# Statistics Primer

2022-08-13

# Contents

# A Note About this Resource

---

**What is this?**

This resource has been written using **R**, **LaTeX** and **R Studio** and published through **GitHub**. As the resource is hosted on GitHub it is available anywhere, any time and on any device. This resource supports the embedding of images, video content, html content (iframes), live code and mathematical formula and makes sharing the content much easier. In addition, this resource is capable of embedding Power BI dashboards, Python and R coded visuals.

**Contents**

# Chapter 1

# Introduction

---

## 1.1 What is Statistics?

Statistics is all about the collection, organization, analysis, interpretation and presentation of data. Statistics is used everywhere from opinion polling in politics to predicting the prices of assets. There are two main branches of statistics: descriptive statistics and inferential statistics.

### 1.1.1 Descriptive Statistics

Descriptive statistics describes or summarises data that have been collected. Measures of central tendency such as (mean, median and the mode) and measures of dispersion (range, interquartile range and standard deviation) are the most important tools.

### 1.1.2 Inferential Statistics

Inferential statistical is used to make prediction about a population using information gathered about a sample. Inferential statistics involves hypothesis testing and regression analysis.

# Chapter 2

# Data Types and Levels of Measurement

---

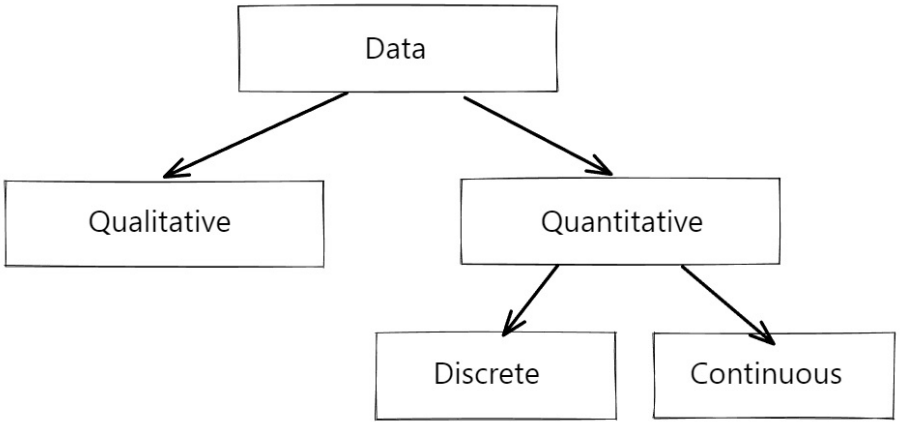
## 2.1 Types of Data

Data can be broadly categorised as **qualitative** (data relating to qualities or characteristics) or quantitative (numerical data relating to sizes or quantities of things).

We can further categorise **quantitative** data as being continuous or discrete.

**Discrete** data involves whole numbers that can't be divided because of what they represent (number of people in a class, number of cars owned). The number of people in a class cannot be 10.5 or 3.14. It must be a whole number because people are not divisible.

**Continuous** data can be divided and measured to some number of decimal places (height, weight, speed in miles per hour). A person's height can be any number (provided it lies within the range of possible human heights) and can be reported to any number of decimal places (150cm or 150.1cm or 150.12cm) depending on how accurate the measurement tool is.




There are also different **levels of measurement**.

## 2.2 Levels of Measurement

The levels of measurement describe how precisely variables are recorded. The different levels of measurement limit which statistics can be used to summarise data and which inferential statistics can be performed. These levels are:

- Nominal
- Ordinal
- Interval
- Ratio

### 2.2.1 Nominal

**Nominal data** is a type of data that is used to label variables. It can be categorised but not ranked (eye colour and gender for instance). The values grouped into these categories have no meaningful order. It is not possible to form a meaningful hierarchy of gender or eye colour.

The only measure of central tendency used with nominal data is the mode.

### 2.2.2 Ordinal

**Ordinal data** is another type of **qualitative data** that groups variables into descriptive categories. The categories used for ordinal data are ordered in some kind of hierarchical scale although the distance between those categories may be uneven or even unknown.

