Scientific Tool

Biostatistics without the mathematics. Part 1–Descriptive statistics

Avijit Hazra

Department of Pharmacology, Institute of Postgraduate Medical Education & Research, Kolkata, India

ABSTRACT

Biostatistics is now an integral part of medical research. Knowledge of statistics is also becoming mandatory to understand most medical literature. The word data denotes the values of variables. It is important to understand the types of data and their mutual interconversion. The raw data for statistical analyses come from experiments or observations and can be numerical or categorical. Numerical variables may be continuous or discrete. Categorical data are described in terms of frequencies, proportions, or percentages. The applications of statistics in medical sciences can be categorized as descriptive statistics, inferential statistics and statistical modeling. Descriptive statistics implies summarizing a collection of data from a population. The observations within a sample tend to cluster around a central location, with more extreme observations being less frequent. The extent to which the observations cluster is summarized by measures of central tendency, while the spread is described by measures of dispersion. The measurement of central tendency include mean, median and mode, while the measurement of dispersion include range, standard deviation, mean deviation and others. The population mean, median, standard deviation, etc., are known as the parameters, while the sample mean, median, standard deviation, etc., are known as the statistics. We can hardly know the true values of parameters. However, we can obtain a reasonable point estimate of a parameter and define an interval in which the true population value is likely to lie with a certain level of confidence. This range is known as the confidence interval (CI). A CI of a parameter that has X% confidence is defined as an interval so that the parameter will lie within this interval with probability X. Conventionally, a 95% CI is used for most analyses. Understanding patterns in data sets and the distribution of the corresponding population are important components of descriptive statistics. The most common distribution is the normal distribution, which is depicted as the well-known symmetrical bell-shaped Gaussian curve. Familiarity with other distributions such as the binomial and Poisson distributions is also helpful. Various graphs and plots have been devised to summarize data and trends visually. Some plots, such as the box-and-whiskers plot and the stem-and-leaf plot are less familiar but provide useful summaries in select situations.

Keywords: Descriptive statistics, Confidence interval, Normal distribution, Boxplot, Stem-and-leaf plot.

Received Date : 31-05-2013
Revised Date : 01-08-2013
Accepted Date : 05-08-2013

DOI: 10.5530/rjps.2013.3.1

Address for correspondence
Dr. Avijit Hazra
Department of Pharmacology
Institute of Postgraduate Medical Education & Research, 244B Acharya J. C. Bose Road, Kolkata – 700020
Telephone: 033-24750764 / 0-98311-88172
Fax: 033-24764656
E-mail: blowfans@yahoo.co.in

www.rjps.in

Biostatistics is a broad discipline encompassing the application of statistical theory and practice to understand living systems. Today, the practice of designing and conducting biomedical observations and experiments, presenting the data accruing therefrom and interpreting the results, would be impossible without applying statistics. There are multiple reasons for this. The major factor is the enormous variability shown by living systems coupled with our inability to understand the sources of such variations and to control them during our observations or experiments. Variations may be due to characteristics of individual subjects, the effects of interventions, measurement errors or simply unknown 'chance' factors. The application of statistics allows adjusting for

these variations to reach meaningful conclusions using representative samples drawn from the population. Another important reason for applying statistics is that, most of the time, researchers in the life sciences are interested in small changes or effects that require mathematical tools to ascertain if these changes or effects are significant enough to be enthused or bothered about. Therefore, in biomedical research, statistical analysis is essential for making sense of inevitable uncertainty. Statistical description is also used in everyday data presentation outside the ambit of biomedical research.

Before the advent of computers and statistical software, researchers and others dealing with statistics had to do most of their analysis by hand, taking recourse to books of statistical formulas and statistical tables. This required one to be proficient in the mathematics underlying statistics. This is no longer necessary since increasingly user-friendly statistical software takes the drudgery out of calculations and obviates the need for looking up statistical tables. Therefore, today, understanding the applied aspects of statistics suffices for the majority of researchers, who do not require delving deep into the mathematical part of statistics, in order to make sense of data that they generate or scrutinize.

The applications of biostatistics can broadly be envisaged as covering three domains – descriptions of patterns in observed values through various descriptive measures (**descriptive statistics**), drawing conclusions regarding populations through various statistical tests applied to sample data (**inferential statistics**), and applications of modeling techniques to understand relationship between variables, sometimes with the goal of prediction (**statistical modeling**). In this article, we will look at the descriptive uses of statistics without delving into mathematical depths. This is not to deny the mathematical underpinnings of statistics – these can be found in statistics textbooks. Our goal here is to present the concepts and look at the applications from the point of view of the applied user of biostatistics.

DATA AND VARIABLES

Data constitutes the raw material for statistical work. They are records of measurement or observations or simply counts. A **variable** refers to a particular character on which a set of data are recorded. Data are thus the values of a variable. Before a study is undertaken it is important to consider the nature of the variables that are to be recorded. This will influence the manner in which observations will be undertaken, the way in which they will be summarized and the choice of statistical tests that will be used.

