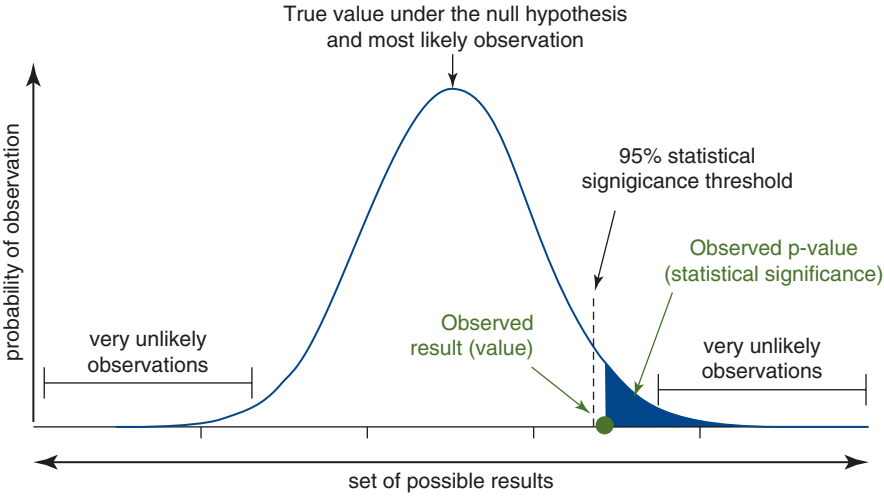


Fig. 10.8 A normal probability distribution curve. Figure Downloaded from the Internet

hypothesis is true. For example, if a researcher wants to test whether a new drug is better than an old one, he applies the *t*-statistic and gets a '*t*-value' of say 2.0. Then he looks for the probability of getting a result of 2.0 or more ($t \geq 2.0$). This value which is shown under the blue area in Fig. 10.8 is the probability and is called the *p*-value.

The *p*-value is automatically calculated by the software. If this is found to be less than or equal to the significance level (usually $p = 0.05$), we have a strong evidence against the null hypothesis and can reject the null hypothesis. However, rejecting the null hypothesis does not imply that it is false. Similarly, when the $p > 0.05$ it is incorrect to say there is 'no effect' or that the 'null hypothesis is true'. When $p \leq 0.05$, the safer inferential statement we can make is that there is at most a 5% probability that our results are compatible with the null hypothesis. The ASA's recent interpretation of the *p*-value clearly says, '*p*-values can indicate how incompatible the data are with a specified statistical model'.

10.1.33 What Are the Most Common Misinterpretations of the *p*-value?

(i) The *p*-value is the probability that the null hypothesis is true when $p > 0.05$.
(ii) 1-minus the *p*-value is the probability that the alternative hypothesis is true.
(iii) When $p > 0.05$, the null hypothesis is false or should be rejected.
(iv) When the *p*-value >0.05 there is no effect.
(v) A statistically significant result is also medically important.

10.1.34 What Is the New Interpretation of the *p*-value by the American Statistical Association (ASA)?

The *p*-value suggested by Fisher has now been interpreted by the American Statistical Association (ASA) [6] as 'Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value'. If the *p*-value is less than the threshold (0.05), then there is a strong evidence against the null hypothesis.

10.1.35 What Are the Six Principles Suggested by the American Statistical Association (2016) on *p*-values?

1. '*p*-values can indicate how incompatible the data are with a specified statistical model'.
2. '*p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. 'Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold'.
4. 'Proper inference requires full reporting and transparency'.
5. 'A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result'.
6. 'By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis'.

10.1.36 Is a Very Low *p*-Value Indicative of a Large Effect Size?

A low *p*-value is not indicative of the effect size (difference between two proportions or means), i.e., that it will be very large. The *p*-value is sample size dependent. Even a very small observed difference turns out to be statistically significant if the sample sizes are large. According to the ASA guidelines inferences should not be simply drawn on the basis of *p*-value alone.

References

1. Indrayan A. Medical biostatistics. 3rd ed. London: Chapman & Hall/CRC Biostatistics Series; 2013.
2. von Hippel PT. Mean, median, and skew: correcting a textbook rule. J Stat Educ. 2005;13:2. https://doi.org/10.1080/10691898.2005.11910556.
3. Maddala GS. Outliers. Introduction to econometrics. 2nd ed. New York: MacMillan; 1992. p. 89. isbn:978-0-02-374545-4.

4. Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organ Res Methods 2013;00: 1–33. Accessed from https://doi.org/10.1177/1094428112470848. Accessed October 4, 2020.
5. Rajali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J Stat Mod Analyt. 2011;2:21–33.
6. Wasserstein RL, Lazar NA. The ASA's statement on P-values: context, process, and purpose. Am Stat. 2016;70:129–33.