Descriptive statistics are used to describe the essential features of the data in a study. It provides simple summaries about the sample and the measures. Together with simple graphics analysis, it forms the basis of virtually every quantitative analysis of data. Descriptive statistics allows presenting quantitative descriptions in a convenient way. In a research study, it may have lots of measures. Or it may measure a significant number of people on any measure. Descriptive statistics helps to simplify large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary.

Frequency Distributions

Frequency distributions are visual displays that organize and present frequency counts (n) so that the information can be interpreted more easily. Along with the frequency counts, it may include relative frequency, cumulative frequency, and cumulative relative frequencies.

- The *frequency* (n) is the number of times a particular variable assumes that value.
- The *cumulative frequency* (N) is the number of times a variable takes on a value less than or equal to this value.
- The *relative frequency* (f) is the percentage of the frequency.
- The *cumulative relative frequency* (F) is the percentage of the cumulative frequency.

Depending on the variable (categorical, discrete or continuous), various frequency tables can be created.

Example 1: favorite color of 10 individuals – categorical variable

List of responses:

Blue	Red	Blue	White	Green
White	Blue	Red	Blue	Black

Frequency Distribution:

Color	**n**	**N**	**f**	**F**
Blue	4	4	0.4	0.4
Red	2	6	0.2	0.6
White	2	8	0.2	0.8
Green	1	9	0.1	0.9
Black	1	10	0.1	1.0
Total	10		1	

Example 2: age of 20 individuals – discrete numerical variable

List of responses:

20	22	21	24	21	20	20	24	22	20
22	24	21	25	20	23	22	23	21	20

Frequency distribution:

Age	**n**	**N**	**f**	**F**
20	6	6	0.3	0.3
21	4	10	0.2	0.5
22	4	14	0.2	0.7
23	2	16	0.1	0.8

24	3	19	0.15	0.95
25	1	20	0.05	1
Total	20		1	

Example 3: height of 20 individuals – continuous numerical variable

List of responses:

1.58	1.56	1.77	1.59	1.63	1.58	1.82	1.69	1.76	1.60
1.73	1.51	1.54	1.61	1.67	1.72	1.75	1.55	1.68	1.65

Frequency distribution:

Interval	**n**	**N**	**f**	**F**
]1.50, 1.55]	3	3	0.15	0.15
]1.55, 1.60]	5	8	0.25	0.4
]1.60, 1.65]	3	11	0.15	0.55
]1.65, 1.70]	3	14	0.15	0.7
]1.70, 1.75]	3	17	0.15	0.85
]1.75, 1.80]	2	19	0.1	0.95
]1.80, 1.85]	1	20	0.05	1
Total	20		1	

Measures of Central Tendency and Measures of Variability

A measure of central tendency is a numerical value that describes a data set, by attempting to provide a "central" or "typical" value of the data (McCune, 2010). As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.
Measures of central tendency should have the same units as those of the data values from which they are determined. If no units are specified for the data values, no units are specified for the measures of central tendency.
The mean (often called the average) is most likely the measure of central tendency that the reader is most familiar with, but there are others, such as the median, the mode, percentiles, and quartiles.
The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

A measure of variability is a value that describes the spread or dispersion of a data set to its central value (McCune, 2010). If the values of measures of variability are high, it signifies that scores or values in the data set are widely spread out and not tightly centered on the mean. There are three common measures of variability: the range, standard deviation, and variance.

Mean

The mean (or average) is the most popular and well-known measure of central tendency. It can be used with both discrete and continuous data. An important property of the mean is that it includes every value in the data set as part of the calculation. The mean is equal to the sum of all the values of the variable divided by the number of values in the data set. So, if we have *n* values in a data set and $(x_1, x_2, \dots, x_n)$

are values of the variable, the sample mean, usually denoted by $\bar{x}$ (denoted by μ, for population mean), is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Applying this formula to example 2 above, the mean is given by:

$$\bar{x} = \frac{20 * 6 + 21 * 4 + 22 * 4 + 23 * 2 + 24 * 3 + 25 * 1}{20} = \frac{435}{20} = 21.75$$

So, the age mean for the 20 individuals is around 22 years (approximately).

Median

The median is the middle value or the arithmetic average of the two middle values of the variable that has been arranged in order of magnitude. So, 50% of the observations are greater or equal to the median, and 50% are less or equal to the median. It should be used with ordinal data. The median (after ordering all values) is as follows:

$$\tilde{x} = \begin{cases} \dfrac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}, \text{if } n \text{ is even} \\ x_{\frac{n+1}{2}}, \quad \text{if } n \text{ is odd} \end{cases}$$

In example 2 above, by ordering the age variable values, we have:

20, 20, 20, 20, 20, 20, 21, 21, 21, **21, 22**, 22, 22, 22, 23, 23, 24, 24, 24, 25

As *n* is even, the median is the average of the middle values. So $\tilde{x} = \frac{21+22}{2} = 21.5$ is the age median for the sample with 20 individuals.