At the most basic level, it is important to distinguish between two types of data or variables. The first type includes those which are defined by some characteristic, or quality, and is referred to as **qualitative variable**. Because qualitative data are best summarized by grouping the observations into categories and counting the numbers in each, they are often referred to as **categorical variables**. The second type includes those that are measured on a numerical scale and is called **quantitative variable**. Since quantitative variables always have values expressed as numbers and the differences between values have numerical meaning, they are also referred to as **numerical variables**. They have also been called **metric variables** as their value is obtained through measurement using an appropriate measuring scale or device.

A qualitative variable can be a **nominal variable** or an **ordinal variable**. A nominal variable covers categories that cannot be ranked; and no category is more valuable than another. The data is generated simply by naming the appropriate category to which the observation belongs. An ordinal variable has categories that follow a logical hierarchy and hence can be ranked. We can assign numbers (scores) to nominal and ordinal categories; although the differences among those numbers do not have numerical meaning. However, category counts do have numerical significance. A quantitative variable can be **continuous** or **discrete**. A continuous variable can, in theory at least, take on any value within a given range, including fractional values. A discrete variable can take on only certain values within a given range; these values are usually integers. Often certain variables, like age or blood pressure, are treated as district variables although strictly speaking they are continuous. A special case may exist for both categorical or numerical variables, when the variable in question can take on only one of two numerical values or belong to only one of two categories; these are known as **binary or dichotomous data**, as opposed to **non-binary** or **polychotomous** data that can take more than two values.

To illustrate the above data types, let us consider the human hair as an example. Hair color would be a categorical variable, but hair length would be a numerical one. Since there is no natural hierarchy of hair color, nor is any ranking possible, hair color would be a nominal variable. However, hair loss may be an ordinal variable if it is expressed as none, partial and total, for example. Hair length would be a continuous variable, but we can treat is as discrete if we are recording it only to the nearest millimeter. It is possible to convert numerical data

to categorical. Thus, after recording hair length we may classify the subject into long, medium or short hair category. If we are interested in only two categories of hair length, say long or short, then this becomes a binary variable.

Numerical data can be recorded on an interval scale or a ratio scale. On an **interval scale**, the differences between two consecutive numbers carry equal significance in any part of the scale, unlike the scoring of an ordinal variable ('ordinal scale'). For example, when measuring distance, the difference between 1 and 2 meters is the same as the difference between 1000 and 1001 meters. Ratio scale is a special case of recording interval data. With interval data the zero value can be arbitrary, such as the position of zero on some temperature scales – the Fahrenheit zero is at a different position to that of the Celsius scale. With **ratio scale**, zero actually indicates the point where nothing is scored on the scale ('true zero'), such as zero on the absolute or Kelvin scale of temperature. Only on a ratio scale, can differences be judged in the form of ratios. 0°C is not zero heat, nor is 26°C twice as hot as 13°C; whereas these value judgments hold with the Kelvin scale. In practice, this distinction is not tremendously important so far as the handling of numerical data in statistical tests is concerned.

Changing data scales is possible so that numerical data may become ordinal, and ordinal data may become categorical (even dichotomous). This may occur because the researcher is not confident about the accuracy of the measuring instrument, is unconcerned about loss of fine detail, or where group numbers are not large enough to adequately represent a variable of interest. It may also make clinical interpretation easier. For example, in ECG monitoring, the extent of ST-segment depression indicates the degree of myocardial ischemia. Although, theoretically a continuous variable, it is generally accepted that ST-segment depression greater than 1.0 mm indicates significant ischemia, so that ST-segment depression less than this value is categorized as 'no ischemia'. This results in some loss of detail, but clinically this is more convenient to deal with and is therefore widely accepted.

When exploring the relationship between variables, some can be considered as dependent (**dependent variable**) on others (**independent variables**). For instance, when exploring the relationship between height and age, it is obvious that height depends on age, at least until a certain age. Thus, age is the independent variable, which influences the value of the dependent variable height. When exploring the relationship between multiple variables, usually in a modeling situation, the value of the **outcome (response) variable** depends on the value of **predictor (explanatory) variables**. In this situation, some variables may be identified that cannot be accurately measured or controlled and only serve to confuse the results. They are called **confounding variables or confounders**. Thus, in a study of antihypertensive drug effect, the change in blood pressure (outcome) would depend on the dose and maybe on the age of the patient (predictors). However, it will also be confounded by salt or sodium intake which cannot be accurately measured or strictly regulated.

