The Nature of Data

18

Clair Smith

18.1 List of Definitions

Data: A collection of data points organized into one or more variables of interest.

Example: The set of all responses to a survey given to a group of people, all measurements taken from the mice in an animal study, etc.

Variable: A measureable characteristic such as blood pressure, age, or gender.

Example: Treatment group, marital status, diabetes status, systolic blood pressure, blood glucose level, etc.

Observation: A single datum. In clinical data this will often be a measurement taken from a person or animal. Also called a data element or data point.

Example: The heart rate of mouse in an animal study, the cancer status of a cell from a person in a cancer study, the range of motion of a knee from a cadaver in a meniscectomy study, and the BMI of one person in a study.

Statistic: A numerical summary of the data points that make up a variable. This can be calculated from a sample.

Example: Mean, variance, median, minimum, and maximum.

Sample: Data that is collected/observed. A small subset of the population of interest is presented.

Example: A random sample of residents of a certain neighborhood and all people hospitalized for a heart attack at one of three local hospitals during a certain period of time.

Population: The group of all subjects researchers are interested in studying.

Example: The set of all women currently living in the USA, the set of all people experiencing lower back pain in the USA, and the set of all mice of a certain species.

18.2 Types of Data

There are two major types of data that researchers typically deal with in health science: continuous and discrete data. The type of data drives which statistics are used in their analyses. Continuous data such as age, height, weight, and BMI have infinitely many possible values. For example, age in years can be any positive real number such as 42 or 37.25. Discrete (also known as categorical) data has a limited number of values it can take on such as race, treatment group, and study site. For example, if possible values of race on a self-reported survey are black, white, and other, then everyone taking the survey will have one of these three values for the variable

C. Smith (✉)
Department of Orthopaedic Surgery,
University of Pittsburgh, Pittsburgh, PA, USA

Department of Physical Therapy,
University of Pittsburgh, Pittsburgh, PA, USA

Bridgeside Point 1, Pittsburgh, PA, USA
e-mail: cns45@pitt.edu

V. Musahl et al. (eds.), *Basic Methods Handbook for Clinical Orthopaedic Research*,
https://doi.org/10.1007/978-3-662-58254-1_18

race. Other less common types of data are count data and censored data. Count data is made up of whole numbers that represent counts such as the number of falls in a year of follow-up or the number of heartbeats per minute. While there are analyses created specifically for count data, it is often treated as continuous data for simplicity. Censored data, such as number of years till death after having a certain procedure, occurs when the event researchers are interested in (such as death) may not occur during the study. Another type of data, longitudinal data, occurs when measurements are taken repeatedly from subjects over a period of time.

A continuous variable is one whose values can be any real number. It is meaningful to measure the distance between values, and arithmetic operations such as addition and multiplication make sense for continuous variables but not for discrete variables. Technically, all continuous variables are measured discretely since we don't have instruments that can be measured continuously. One can think of continuous data as discrete with lots of levels or categories. For example, blood pressure is typically measured to 2 mmHg because measuring with higher precision would be difficult with the instruments that are used. However, we still treat blood pressure as continuous since there would be far too many levels to treat it as discrete. Sometimes discrete variables with many levels such as the visual analog scale (VAS) for measuring pain or variables measured with the Likert scale are treated as continuous for ease of analysis. When treating discrete variables as continuous, you are assuming that each level of the variable is equidistant apart. Another example of a continuous variable is the Lysholm scale for assessing ACL injuries which gives a score from 0 to 100 with higher scores indicating fewer symptoms.

There are two major types of discrete data: nominal and ordinal. Nominal data has no inherent ordering such as gender, race, and marital status. Ordinal data can be ordered from low to high such as injury severity, level of education, and household income. The differences between the levels of ordinal variables are not necessarily equal. For example, the difference between the mild and moderate level of an injury severity variable may not be the same as the difference between the moderate and severe level. For both types of discrete data, each observation must belong to exactly one level of the discrete variable, and the levels should cover all possible values that exist in the data set. For example, if the discrete variable race has levels black, white, and other, then each observation in the data set must be categorized as black, white, or other. If a discrete variable has only two levels, then it is called a dichotomous variable. Examples include gender and disease status (the disease is either present or not present).