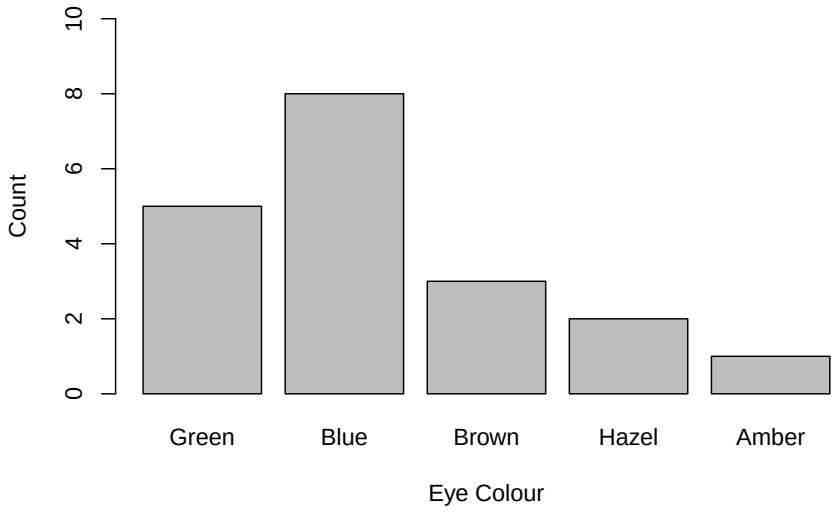



Figure 2.1: Eye colour is an example of nominal data.

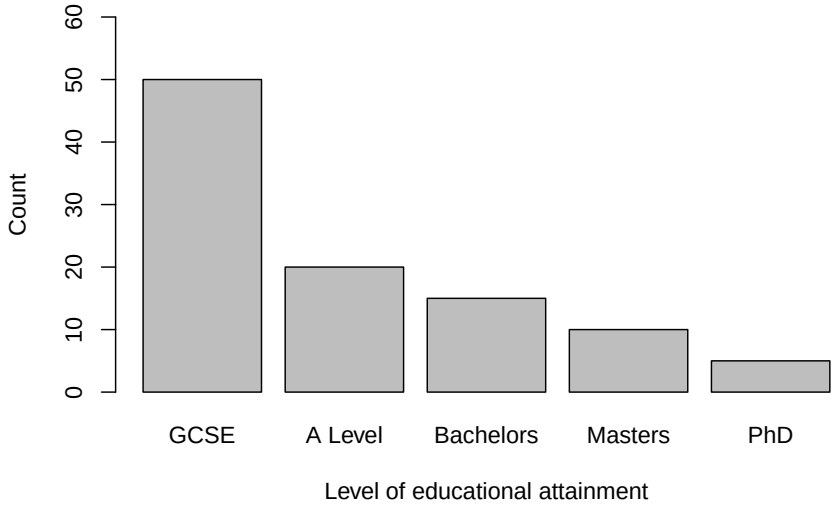



Figure 2.2: The highest level of educational attainment has a heirarchical scale but the distance between categories is unclear.

Ordinal variables often include ratings about opinions that can be categorised (strongly agree, agree, don't know, disagree, strongly disagree).

The descriptive statistics which can be used with ordinal data are the mode and the median.

Ordinal data can also be described with a measure of dispersion, namely, range.

### 2.2.3 Interval

**Interval data** is a type of quantitative data that groups variables into categories. Values can be ordered and separated using an equal measure of distance.

An example of interval level data is temperature data recorded in Celsius or Fahrenheit. The values on either scale are ordered and separated using an equal measure of distance (the distances between notches on a thermometer are always equally spaced).

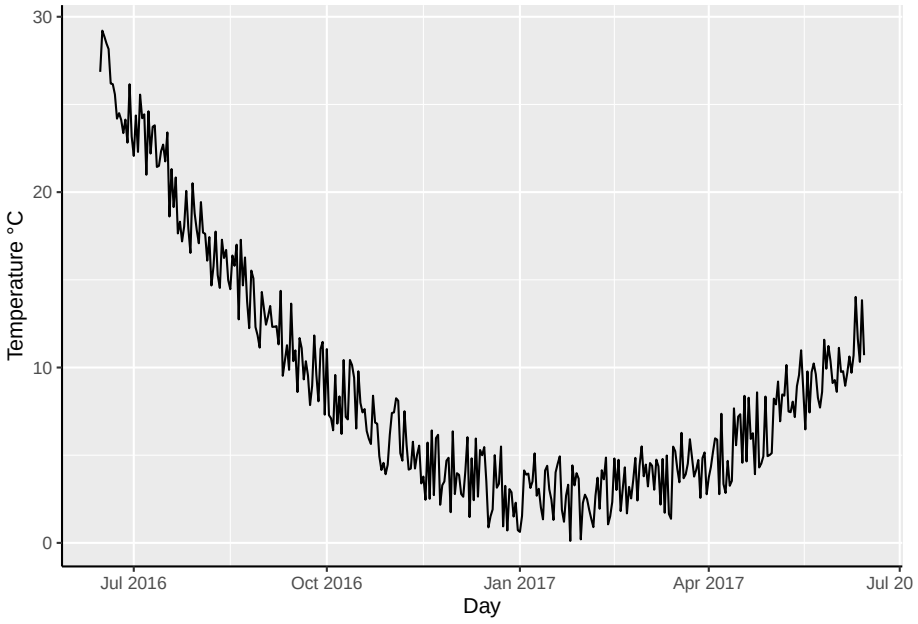



Figure 2.3: Temperature in Celsius is interval data. The values are ordered and separated by an equal interval. The distance between 0°C and 1°C is the same as the distance between 2°C and 3°C.