Numerical or categorical variables may sometimes need to be **ranked**, that is arranged in ascending order and new values assigned to them in serial order. Values that tie are each assigned average of the ranks they encompass. Thus, a data series 2, 3, 3, 3, 3, 5, 7, 9, 15 can be ranked as 1, 3.5, 3.5, 3.5, 3.5, 6, 7, 8, 9, since the four 3s encompass ranks 2, 3, 4, 5 giving an average rank value of 3.5. Note that when a numerical variable is ranked, it gets converted to an ordinal variable. Ranking obviously does not apply to nominal variables because their values do not follow any order.

DESCRIPTIVE STATISTICS

Descriptive statistics means summarizing a collection of data from a group. Traditionally, summaries of sample data ('statistics') have been denoted by Roman letters (e.g., $\bar{x}$ for mean, SD for standard deviation, etc.) while summaries of population data ('parameters') have been denoted by Greek letters (e.g., μ for mean, σ for standard deviation, etc.).

For numerical data, the individual observations within a sample or population tend to cluster around a central location, with more extreme observations being less frequent. The extent to which observations cluster is summarized by **measures of central tendency** while the spread is described by **measures of dispersion**.

MEASURES OF CENTRAL TENDENCY

The common measures of central tendency include mean, median, and mode. The **mean** (or more correctly, the **arithmetic mean**) is calculated as the sum of the individual values in a data series, divided by the number of observations. The mean is the most commonly used measure of central tendency to summarize a set of numerical observations. It is usually stable and reliable. However, the presence of extreme values (outliers) can distort the mean. It should not ordinarily be used in describing categorical variables because of the arbitrary nature of category scoring. It may, however, be used to summarize category counts. Note also that the mean value need not be an actual value in the sample from which it is derived.

The median denotes the point in a data series at which half the observations are larger and half are smaller than it. If the values in a data series are arranged (either in ascending or descending order), then the **median** is the middle value (for an odd number of observations) or the average of the two middle values (for an even number of observations). It is a useful summary measure, particularly if the distribution of the data is not symmetrical, since it is less sensitive to extreme values than the mean. The median value is also known as the 50^{th} percentile value.

The **mode** is the most frequently occurring value in a data series. It is not often used, for the simple reason that it is difficult to pinpoint a mode if no value occurs with a frequency markedly greater than the rest. Two or more values may occur with equal frequency, making the data series **bimodal** or **multimodal**.

The relationship between the three measures of central tendency depends on the shape of the data distribution. In a unimodal symmetrical distribution (such as the normal distribution shown in **Figure 1**), all three measures are identical, but in a skewed distribution they will usually differ. The mode would simply be the most frequently occurring value (the highest point on the distribution curve); the mean is pulled to one side by the influence of a relatively small number of very high or very low values; and the median lies between the two, dividing the distribution into two equal areas under the curve.

Two other measures of central tendency are geometric mean and harmonic mean. The **geometric mean** of a series of *n* observations is the *n*th root of the product of all the observations. It is always equal to or less than the arithmetic mean. It is not often used, but is a more appropriate measure of central location when data has been recorded on a logarithmic scale. Interestingly, the logarithm of the geometric mean is the arithmetic mean of the logarithms of the observations. As such, the geometric mean may be calculated by taking the antilog of the arithmetic mean of the log values of the observations. The **harmonic mean** of a set of non-zero positive numbers is obtained as the reciprocal of the arithmetic mean of the reciprocals of these numbers. It is seldom used in biostatistics.

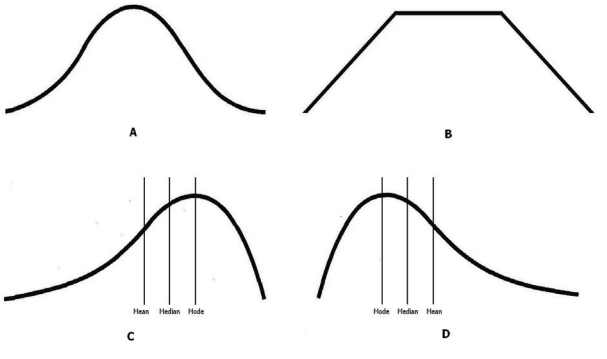

Figure 1: Examples of frequency distributions – A) Symmetric and normal B) Symmetric but not normal C) Asymmetric, negatively skewed D) Asymmetric, positively skewed.

Oftentimes data is presented as a **frequency table**. If the original data values are not available, a **weighted average** can be estimated from the frequency table by multiplying each data value by its frequency (the number of cases in which that value occurs), summing up the products and dividing the product sum by the sum of the frequencies (total number of observations). A frequency table of numerical data may report the frequencies for class intervals (the entire range covered by the observations being broken up into a convenient number of intervals) rather than for individual data values. In such cases, we can calculate the weighted average by using the mid-points of the class intervals and their corresponding frequencies. However, in this instance the weighted mean may vary slightly from the arithmetic mean of all the raw observations.