18.3 Data Description

Summarizing discrete data is simpler than summarizing continuous data. Discrete data is often described by reporting the frequency and proportion (or percent) of people belonging to each level of the discrete variable. For example, say you were reporting on disease severity in a study of 50 people. Your description of the variable "disease severity" could be 23 (46%) mild, 12 (24%) moderate, and 15 (30%) severe if 23 people in the study had mild disease, 12 had moderate, and 15 had severe. If the variable is dichotomous, then it is acceptable to report only the frequency and proportion in one level of that variable. For example, if researchers were summarizing the dichotomous variable "gender" and putting it in a table of demographic information for a study, then they could simply report the frequency and proportion of women in the study sample. Researchers would not need to include the number of men in the study since this can be deduced by subtracting the number of women in the study from the sample size. Some researchers depict the proportions of subjects in each level of a discrete variable in a bar graph. This may be appropriate if the publication has no other figures. However, when other figures are present, graphing proportions is superfluous as they are described adequately by frequencies and proportions alone.

The relationship between two categorical variables is best captured by a 2 × 2 table. In such a table, the rows are levels of one categorical variable, and the columns are levels of the other categorical variable. The cells of the table contain the number of people in the study belonging to the corresponding levels of the row and column variables. The last row and last column are typically reserved for totals (also known as margins).

Continuous data contains more information, has more properties, and requires more statistics to describe it than discrete data. There are different statistics to measure the location (or center), spread (or dispersion), and shape of the distribution of values from a continuous variable.

Measures of location seek to describe the central tendency of the data with a representative value from it. Examples are the mean (or average), median, and mode. The mean of a continuous variable is the sum of all the values divided by the number of values present in that sum. Put into symbols the mean is $\frac{\sum_{i=1}^{n} x_i}{n}$, where x_i represents the ith value, n is the sample size, and $\sum_{i=1}^{n} x_i$ says to sum all the values from 1 to n. The letter i is called an index. The median is the middle value of the ordered data. If the values are ordered from smallest to largest (or largest to smallest), then the median is the value that has an equal number of observations on either side of it. If the sample size is even, then the median is found by averaging the two middle values of the ordered data. Instead of listing out the ordered values by hand and visually finding the middle value, there is a simple formula that can be used to find the position of the median. Say there are n people in the study and the age of each person is listed from smallest to largest. If n is odd, then the position of the median age is $\frac{n+1}{2}$. Note that this equation will produce the *position* of the median, not the value of the median itself. If n is even, then the median is the average of the numbers in the $\frac{n}{2}$ and $\frac{n}{2}+1$ positions of the ordered list of values [1]. The median and mean can only be used to describe continuous data. The mode of a variable is the most frequently occurring value and can be used to describe the central tendency of continuous or discrete data.

Another name for the median is the second quartile. The quartiles split the list of ordered values into fourths. The first quartile is also called the 25th percentile and is often denoted Q_1. If the data is listed in order, then 25% of the values will be below or equal to the first quartile. The median is the 50th percentile (or second quartile) since half of the values are less than or equal to it and it is denoted Q_2. The third quartile, Q_3, is also called the 75th percentile, and 75% of the values are less than or equal to it.

Measures of spread describe how tightly clustered the values are around the mean of continuous data. Examples are the variance, standard deviation, range, and interquartile range. The variance is the average squared distance from the mean. It is calculated by adding up all of the squared differences between the mean and each data point then dividing this sum by the number of data points minus one. If n is the sample size and $\bar{x}$ is the mean of the sample, then the following is an equation for finding the variance of the sample: $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$. Note that Σ is a symbol meaning "the sum of" and x_i represents the ith value in the sample. When put together as in $\sum_{i=1}^{n}(x_i-\bar{x})^2$, this means take each value from the first (i = 1) to the last (i = n), subtract the mean from it, then square it, then add all these squares together. This equation is similar to the equation for the mean except that it is divided by $n - 1$ instead of n. Dividing by $n - 1$ leads to a less biased statistic than dividing by n. The standard deviation is the square root of the variance and

thus the average distance from the mean. The range is simply the maximum (largest) value minus the minimum (smallest) value. The interquartile range (IQR) is the third quartile minus the first quartile. The IQR describes the spread of the central 50% of the data. For all measures of spread, a higher value indicates that observations are more spread around the mean and smaller values indicate they are more tightly clustered about the mean.