Mathematical operations can be carried out on this type of data, for instance, subtracting one value from another to find the difference. Interval data lacks a **true zero**.

True zero indicates a lack of whatever is being measured. The Celsius scale doesn't qualify as having a true zero since the zero point in a thermometer is arbitrary. When the Celsius scale was first created by Anders Celsius 0°C was selected to match the boiling point of water and a value of 100 °C was the freezing point of water. The scale was later reversed. Thermometers measure heat and at 0°C there is still heat, maybe not a great deal of it but heat is still measurable meaning 0°C is not a true zero. The thermodynamic Kelvin Scale has a true zero - where particles have no motion and can become no colder (there is a true absence of heat).

A range of descriptive statistics can be used to describe interval data. The measures of central tendency applicable to interval data are the **mode, median** and the **mean**. The measures of dispersion applicable to interval data are the **range, standard deviation** and the **variance**.

### 2.2.4 Ratio

**Ratio data** is a form of quantitative data. It measures variables on a continuous scale with an equal distance between adjacent values (weight, height). Ratio data has a true zero. Ratio data is the most complex of the four data types.

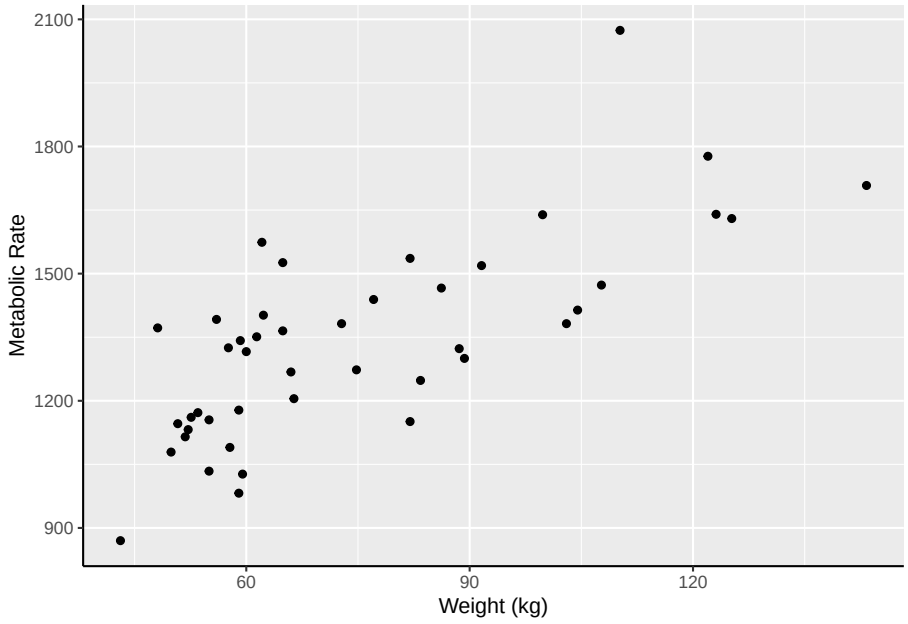



Figure 2.4: The scatterplot above shows metabolic rate plotted against body weight (kg). These are examples of ratio data. This data comes from the Introduction to Statistics with R (ISwR) library for R.

Ratio data can be analysed with descriptive statistics including the **mode, median** and **mean**. **Range, standard deviation, variance** and the **coefficient of variation** can all be used to describe the dispersion of ratio data.

# Chapter 3

# Levels of Measurement Quiz

---

If you would like to try and test your knowledge of the various levels of measurement outlined in this chapter you can take the Levels of Measurement Quiz below. This quiz isn't scored or recorded anywhere.

# Chapter 4

# Describing Data

---

## 4.1 Frequency

The **frequency** of an observation is the number of times it occurs or is recorded.A frequency table, like the one shown below detailing exam grades, is a commonly used method of depicting frequency.

Table 4.1: Frequency Table

| Grade | Frequency |
|---|---|
| A | 15 |
| B | 20 |
| C | 25 |
| D | 21 |
| E | 14 |

The total of all frequencies so far in a frequency distribution is the **cumulative frequency**. It is the 'running total' of frequencies.