MEASURES OF DISPERSION

The spread, or variability, of a data series can be readily described by the **range**, which is the interval between minimum and maximum values. However, the range does not provide much information about the overall distribution of observations, and obviously responds only to the two extreme values.

A more useful estimate of the dispersion can be obtained by first arranging the values in ascending order and then grouping them into 100 equal parts (in terms of the number of values). The partition values are called **centiles** or **percentiles**. For example, a value for which 10% of the observations are less than it is known as the 10^{th} percentile. Thus, we can have the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, or any other percentile. It is then possible to state the range covered by any two of these percentiles such as the 10^{th} to 90^{th} or 25^{th} to 75^{th} percentile range. It may be noted that the median represents the 50^{th} percentile. If we estimate the range of the middle 50% of the observations about the median by using the 25^{th} and 75^{th} percentile values, we have the **interquartile range**. The interquartile range is another useful estimate of dispersion, especially for skewed distributions. If the dispersion in the data series is less, we can use the range defined by 10^{th} and 90^{th} percentile values to denote the dispersion.

A still better method of measuring variability about the central location is to estimate how closely the individual observations cluster about it. This leads to the mean square deviation or **variance**, which is calculated as the sum of the squares of the differences of individual

deviations from mean, divided by the number of observations. The squaring removes the effect of negative values. The **standard deviation** (SD) of a data series is simply the square root of the variance. Note that the variance is expressed in squared units, which is difficult to comprehend, but the standard deviation retains the original unit of observation. The standard deviation is particularly useful in normal distributions, because the proportion of values in the normal distribution (i.e., the area under the curve) is a constant for a given number of standard deviations above or below the mean of the distribution, as shown later.

The formulae for the calculation of variance (and standard deviation) of a population has the value 'n' as the denominator. However, the expression (n – 1) is used when calculating the variance (and standard deviation) of a sample. The quantity (n – 1) denotes the **degrees of freedom**, which is the number of independent observations or choices available. For instance, if a series of four numbers is to add up to 100, we can assign different values to the first three, but the value of the last is fixed by the first three choices and the condition imposed that the total must be 100. Thus, in this example, the degrees of freedom can be stated to be 3. The degrees of freedom is used when calculating the variance (and standard deviation) of a sample because the sample mean is an estimate of the predetermined population mean, and, in the sample, each observation is free to vary except the last one which must be a defined value.

MEASURES OF PRECISION

The **coefficient of variation** (CV) of a data series denotes the SD expressed as a percentage of the mean. Thus, it denotes the relative size of the SD with respect to the mean. The CV is calculated by dividing the SD by mean and multiplying by 100. An important source of variability in biological observations is measurement imprecision and CV is often used to quantify this imprecision. It is thus commonly used to describe variability of measuring instruments, and it is generally taken that a CV of less than 5% is acceptable reproducibility. CV can be conveniently used to compare variability between studies, as, unlike standard deviation, its magnitude is independent of the units employed.

Another measure of precision for a data series is the **standard error of the mean** (SEM), which is simply calculated as the SD divided by the square root of the number of observations. The SEM is primarily used to construct confidence intervals of population mean. Its use to depict dispersion of data in place of SD is erroneous. The standard error is a measure of precision and not dispersion. It is meant to provide an estimate of a population parameter from a sample statistic in terms of the confidence interval.

It is self-evident that when we observe a sample, and calculate the sample mean, this will not be identical to the population ('true') mean. However, if our sample is sufficiently large and representative of the population, and we have made our observations or measurements carefully, then the sample mean would be close to the true mean. If we keep taking repeated samples, and calculate a sample mean in each case, the distribution of these sample means would be expected to have less dispersion than that of all the individual observations in the samples. In fact, it can be shown that the sample means would have a symmetrical distribution, with the true population mean at its central location, and the standard deviation of this distribution would be nearly identical to the SEM calculated from individual samples. This is the essence of the **central limit theorem** in probability theory.

In general, we are not interested in drawing multiple samples, but rather would like to know how reliable our one sample is in describing the population. We use standard error to define a range in which the true population value is likely to lie, and this range is the **confidence interval**, with its two terminal values being called **confidence limits**. The width of the confidence interval depends on the standard error and the extent of confidence required. Conventionally, the 95% confidence interval (95% CI) is most commonly used. From the properties of a normal distribution curve (see below) it can be shown that the 95% CI of the mean would cover a range 1.96 standard errors on either side of the sample mean, and will have a 95% probability of including the population mean; while 99% CI will span 2.58 standard errors on either side of the sample mean and will have 99% probability of including the population mean. Thus, a fundamental relation that needs to be remembered is

$$95\%\ \text{CI of mean} = \text{Sample mean} \pm 1.96 \times \text{SEM}$$

It is evident that the confidence interval would be narrower if SEM is smaller. The larger the sample size, the smaller is the SEM. The confidence interval is correspondingly narrower and thus more 'focused' on the true mean. Large samples therefore increase precision.