Fact Box 18.1

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n.}$$

$$\text{Position of median for odd } n = \frac{n+1}{2}$$

$$\text{Position of median for even } n = \frac{\frac{n}{2} + \left(\frac{n}{2}+1\right)}{2}$$

Measures of the shape of a distribution describe the overall trend of the data. As an example, a distribution could have mostly large values with a few extreme outliers, or it could have values evenly distributed across the range. Some examples of distribution shapes are symmetric, normal, bimodal, and left or right skewed. The mean, median, and mode of a continuous variable are equal if the distribution of its values is symmetric. In terms of symmetric data, the relative position of observations is the same on either side of the median. Right skewed (or positively skewed) data occurs when observations above the median are farther in absolute value than observations below the median. Another way of saying this is that the distribution has a long tail to the right. In right skewed data, the majority of the values are relatively small and close together, and a minority of the values are extreme or much larger in value than the rest. Left skewed (or negatively skewed) data occurs when observations below the median are farther in absolute value than observations above the median. Left skewed data has a long tail to the left since most of the values are large and there are a few extreme observations that are much smaller than the rest. The mean is greater than the median in right skewed data and less than the median in left skewed data. There is a skewness index that measures the degree of skewness in the data. The index is zero if the data is symmetric, greater than zero if the data is right skewed, and less than zero if the data is left skewed [3]. A *normal distribution* is a symmetrical hill or bell shape with the majority of the values close to the central value (the mean) and a few extreme observations on either side of the mean (i.e., in the tails of the distribution). A bimodal distribution looks like the two humps of a camel; it has two central values. When the data is symmetric, the best numerical summaries are the mean and standard deviation. When the data is skewed, it is best to use the median and interquartile range or interquartile deviation (half of the interquartile range).

Fact Box 18.2

$$\text{Variance} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

$$\text{Standard deviation} = \sqrt{variance}$$

$$\text{IQR} = Q_3 - Q_1$$

Kurtosis is another measure of shape that describes how flat or steep the distribution of values is compared to a bell-shaped (or *normal*) distribution. If there are a lot more observations in the tails of a distribution compared with a normal distribution, then the graph appears flatter than a bell shape. If there are many fewer observations in the tails of a distribution compared with a normal distribution, then the graph appears more peaked than a bell shape [2].

All of the descriptive measures discussed thus far are statistics. Statistics are calculated from a sample that is drawn from the population of interest. Suppose the goal of a study is to determine whether a new surgical technique for repairing a joint leads to a better clinical outcome than the standard procedure. The population of interest in this case would be the set of all people with the joint injury who would be eligible for this surgery. In order to determine whether the new technique is an improvement over the old technique, researchers must look at the outcomes from a sample of people with the joint injury. It is not possible to observe all people with this injury (the population of interest), so a sample must be taken. Typically, the sample is chosen in such a way that every member of the population has an equal opportunity of being picked for the sample. A sample that is created in this way is called a random sample because each member is chosen at random. This helps to ensure that the sample is representative of the population, e.g., if half of the population is women, then roughly half of the random sample drawn from the population should be women. The researchers would then use statistics such as means and standard deviations calculated from this sample to summarize outcomes of the two surgery techniques. Such an outcome may be the range of motion of the repaired joint after it has healed. Since range of motion is a continuous measure, researchers would use a mean or median and standard deviation or interquartile range to summarize it. This example illustrates using a sample to make inference about a population, the main goal of statistics.

Since a sample does not include all members of a population, there are multiple ways to draw a sample from a given population. The number of people in the population of interest is typically denoted by N, and the number of people in a given sample drawn from the population is denoted by n. Figure 18.1 depicts three different samples of size n drawn from a population of size N.

Suppose the three samples in the figure were drawn in a sequence: n individuals were selected at random from the population to form the first sample, measurements were taken on them, and then they were returned to the population pool. This way of sampling is called *sampling with replacement*. This process was then repeated for the second and third samples of size n. If we calculated the mean of the measurements taken on each of the three samples, we would get three different means even though the samples are the same size and are taken from the same population. This occurs because the three samples consist of different individuals. It is possible that there is overlap between the three samples, that is, some individuals may occur in two or more of them. This is possible because the samples were drawn from the population with replacement: they were selected, their measurements taken, and then they were returned to the population pool. Thus, every time a sample is drawn from the population, a different sample mean is calculated, but each of these means will be a good estimate of the true mean of the entire population (given that the sample size, n, is sufficiently large). The mean calculated from the sample of size n is an example of a sample statistic, and the mean calculated from the population of size N is an example of a population parameter. Sample statistics are estimates of population parameters. Population parameters are usually unknown since we cannot measure an entire population but we estimate these parameters by taking a random sample of the population and calculating sample statistics. The larger the sample size, the more confident researchers are that the sample statistics are good approximations of the population parameters.