Table 4.2: Cumulative Frequency Table

| Grade | Frequency | Cumulative Frequency |
|---|---|---|
| A | 15 | 15 |
| B | 20 | 35 |
| C | 25 | 60 |

| Grade | Frequency | Cumulative Frequency |
|---|---|---|
| D | 21 | 81 |
| E | 14 | 95 |

The **relative frequency** is the ratio of the category frequency to the total number of outcomes. For grade A, the relative frequency is:

$$\frac{15}{15+20+25+21+14} = 0.16.$$

The table can be extended to include the relative frequency.

Table 4.3: Relative Frequency Table

| Grade | Frequency | Relative Frequency |
|---|---|---|
| A | 15 | 0.16 |
| B | 20 | 0.21 |
| C | 25 | 0.26 |
| D | 21 | 0.22 |
| E | 14 | 0.15 |

The **relative frequency** can be reported as a percentage by multiplying the values by 100%. For grade A, the relative frequency reported as a percentage is: 100% x 0.16 = 16%.

## 4.2 Measures of Central Tendency

Measures of central tendency help find the middle, or the average, of a data set.The measures of central tendency are the mean, median and mode.

### 4.2.1 Mean

The mean is the sum of the recorded values divided by the number of values recorded.

#### 4.2.1.1 Example

Find the mean of this list of numbers:

$$2, 3, 3, 4, 20.$$

$$\text{Mean} = \frac{\text{Sum of recorded values}}{\text{Number of values recorded}},$$
$$\text{Mean} = \frac{2 + 3 + 3 + 4 + 20}{5},$$
$$\text{Mean} = \frac{32}{5},$$
$$\text{Mean} = 6.4.$$

### 4.2.2 Median

The **median** is the middle number in a sorted, ascending or descending, list of values.

If there are an odd number of values the median is simply the middle value.

For an even number of values there will be two values in the center. Those values are summed and divided by two.

The median is sometimes used as opposed to the mean when there are outliers that might skew the average of the values.

#### 4.2.2.1 Example

Find the median of this list of numbers:

$$2, 3, 3, 4, 20.$$

There are 5 values listed in ascending order and the middle value is the third value in the list so the median is 3.

Note: In the previous example the mean was 6.4. It was skewed by the outlier (20). The median remains closer to what might be considered to be the middle of the data set.

#### 4.2.2.2 Example

Find the median of this list of numbers:

$$3, 5, 4, 4, 2, 8, 7, 1.$$

The list should be sorted:

$$1, 2, 3, 4, 4, 5, 7, 8.$$

There are an even number of values so there will be two middle values. The middle values are 4 and 4. Sum them and divide by two to get the median: 4.

### 4.2.3 Mode

The mode of a set of data values is the value that appears most often. It is the value that is most likely to be sampled. There can be multiple modes or no modes.

#### 4.2.3.1 Example

Find the mode of this list of numbers:

$$1, 2, 2, 2, 3, 3, 4.$$

A simple way to find the mode is to make a frequency table with the unique values on the left hand side and their frequency on the right hand side. We can tally up how many times each number occurs. Whichever has the greatest frequency is our mode.

Table 4.4:

| Value | Frequency |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 3 | 2 |
| 4 | 1 |

The mode is 2.

#### 4.2.3.2 Example

Find the mode of this list of numbers:

$$7, 3, 5, 3, 4, 3, 5, 6, 8, 5.$$

Table 4.5:

| Value | Frequency |
|---|---|
| 3 | 3 |

| Value | Frequency |
|---|---|
| 4 | 1 |
| 5 | 3 |
| 6 | 1 |
| 7 | 1 |
| 8 | 1 |

This is **bimodal**, it has two modes, 3 and 5.

#### 4.2.3.3 Example

Find the mode of this list of numbers:

$$1, 2, 3, 4, 5, 6.$$

Table 4.6:

| Value | Frequency |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |
| 6 | 1 |

Every value is unique and occurs only once so this data has no mode.

### 4.2.4 Using Excel

It is useful to calculate descriptive statistics by hand for understanding but for larger data sets it is not always possible to arrange data and perform calculations by hand.

Excel has a number of functions designed to perform descriptive statistics.

**Frequency**

=FREQUENCY(start:end,bins_array)

The frequency() function will return a frequency table describing your data. It takes two arguments, the first being the array of values and the second being an array describing the upper boundary of the bins used.

**Average**

=AVERAGE(start:end)

The mean is calculated using the average() function. There are several other functions relating to means: geomean(), harmean() and trimmean(). Take care not to use these as they are quite different from calculating the mean that has been described here.

**Median**

=MEDIAN(start:end)

The median is calculated using the median() function.

**Mode**

=MODE.SNGL(start:end)

=MODE.MULT(start:end)

There are several functions for calculating the mode: mode(), mode.sngl() and mode.mult(). mode() was used in Excel 2007 and may still appear as an option in some versions of Excel. mode.sngl() will return one mode and mode.mult() will return multiple modes (if there are multiple modes).