Mode

The mode is the most common value (or values) of the variable. A variable in which each data value occurs the same number of times has **no mode**. If only one value occurs with the greatest frequency, the variable is **unimodal**; that is, it has one mode. If exactly two values occur with the same frequency, and that is higher than the others, the variable is **bimodal**; that is, it has two modes. If more than two data values occur with the same frequency, and that is greater than the others, the variable is **multimodal**; that is, it has more than two modes (McCune, 2010). The mode should be used only with discrete variables.

In example 2 above, the most frequent value of age variable is "20". It occurs six times. So, "20" is the mode of the age variable.

Percentiles and Quartiles

The most common way to report relative standing of a number within a data set is by using percentiles (Rumsey, 2010). The P_{th} percentile cuts the data set in two so that approximately P% of the data is below it and (100−P)% of the data is above it. So, the percentile of order *p* is calculated by (Marôco, 2011):

$$P_p = \begin{cases} X_{int(i+1)} & \text{if } i = \dfrac{np}{100} \text{ is not integer} \\ \dfrac{X_i + X_{i+1}}{2} & \text{if } i = \dfrac{np}{100} \text{ is integer} \end{cases}$$

where n is the sample size and $int(i+1)$ is the integer part of $i+1$.
It is usual to calculate the P_{25} also called first quartile (Q_1), P_{50} as second quartile (Q_2) or median and P_{75} as the third quartile (Q_3).

In example 2 above, we have:

20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 22, 22, 22, 22, 23, 23, 24, 24, 24, 25

Thus,

- 25th percentile (P_{25}) or 1st quartile (Q_1): as $i = \frac{20*25}{100} = \frac{500}{100} = 5$ is integer,

$$P_{25} = Q_1 = \frac{X_5 + X_6}{2} = \frac{20+20}{2} = 20$$

- 50th percentile (P_{50}) or median: as $i = \frac{20*50}{100} = \frac{1000}{100} = 10$ is integer,

$$P_{50} = Q_2 = \tilde{x} = \frac{X_{10} + X_{11}}{2} = \frac{21+22}{2} = 21.5$$

- 75th percentile (P_{75}) or 3rd quartile (Q_3): as $i = \frac{20*75}{100} = \frac{1500}{100} = 15$ is integer,

$$P_{25} = Q_3 = \frac{X_{15} + X_{16}}{2} = \frac{23+23}{2} = 23$$

Range

The range for a data set is the difference between the maximum value (greatest value) and the minimum value (lowest value) in the data set; that is

$$range = maximum\ value - minimum\ value$$

The range should have the same units as those of the data values from which it is computed.
The interquartile range (IQR) is the difference between the first and third quartiles; that is, $IQR = Q_3 - Q_1$ (McCune, 2010).

In example 2 above, minimum value=20, maximum value=25. Thus, the range is given by 25-20=5.

Standard Deviation and Variance

The variance and standard deviation are widely used measures of variability. They provide a measure of the variability of a variable. It measures the offset from the mean of a variable. If there is no variability in a variable, each data value equals the mean, so both the variance and standard deviation for the variable are zero. The greater the distance of the variable' values from the mean, the greater is its variance and standard deviation.
The relationship between the variance and standard deviation measures is quite simple. The standard deviation (denoted by σ for population standard deviation and s for sample standard deviation) is the square root of the variance (denoted by σ^2 for population variance and s^2 for sample variance).
The formulas for variance and standard deviation (for population and sample, respectively) are:

- Population variance: $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$, where x_i is the i^{th} data value from the population, μ is mean of the population, and N is the size of the population
- Sample variance: $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$, where x_i is the i^{th} data value from the sample, $\bar{x}$ is mean of the sample and n is the size of the sample
- Population standard deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$
- Sample standard deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$

Charts and Graphs

Data can be summarized in a visual way using charts and/or graphs. These are displays that are organized to give a big picture of the data in a flash and to zoom in on a particular result that was found. Depending on the data type, the graphs include pie charts, bar charts, time charts, histograms or boxplots.

Pie Charts

A pie chart (or a circle chart) is a circular graphic. Each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. The sum of all slices of the pie should be 100% or close to it (with a bit of round-off error). The pie chart is used with categorical variables or discrete numerical variables.
Figure 1 represents the example 1 above.

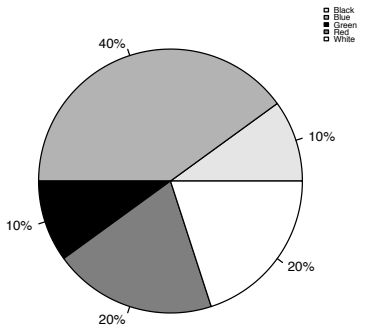

Figure 1 Pie chart example

Bar Charts

A bar chart (or bar graph) is a chart that presents grouped data with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column bar chart. In general, the x-axis represents categorical variables or discrete numerical variables.
Figure 2 and Figure 3 represent the example 1 above.