Confidence intervals can be used to capture most population parameters from sample statistics like means, medians, proportions, correlation coefficients, regression coefficients, odds ratios, relative risks, and others. In all cases, the principles and the general pattern of estimating the confidence interval remain the same, that is

95% CI of a parameter = Sample statistic ± 1.96 X Standard error for that statistic

The formula for estimating standard error however varies for different statistics, and in some instances is quite elaborate. The situation therefore is usually managed by relying on computer software to do the calculations.

FREQUENCY DISTRIBUTIONS

It is useful to summarize a set of observations with a frequency distribution. The summary may be in the form of a table or a graph (plot). Many frequency distributions are encountered in medical literature and it is important to have a clear idea of the more commonly encountered ones.

Majority of distributions that quantitative clinical data follow are **unimodal**, that is the data has a single peak (mode) with a tail on either side. The more common of these unimodal distributions are symmetrical. However, some are **skewed** with a substantially longer tail on one side (**Figure 1**). The type of skew is determined by which side tail is longer. A **positively skewed** distribution has a longer tail on the right; with the majority of values being relatively low with a smaller number of extreme high values. A **negatively skewed** distribution has a longer tail to the left; with the extreme values being markedly low in comparison to the rest of the dataset. In this instance the mean, being unduly influenced by the extreme low values on the left, will be smaller than the median. On the other hand, in a **positively skewed** distribution the mean will be greater than the median because the mean is strongly influenced by the extreme values in the right-hand tail.

Manipulation of a dataset in order to alter its distribution is called **data transformation**. There are many different transformations, such as logarithmic, square root, reciprocal, logit transformation, and so on. There are certain advantages in working with symmetrical rather than asymmetrical data sets, and the most commonly used transformation to make positively skewed data symmetrical is the **logarithmic transformation**. In this, every value in the dataset is replaced by its logarithm. Logarithms are defined to a base, the most common being base *e* (natural logarithm) or base 10 (common logarithm). The end result is independent of the base chosen, provided the same base is used throughout. Notice that in log transformation, the differences in the transformed values are larger at the lower end of the scale. The logarithmic transformation stretches out the lower end and compresses the upper end of a distribution, with the result that positively skewed data will tend to become more symmetrical in shape. Calculations and statistical tests can be carried out on the transformed data before converting the results back to the original scale. A linear relationship between variables is desirable in regression analysis, and the logarithmic transformation is also useful in linearizing data, if an exponential relationship exists between two variables (**Figure 2**).

It is possible that Datasets may have more than one peak (mode). Such data can be difficult to manage and it may be the case that neither the mean nor the median is a representative measure of the central tendency. However, it is important to remember that **bimodal** or **multimodal** distributions are rare and may even be artifacts. A distribution with two peaks may actually be reflecting a combination of two unimodal distributions, for instance one for each gender or different age groups. In such cases, appropriate subdivision, categorization, or even recollection of the data may be required to eliminate multiple peaks.

THE NORMAL DISTRIBUTION

Many biological variables tend to cluster around a central value, with a symmetrical positive and negative dispersion about this point. The more extreme values become less frequent the further they lie from the central point. These features describe a normal distribution (**Figure 3**); the term 'normal' probably relating to the wide prevalence of this distribution. It is also referred to as a Gaussian distribution after the German mathematician, Karl Friedrich Gauss (1777–1855), although Gauss was not the first person to describe such a distribution. Some of the properties of a normal distribution are:

- Unimodal, bell-shaped distribution
- Symmetric about the mean

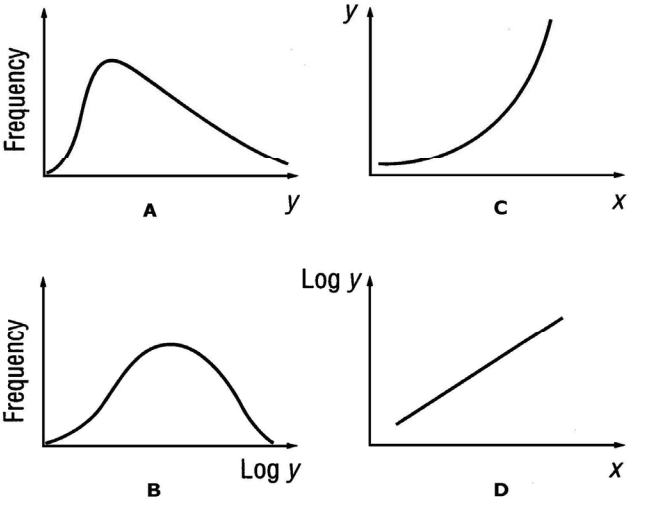

Figure 2: Two uses of logarithmic transformation of data. A & B: making positively skewed data normally distributed, and C & D: linearizing exponential relationship between two variables.