Neither mode() nor mode.sngl() will warn you if there are multiple modes so mode.mult() is usually the safest option.

#### 4.2.4.1 Example

In the example below, the variable name is in cell A1 and the values are in cells A2 to A18. To calculate the average, type "=AVERAGE(A2:A18)" in another cell. It doesn't matter which cell but in this example A19 has been used. Press enter to return the value.

## 4.3 Measures of Dispersion

**Dispersion** (or variability) describes how far apart data points lie from each other and the center of a distribution. The **range, interquartile range, variance** and **standard deviation** are all measures of dispersion and they describe how far apart data points lie from one another and the center of a distribution.

### 4.3.1 Range

The range is the difference between the highest and lowest values and is calculated by subtracting the minimum value from the maximum value.

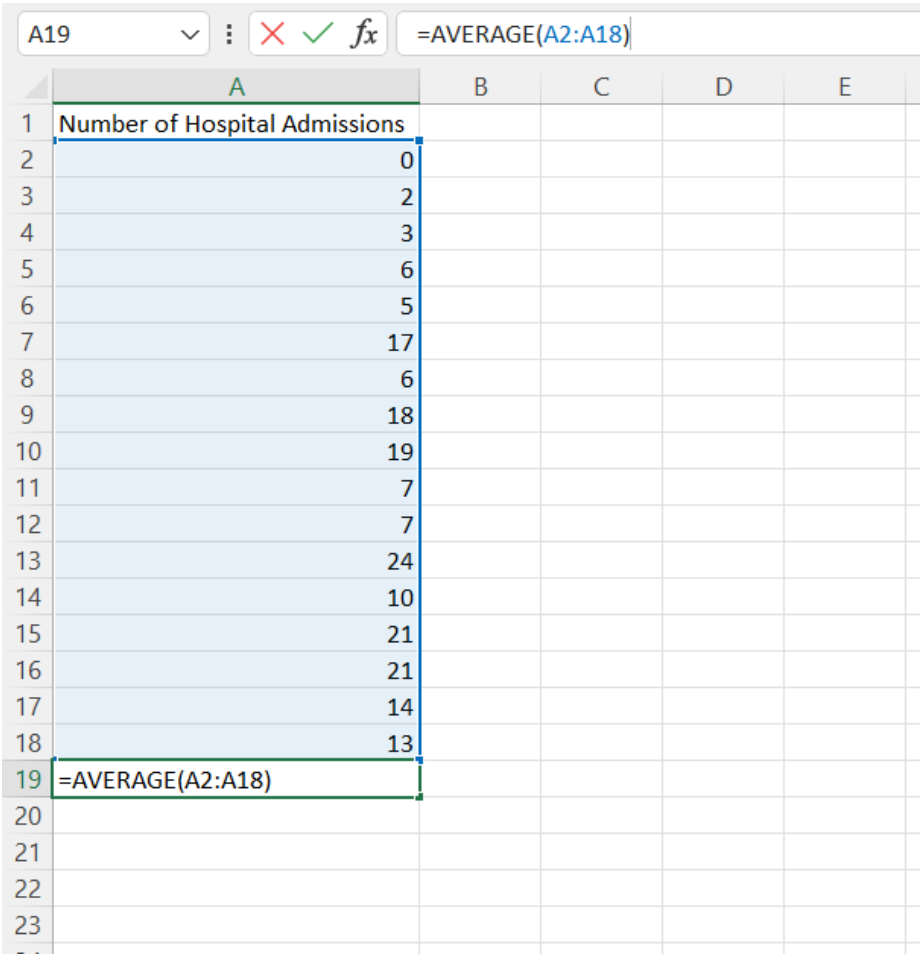



Figure 4.1: Screen shot showing excel spreadsheet with a list of values and the average being calculated using the average() function.

#### 4.3.1.1 Example

Calculate the range for the following set of numbers:

$$23, 42, 75, 19, 74.$$

First, arrange the values in ascending order:

$$19, 23, 42, 74, 75.$$

The maximum value is 75 and the minimum is 19.

$$\text{Range} = 75 - 19,$$
$$\text{Range} = 56.$$

### 4.3.2 Interquartile Range

The **interquartile range** (IQR) describes the spread of the middle half of a distribution. How the interquartile range is calculated depends on whether there are an even or an odd number of values in a dataset.

For an even number of values the dataset in split half. The medians for the two new subsets of data are calculated. The positive difference of those medians is the interquartile range. For an odd number of values either the inclusive or the exclusive method of finding the interquartile range must be used.