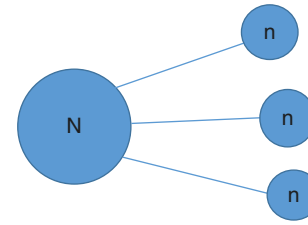

Fig. 18.1 Samples of size n drawn from a population of size N

18.4 Visual Displays

It is good practice to plot the data before summarizing and performing statistical tests on it. This will give the researcher a sense of the type of data available. There are various methods for describing and analyzing data, and which method to be used depends on the nature of the data.

A stem-and-leaf plot is a simple visual display of data points that shows the distribution or shape of the values of a continuous variable. An advantage of this plot is that it includes the value of each individual observation. This plot is appropriate when there are a small number of observations. As an example, suppose there is a small sample of 15 subject's BMI measurements that have been rounded to the nearest integer. BMI is a continuous variable and its units are kg/m^2. The first step of making a stem-and-leaf plot of these values would be to list them in order:

$$18, 19, 23, 24, 24, 24, 25, 25, 26, 26, 27, 28, 30, 32, 37$$

The stem of the plot is made up of the leading number, and the leaves are made up of the trailing number. Both the numbers in the stem and in the leaves are ordered smallest to largest.

1 | 89
2 | 3444556678
3 | 027

It can be seen from this plot that the BMI measurements are distributed in roughly a hill shape: most of the values are in the middle and there are a few in either tail. If there are more observations, more than one line can be added for each digit in the stem.

A histogram is a graph that shows the shape of the distribution of values of a continuous variable. The horizontal (or x) axis has the values of the variable, and the vertical (or y) axis has the frequency or proportion of observations. The height of each rectangle represents the proportion or frequency of observations whose values fall within the range specified by the width of the rectangle. If the distribution of values of a variable is symmetric, then cutting the histogram along the median will result in each half being a mirror image of the other. A common example of a symmetric distribution is a hill or bell-shaped distribution. Figure 18.2 shows a histogram for the 15 BMI measurements in the last example.

The width of the rectangles in this histogram is five observations, and the vertical axis is the frequency of occurrences. The x-axis shows the range of values for each rectangle. The histogram shows the general *hill-shaped* trend of the data: most of the data (9 observations) fall within the range of 23–28 kg/m^2. This histogram shows a similar shape as the stem-and-leaf plot turned on its side. If the width of the rectangles is too large, important information about the shape of the distribution can be lost. The smaller the width of the rectangles, the more detail about the shape of the distribution will be shown. Most statistical programs will automatically choose a width that is appropriate for the data.

Another visual display for continuous data is the box plot. The box plot shows the interquartile range, the median, and any extreme observations (i.e., observations that have values that are much larger or much smaller than the rest of the data). If there is a lot of variability in the data, then the box and whiskers will be elongated. If there is not a lot of variability, then the box and whiskers will appear squatter. Figure 18.3 shows a box plot of the BMI example data.

The first quartile of the BMI data is the bottom line of the box, the median is the middle line in the box, and the third quartile is the top line of the box. When the third quartile is farther from the median than the first quartile, the data is *right skewed*, and

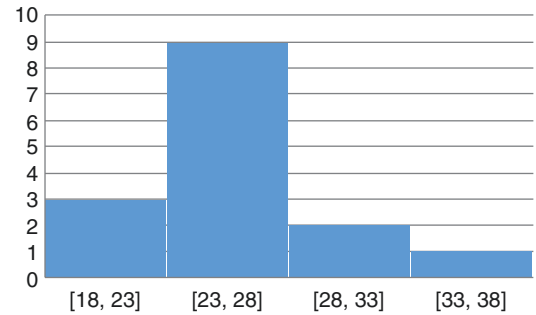

Fig. 18.2 Histogram of a sample of 15 BMI measurements

when the first quartile is farther from the median, the data is *left skewed*. The histogram and box plot of the BMI data in Fig. 18.3 show a slight right skew in the shape of the distribution. The whiskers of the box plot are drawn to the smallest and largest observations in the sample that are not outliers. Outliers are defined as values that are greater than Q_3 + 1.5*(IQR) or less than Q_1 – 1.5*(IQR). The dots in the box plot are extreme outliers, which are defined to be larger than Q_3 + 3*(IQR) or smaller than Q_1 – 3*(IQR) [3].

The box plot is a good visual display to use when comparing a continuous variable between different groups of a categorical variable since they can be plotted side-by-side on the same set of axes. This allows direct comparison of the distributions of the continuous variable across various levels of the categorical variable. As an example, consider the BMI data again, but suppose there is information on whether the subjects were over 25 years old or under 25 years old. Figure 18.4 is an example of a way to visualize the relationship between a continuous variable (BMI) and a categorical variable (age category).