- Flattens symmetrically as the variance is increased
- Kurtosis is zero ('kurtosis' refers to how peaked a distribution is)

In a normal distribution curve, the mean, median and mode coincide. The area delimited by one standard deviation on two sides of the mean includes 68% of the total area under the curve, that for two standard deviations includes 95.4%, and for three standard deviations includes 99.7%; 95% of the values lie within 1.96 standard deviations on two sides of the mean. It is for this reason that the interval denoted by mean ± 1.96 X SD, is often taken as the normal range or **reference range** for many physiological variables. Anthropometric measurements (e.g., weight, height, waist circumference), biochemical evaluations (e.g., plasma glucose, liver function tests, urea, creatinine, serum electrolytes) and psychometric parameters (e.g., intelligence quotient scores) are common examples of variables that tend to follow normal distribution.

If we look at the formula for the normal distribution given below, it is evident that there are two parameters that define the curve, namely μ (the mean) and σ (the standard deviation):

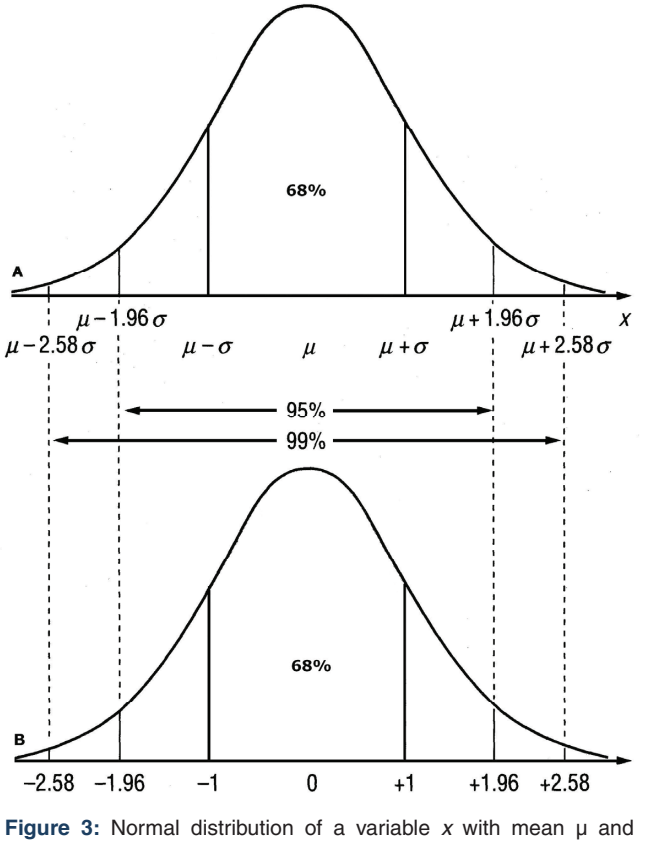

Figure 3: Normal distribution of a variable *x* with mean μ and standard deviation σ. The bottom panel shows *z*-transformation of *x* to derive the standard normal curve with mean 0 and standard deviation 1.

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-(x-\mu)^2/2\sigma^2},\ -\infty<x<\infty$$

The **standard normal distribution curve** is a particular normal distribution for which the probabilities that x will be in any interval have been extensively calculated. It is a symmetrical bell-shaped curve with a mean of 0 and a variance (or standard deviation) of 1. This is also known as the **z distribution**. The **standardized normal deviates** or **z values** or **z scores** of a random variable x can be calculated as follows:

$$z=\frac{x-\mu}{\sigma}$$

The z value tells us how many standard deviations the corresponding value of x lies above or below the mean of the normal distribution. Tables of z scores (in statistics books or generated using computer software) can be used to find out what proportion of any normal distribution lies above any given z score. We can also do the converse, that is, use z scores to find the score that divides the distribution into specified proportions. The z scores also allow us to determine the probability of a randomly picked element being above or below a particular score.

As the number of observations increase (say, n > 30), the shape of the distribution of sample means will approximate a normal distribution curve even if the distribution of the variable in question is not normal. This is explained by the central limit theorem, and is one reason why the normal distribution is so important in biomedical research.

Many statistical techniques require assumption of normality of the dataset. It is not mandatory for the sample data to be normally distributed, but it should represent a population that is normally distributed.

BINOMIAL DISTRIBUTION

A binomial distribution can exist if a population has a characteristic that belongs to one of two mutually exclusive categories. This distribution describes the probability of 'success' of an event in a fixed number of observations. It is frequently used to model the number or proportion of successes in a sample of size n drawn with replacement from a population of size N. The binomial variate satisfies the following properties:

- A Bernoulli (success – failure) experiment is performed n times
- The trials are independent
- The probability of success on each trial is a constant p; the probability of failure is q = 1–p
- The random variable X counts the number of successes in the n trials

The probability of an event in a binomial distribution can be calculated from the binomial distribution formula. We use the properties of the binomial distribution when drawing inferences about proportions of successes. The normal approximation to the binomial distribution is often used when analyzing proportions.