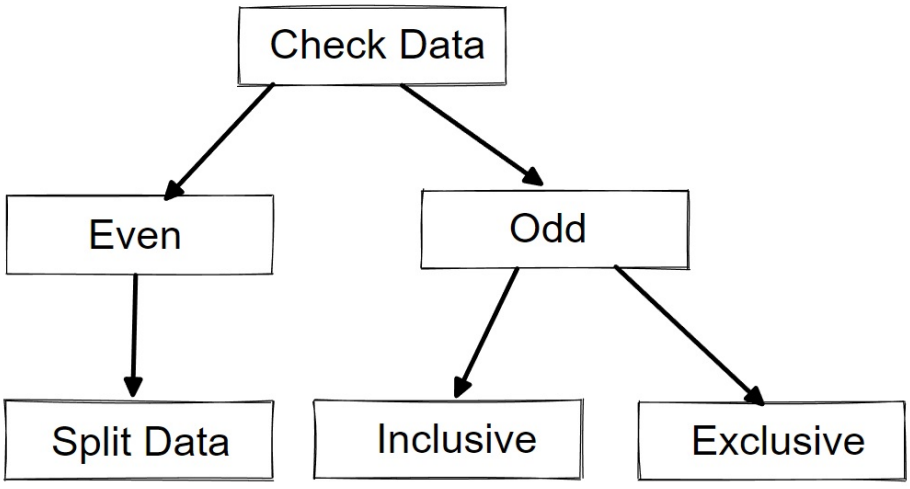



Figure 4.2: Tree diagram showing process of deciding how to calculate the IQR

The algorithm for the **exclusive method** is detailed below:

1. Arrange the data in numeric order.
2. Remove the median and split the data about its center.
3. Find the medians of the two newly appended subsets of data.
4. Calculate the difference.

The algorithm for the **inclusive method** is detailed below:

1. Arrange the data in numeric order.
2. Remove the median and split the data about its center.
3. Append the two new subsets of data with the median.
4. Find the medians of the two newly appended subsets of data.
5. Calculate the difference.

#### 4.3.2.1 Example

**Even**

Find the interquartile range for the list of numbers below:

$$6, 7, 8, 8, 7, 6, 9, 5, 10, 4.$$

There are an even number of values. Arrange them in numeric order:

$$4, 5, 6, 6, 7, 7, 8, 8, 9, 10.$$

Split the values about their center into two sub sets of data.

$$(4, 5, 6, 6, 7), (7, 8, 8, 9, 10).$$

Find the medians of each of these sub sets. The first subset has a median of 6 while the second has a median of 8.

The interquartile range is:

$$\text{IQR} = 8 - 6 = 2.$$

Note: To calculate the interquartile range the smaller median value is always subtracted from the larger.

**Odd (Exclusive Method)**

Find the interquartile range for the list of numbers below:

$$2, 3, 2, 4, 3, 5, 4, 4, 2.$$

Arrange the values in numeric order:

$$2, 2, 2, 3, 3, 4, 4, 4, 5.$$

Remove the median (3) and split the data as before:

$$(2, 2, 2, 3), (4, 4, 4, 5).$$

The interquartile range is:

$$\text{IQR} = \text{Median of sub set 2} - \text{Median of sub set 1},$$
$$\text{IQR} = \frac{4+4}{2} - \frac{2+2}{2} = \frac{8}{2} - \frac{4}{2} = 4 - 2 = 2.$$

**Odd (Inclusive Method)**

Find the interquartile range of the list of numbers below:

$$2, 3, 2, 4, 3, 5, 4, 4, 2.$$

Sort in numeric order as before:

$$2, 2, 2, 3, 3, 4, 4, 4, 5.$$

Split the data as before but append each subset of data with the median (at the end and start of each subset respectively):

$$(2, 2, 2, 3, 3), (3, 4, 4, 4, 5).$$

Find the medians of each of the subsets and calculate the interquartile range. The median of the first subset is 2 and the median of the second subset is 4.

$$\text{IQR} = 4 - 2 = 2$$

The interquartile range is a useful measure of variability for skewed distributions. It can show where most values lie and how clustered they are. It is useful for datasets with outliers as it is based on the middle half of the distribution and less influenced by extreme values. Exclusive calculations result in a wider interquartile range than inclusive calculations.

### 4.3.3 Variance and Standard Deviation

The standard deviation describes to what extent a set of numbers lie apart (their spread). It is the square root of variance which is also an indicator of the spread of values.

**Variance**

To calculate the variance:

1. Start by finding the mean of the values in the dataset.
2. Find the difference between each recorded value and the mean.
3. Square those differences.
4. Sum the squared differences.
5. Divide the sum by the number of values recorded for population variance or the sum of the number of values minus 1 for sample variance.