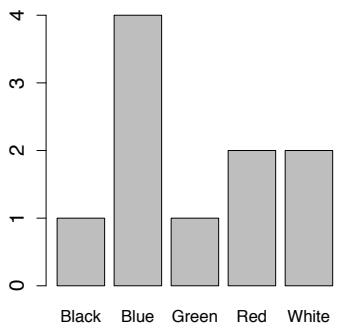

Figure 2 Bar graph example (with frequencies)

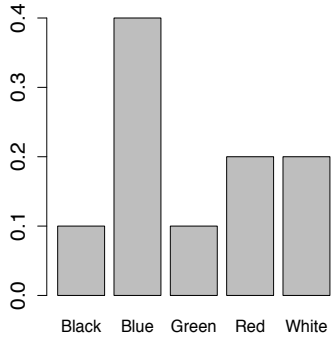

Figure 3 Bar graph example (with relative frequencies)

Time Charts

A time chart is a data display whose main point is to examine trends over time. Another name for a time chart is a line graph. Typically a time chart has some unit of time on the horizontal axis (year, day, month, and so on) and a measured quantity on the vertical axis (average household income, birth rate, total sales, or others). At each time's period, the amount is shown as a dot, and the dots are connected to form the time chart (Rumsey, 2010).

Figure 4 is an example of a time chart. It represents the number of accidents, for instance, in a small city along some years.

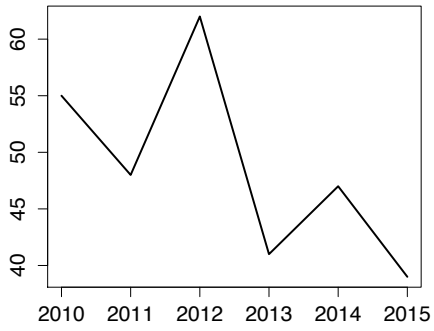

Figure 4 Time Chart example

Histogram

A histogram is a graphical representation of numerical data distribution. It is an estimate of the probability distribution of a continuous quantitative variable. Because the data is numerical, the categories are ordered from smallest to largest (as opposed to categorical data, such as gender, which has no inherent order to it). To be sure each number falls into exactly one group, the bars on a histogram touch each other but don't overlap (Rumsey, 2010). The height of a bar in a histogram may represent either frequency or a percentage (Peers, 2006).

Figure 5 accounts for the histogram of example 3 above.

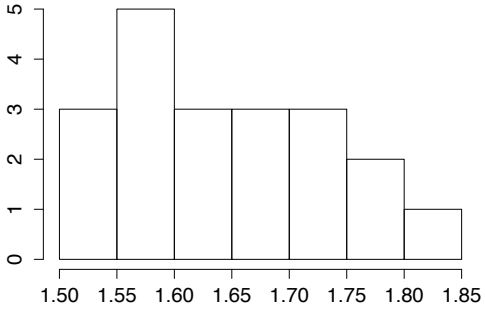

Figure 5 Histogram example

Boxplot

A boxplot or box plot is a convenient way of graphically depicting groups of numerical data. It is a one-dimensional graph of numerical data based on the five-number summary, which includes the minimum value, the 25th percentile (also known as Q_1), the median, the 75th percentile (Q_3), and the maximum value. In essence, these five descriptive statistics divide the data set into four equal parts (Rumsey, 2010).

Some statistical software adds asterisk signs (∗) or circle signs (o) to show numbers in the data set that are considered to be, respectively, outliers or suspected outliers — numbers determined to be far enough away from the rest of the data. There are two types of outliers:

- *Outliers* are either 3×*IQR* or more above the third quartile or 3×*IQR* or more below the first quartile.
- *Suspected outliers* are slightly more central versions of outliers: either 1.5×IQR or more above the third quartile or 1.5×IQR or more below the first quartile.

Figure 6 is a boxplot's representation.

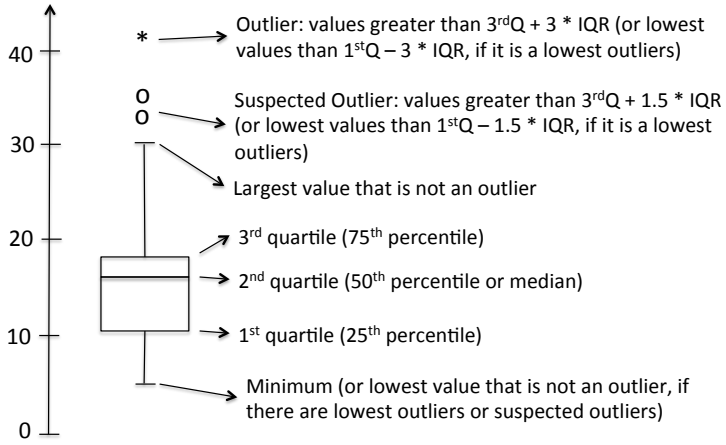

Figure 6 Boxplot

REFERENCES

McCune, S. (2010). *Practice Makes Perfect Statistics*. 1st Edition. United States: McGraw-Hill.

Peers, I. (2006). *Statistical analysis for education and psychology researchers: Tools for researchers in education and psychology*. Routledge.

Rumsey, D. (2010). St*atistics Essentials For Dummies*. New Jersey, Wiley Publishing, Inc.