If the number of observations is very large, and the probability of a particular event is very small, then calculation of binomial probabilities can become quite tedious. An approximation, called Poisson distribution, can be used in such cases.

POISSON DISTRIBUTION

The Poisson distribution (named after Siméon Denis Poisson) is expressed by an exponential formula, that can be used to calculate the probability of 'rare' events that occur randomly but at a fixed 'mean' rate. Its assumptions are:

- Events occur randomly
- Events occur independent of one another
- Events occur at a uniform long-term rate

The mean equals the variance in a Poisson distribution. The distribution is skewed to the right when this mean is small, but approximates a normal distribution as mean becomes larger. Many interesting phenomena can be modeled by the Poisson distribution. Two frequently stated examples from biology are the number of suicides in a population and the number of mutations in a DNA strand.

PRESENTING DATA

Once summary measures of data have been calculated, they need to be presented in tables and graphs. Regarding data presentation in tables, it is helpful to remember the following:

- The mean is to be used for numerical data and for symmetric (non-skewed) distributions
- The median should be used for ordinal data or for numerical data if the distribution is skewed
- The mode is generally used only for examining bimodal or multimodal distributions
- The range may be used for numerical data to emphasize extreme values
- The standard deviation is to be used along with the mean
- Interquartile range or percentiles should be used along with the median
- Standard deviations and percentiles may also be used when the objective is to depict a set of norms ('normative data')
- The coefficient of variation may be used if the intent is to compare variability between datasets measured on different numerical scales
- 95% confidence intervals should be used whenever the intent is to draw inferences about populations from samples

For presenting data graphically, it is usually necessary to obtain the frequency distribution or relative frequency distribution (e.g., percentages) of the data. This can then be utilized to draw different types of graphs (or charts or plots or diagrams). Some useful graphs are as follows:

Pie chart: This depicts frequency distribution of categorical data in a circle (the 'pie'), with the sectors of the circle proportional in size to the frequencies in the respective categories. A particular category can be emphasized by pulling out that sector. All sectors are pulled out in an 'exploded' pie chart. Pie charts can be made highly attractive, by using color and three-dimensional design enhancements, but become cumbersome if there are too many categories.

Bar chart (also called **column chart**): This depicts categorical or discrete numerical data as a series of vertical or horizontal bars, with the bar heights being proportional to the frequencies. The separation between bars is of little significance other than to indicate that the bars denote discrete values or categories. Usually the separation distance is kept equal. Bars depicting subcategories can be stacked one on top of another (**compound, segmented or stacked bar chart**). Two or more data series can be depicted on the same bar chart by placing corresponding bars side by side – different patterns or colors are used to distinguish the different series (**clustered or multiple bar chart**). It is believed that the first bar chart appeared in the 1786 book, 'The Commercial and Political Atlas', by William Playfair.

Histogram: This is similar to a bar chart but is used for summarizing continuous numerical data and hence there should not be any gaps between the bars. The bar widths correspond to the class intervals. The alignment of the bars can be vertical or horizontal. A histogram is popularly used to depict the frequency distribution in a large data series. The class intervals should be so chosen that the bars are narrow enough to illustrate patterns in the data but not so narrow that they become too many in number. A histogram should be labeled carefully to clearly depict where the boundaries lie.

Dot plot: This depicts frequency distribution like histograms and can also be used for summarizing discrete numerical data. Instead of bars, it has a series of dots for each value or class interval – each dot

representing one observation. The alignment can be vertical or horizontal. Dot plots are conceptually simple but become cumbersome for large data sets. In the example given in **Figure 4**, note that the data is not showing a clear distribution.

Stem-and-leaf plot: This plot was introduced by the renowned statistician John Wilder Tukey in his 1970 book, *Exploratory Data Analysis*. It is sort of mixture of a diagram and a table and was devised to depict frequency distribution as well as individual data values for numerical data. The data values are examined to determine their last significant digit (the 'leaf' item) and this is 'attached' to the previous digits (the 'stem' item). The stem items are arranged in ascending or descending order vertically and a vertical line is usually drawn to separate the stem from the leaf. The number of leaf items should total up to the number of observations. An example of a stem-and-leaf plot is provided in **Figure 5**. The figures to the left of the vertical line constitute the stem, while those to the right comprise the leaf. The number of digits in the leaf equals the number of observations in the data set. Note that the plot gives an idea of the underlying distribution while retaining all the individual values. However, it becomes cumbersome with large data sets.