### Standard Deviation

Taking square root of the variance corrects for the fact that all the differences were squared, resulting in the standard deviation.

The plot below shows three distributions of values, each with a mean of 30 but with different standard deviations. In statistics there is a rule called the empirical rule that states that 68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively.

To make sense of this through an example, the plot below shows some simulated data for test scores. Three groups given the same test could achieve the same average score but with greater or lesser spreads of scores.

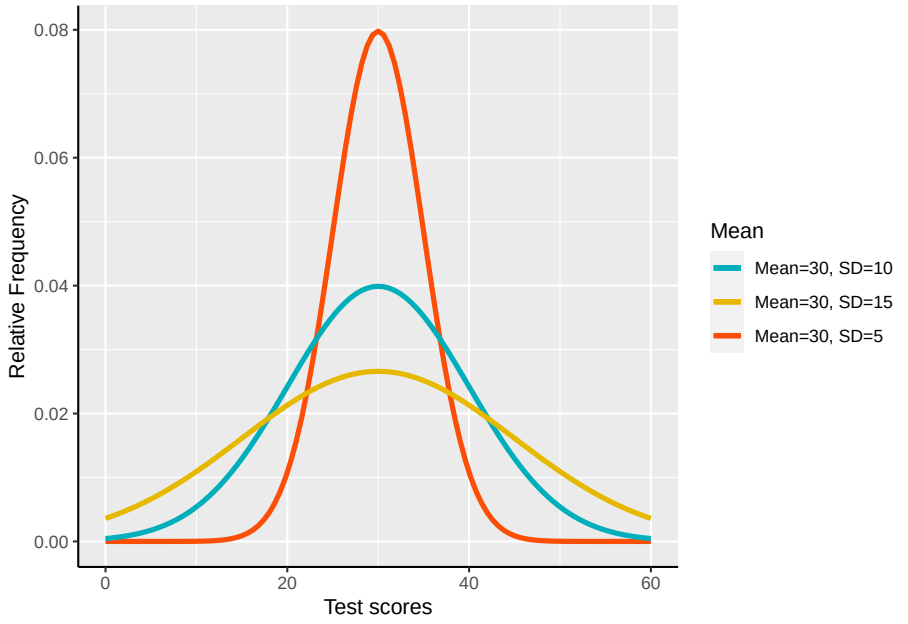



Figure 4.3: Plot showing several distributions of simulated test score data the same means but differing standard deviations.

For a mean of 30 and standard deviation of 5: 68% of the values will lie within the range 25-35.

For a mean of 30 and standard deviation of 10: 68% of the values will lie within the range 20-40.

For a mean of 30 and standard deviation of 15: 68% of the values will lie within the range 15-45.

This is particularly clear with a mean of 30 and a standard deviation of 5 as most of the values are tightly packed within the range 25-35.

#### 4.3.3.1 Example

Calculate the sample estimate of variance and sample estimate of standard deviation for the following list of values:

$$2, 4, 4, 5, 6.$$

Start by finding the mean of the values in the dataset:

$$\text{Mean} = \frac{2+4+4+5+6}{5} = 4.2.$$

Find the difference between each recorded value and the mean.

Table 4.7:

| Value | Difference |
|---|---|
| 2 | 2 - 4.2 = -2.2 |
| 4 | 4 - 4.2 = -0.2 |
| 4 | 4 - 4.2 = -0.2 |
| 5 | 5 - 4.2 = 0.8 |
| 6 | 6 - 4.2 = 1.8 |

Square the differences.

Table 4.8:

| Value | Difference | Squared Difference |
|---|---|---|
| 2 | -2.2 | 4.84 |
| 4 | -0.2 | 0.04 |
| 4 | -0.2 | 0.04 |
| 5 | 0.8 | 0.64 |
| 6 | 1.8 | 3.24 |

Sum the squared differences.

$$\text{Sum} = 4.84 + 0.04 + 0.04 + 0.64 + 3.24 = 8.8.$$

Divide the sum by the number of values recorded minus one to get the sample estimate of variance.

$$\text{Variance}_S = \frac{8.8}{5-1} = 2.2.$$

To get the sample estimate of the standard deviation take the square root of this value:

$$\text{Standard Deviation}_S = \sqrt{\text{Variance}_S} = \sqrt{2.2} = 1.48.$$

### 4.3.4 Using Excel

Calculating the variance and standard deviation by hand is a long process and due to the number of steps involved it is prone to error. Excel, SPSS, Python and R all have functions which allow users to calculate these descriptive statives and their use is highly recommended over calculating the statistics by hand.