From Fig. 18.4 it can be seen that subjects who are over 25 years old have a higher BMI than people who are 25 years old or younger.

A scatter plot is useful for understanding the relationship between two continuous variables and revealing potential outliers. Values of one variable are plotted on the horizontal axis, and values of the other variable are on the vertical axis. A scatter plot is a quick way to discover potential trends in the data. For example, if higher values of one variable tend to occur with higher values of the other, then the scatter plot will show this positive relationship. If there is no relationship between the two variables, then the scatter plot will show a random scattering of points that don't indicate any specific pattern. If most of the points are clustered tightly together, while one or two points are clearly outside of this cluster, then these points are potential outliers and should be checked for accuracy. Figure 18.5 is an example of a scatter plot using the BMI data. A second variable, age, has been added to the vertical axis.

Figure 18.5 shows that as age increases so does BMI. In other words, there is a positive relationship between age and BMI. An example of an outlier for this data is the point (32, 20), i.e., the point with BMI = 32 and age = 20. While

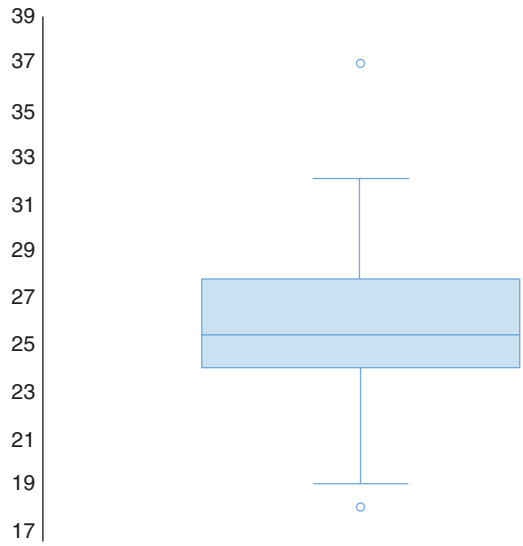

Fig. 18.3 Box plot of 15 BMI measurements

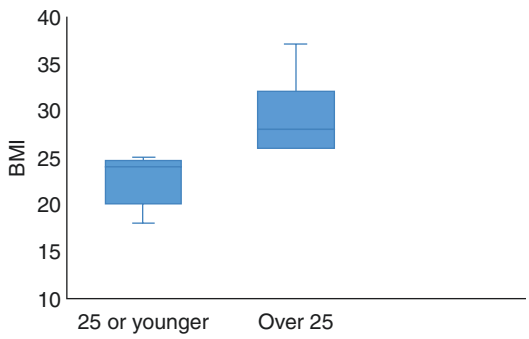

Fig. 18.4 Side-by-side box plots of 15 BMI measurements

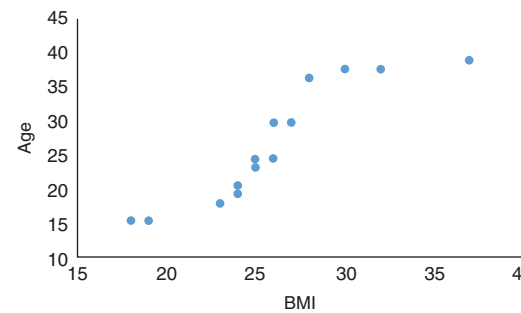

Fig. 18.5 Scatter plot of 15 subjects' age and BMI

the point is medically feasible, if it were in the plot in Fig. 18.5, we would want to check on its accuracy, because it is so far away from the increasing trend of the rest of the points. Two continuous variables could also have a negative relationship if it were the case that as one increased the other decreased. If the scatter plot appears to show a positive relationship for some values and a negative relationship for others, we would say the relationship appears to change direction. A statistic called a correlation coefficient classifies the strength of the association between two continuous variables.

18.5 Conclusion

The first steps of data analysis should be to determine what types of variables are present and to describe them with appropriate summary statistics and visual displays. Different types of data have different properties, and these properties determine which statistical tests are appropriate for answering the questions of interest. Statistical tests are based on probability theory and allow the researchers to draw conclusions about a population based on a sample from that particular population. This is called statistical inference and is the overarching goal of statistical analysis.

References

1. D'Agostino RB, Sullivan LM, Beiser AS. Introductory applied biostatistics. Belmont: Thomson Brooks/Cole; 2006.
2. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. 10th ed. Hoboken: Wiley; 2013.
3. Rosner B. Fundamentals of biostatistics. 8th ed. Boston: Cengage Learning; 2016.