Box-and-whiskers plot (or **box plot**): This was also introduced by Tukey in his 1970 book *Exploratory Data Analysis*. This is a graphical representation of numerical data based on the five-figure summary– minimum, 25th percentile, median (50th percentile), 75th percentile and maximum values. A rectangle is drawn extending from the lower quartile to the upper quartile, with the median dividing this 'box' but not necessarily equally. Lines ('whiskers') are drawn from the ends of the box to the extreme values. Outliers may be indicated beyond the extreme values by dots or asterisks – in such 'refined' box plots, the whiskers have lengths not exceeding 1.5 times the interquartile range. The whole plot may be aligned vertically or horizontally. Box plots are ideal for summarizing large samples and are being increasingly used. Multiple box plots, arranged side by side, allow ready comparison of data sets. A horizontal box plot depicting the five number summary of numerical data is shown in **Figure 6**. Note that this particular dataset is not symmetrical but is skewed to the left.

Finally, an increasingly important plot these days is the **forest plot**. This is intended to illustrate the strength of treatment effects in multiple studies addressing the same outcome or multiple subgroups within the same study. Most commonly, a forest plot is used to summarize the results of a meta-analysis in terms of odds ratios (or relative risks) and their confidence intervals. As shown in **Figure 7**, each study is represented by a small box (or circle), the size of which is proportional to the weight given to the study, which is usually dependent on its sample size. The mid-point of this box is positioned at the value of its treatment effect (in terms of the point estimate i.e., the odds ratio or the relative risk) with horizontal lines extending on both sides to the 95% confidence limits. A diamond placed at the bottom with its mid-point being positioned at the pooled point estimate depicts the overall result. The width of the diamond indicates the 95% CI for the pooled results. There is also a vertical line corresponding to the position of no treatment effect. The term 'forest' probably relates

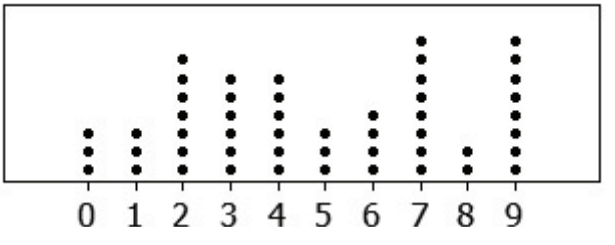

Figure 4: A simple dot plot depicting a series of discrete scores.

100	002266
110	0002224468
120	000002222244444444666666688888
130	000222224444446666668888
140	00000222224444446666
150	0022244668

Key: 120 | 4 means 124 mmHg

Figure 5: A stem-and-leaf plot depicting systolic blood pressure recordings (recorded as even values only) in 100 individuals.

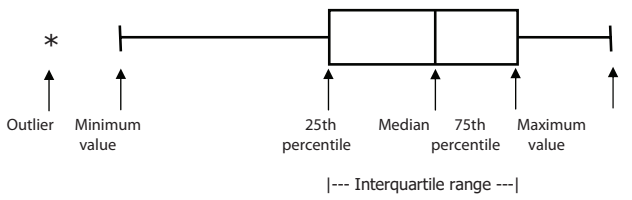

Figure 6: A horizontal box plot depicting the five number summary of numerical data. Note that this particular dataset is not symmetrical but is skewed to the left.

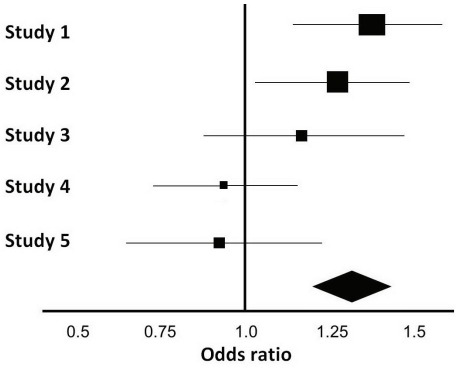

Figure 7: The typical appearance of a forest plot.

to the collection of lines in the plot. It is customary to indicate the names of the studies in a column to the left hand side of the plot with additional summary data, if considered necessary.

We have looked at the commonly used plots used for summarizing data and depicting underlying patterns. Many other plots are used in biostatistics for depicting data distributions, time trends in observations, relationships between two or more variables, exploring goodness-of-fit to hypothesized data distributions and drawing inferences by comparing data sets.

FURTHER READING

1. Sprent P. Statistics in medical research. Swiss Med Wkly 2003; 133: 522–9.
2. Chan YH. Data presentation. Singapore Med J 2003; 44: 280–5.
3. Dawson B, Trapp RG. Basic & clinical biostatistics. 4th ed. New York: McGraw-Hill; 2004.
4. MacDonald TH. Basic concepts in statistics and epidemiology. Oxford: Radcliffe Publishing; 2007.
5. Samuels ML, Witmer JA, Schaffner AA. Freund JE. Statistics for the life sciences. 4th ed. Boston: Prentice Hall (Pearson Education); 2012.