**Range**

=MAX(start:end)-MIN(start:end)

**Standard Deviation**

=STDEV.S(start:end) =STDEV.P(stard:end)

stdev.s() estimates standard deviation based on a sample. stdev.p() calculates standard deviation based on the entire population given as arguments.

**Variance**

=VAR.S(start:end) =VAR.P(start:end)

var.s() estimates variance based on a sample. var.p() calculates variance based on the entire population given as arguments.

# Chapter 5

# Descriptive Statistics Quiz

---

If you would like to try and test your knowledge of the various levels of measurement outlined in this chapter you can take the Measures of Central Tendency and Measures of Dispersion Quiz below. This quiz isn't scored or recorded anywhere.

# Chapter 6

# Comparing Data

---

Statisticians use several statistical measures like the percentage difference, percentage change and percentage error to evaluate the differences between measured values. All three differ in what they measure.

Percentage difference is the difference between two values divided by the average of two values multiplied by 100%. This is typically used to understand how close two values are to one another.

If the data is tracking values over time (comparing old values to new values) then you should calculate the percentage change instead of the percentage difference. There is a key difference between the two. While percentage change aims to measure change over time, the percentage difference seeks to understand the difference between two averages.

## 6.1 Percentage Difference

The percentage difference is used to compare two values.

$$\text{Percentage Difference} = 100\% \frac{|Value_1 - Value_2|}{\frac{Value_1 + Value_2}{2}}$$

The | symbol in the formula below indicates that the 'absolute' value (the result of the calculation if you were to ignore whether the result was positive or negative) of the calculation should be taken. For instance:

$$|3 - 2| = 1,$$

$$|10-5| = 5,$$
$$|7-9| = 2,$$

7-9 is equal to -2 but when we take the 'absolute' value (|7-9|) we ignore the negative sign and report the result as 2.

### 6.1.1 Example

One researcher produced thirteen research reports in 2022 another produced 11. What is the percentage difference?

$$\text{Percentage Difference} = 100\% \frac{|13-11|}{\frac{13+11}{2}} = 100\% \frac{2}{\frac{24}{2}} = 16.7\%$$

## 6.2 Percentage Change

Percentage change is about comparing old to new values. The formula for calculating a percentage change is given below:

$$\text{Percentage Change} = 100\% \frac{\text{New Value} - \text{Old Value}}{\text{Old Value}}.$$

### 6.2.1 Example

What is the percentage change in the population in these three places between 2018 and 2019?

Table 6.1:

| Location | 2018 | 2019 | Percentage Change (%) |
|---|---|---|---|
| Somewhere | 50 | 36 | -28 |
| Anywhere | 50 | 50 | 0 |
| Elsewhere | 50 | 58 | 16 |

Use the formula above to calculate the percentage change for Upper Braniel:

$$\text{Percentage Change} = 100\% \frac{\text{New Value} - \text{Old Value}}{\text{Old Value}},$$

$$\text{Percentage Change} = 100\% \frac{36-50}{50} = 100\% \frac{-14}{50} = -28\%.$$

A negative percentage change indicates a percentage decrease while a positive percentage change indicates a percentage increase.

## 6.3 Percentage Point Change

Note that subtracting one percentage from another gives the percentage point change rather than the percentage change.

### 6.3.1 Example

What is the percentage point change in the population in these locations between 2018 and 2019?

Table 6.2:

| Location | 2018 | 2019 | Percentage Change (%) |
|---|---|---|---|
| Somewhere | 3.5 | 2.5 | -1 |
| Anywhere | 3.4 | 3.3 | -0.1 |
| Elsewhere | 3.3 | 3.8 | 0.5 |

# Chapter 7

# Correlation

---

## 7.1 Correlation

Correlation refers to the relationship (association) between two or more variables. A correlation coefficient (usually Pearson's r or Spearman's rho) quantifies the relationship:

- r = 1 represents a perfect positive association
- r = -1 represents a perfect negative association
- r = 0 represents a lack of association

## 7.2 Examples

### 7.2.1 Positive Association

The plot above shows a positive association between the number of hours worked and wages earned. As the number of hours worked increases so too do the wages.

### 7.2.2 Negative Association

The plot above shows a negative association between the spending and savings. As more money is spent, less is saved.

### 7.2.3 No Association

Ice cream sales plotted against dog food production. There is no association and r=0.

The plot above shows a positive association between the number of hours worked and wages earned. As the number of hours worked increases so too do the wages.

## 7.3 Understanding Correlation

It is important to note that correlation does not imply causation. Just because two variables are related, does not mean that one causes the other.

The number of drownings and the sales of ice cream are highly correlated but that doesn't mean one causes the other.

Drowning's and sales of ice cream are higher in the Summer months when weather is warmer. A third variable (good weather) causes both.