

# Statistics Primer

2022-08-13



# Contents

<b>Statistics Primer</b>	<b>5</b>
<b>1 Introduction</b>	<b>7</b>
1.1 What is Statistics? . . . . .	7
<b>2 Data Types and Levels of Measurement</b>	<b>9</b>
2.1 Types of Data . . . . .	9
2.2 Levels of Measurement . . . . .	10
<b>3 Levels of Measurement Quiz</b>	<b>15</b>
<b>4 Describing Data</b>	<b>17</b>
4.1 Frequency . . . . .	17
4.2 Measures of Central Tendency . . . . .	18
4.3 Measures of Dispersion . . . . .	22
<b>5 Descriptive Statistics Quiz</b>	<b>31</b>
<b>6 Comparing Data</b>	<b>33</b>
6.1 Percentage Difference . . . . .	33
6.2 Percentage Change . . . . .	34
6.3 Percentage Point Change . . . . .	35

<b>7</b>	<b>Data Visualisation</b>	<b>37</b>
7.1	Data Visualisation . . . . .	37
7.2	Visual Cues . . . . .	37
7.3	Relationships in Data . . . . .	37
7.4	Why Visualise Data? . . . . .	42
7.5	Data Visualisation Tools . . . . .	42
7.6	Dynamic Visualisations (Dashboards) . . . . .	43
<b>8</b>	<b>Correlation</b>	<b>45</b>
8.1	Correlation . . . . .	45
8.2	Examples . . . . .	45
8.3	Understanding Correlation . . . . .	46
<b>9</b>	<b>Samples</b>	<b>47</b>
9.1	What is a Sample? . . . . .	47
<b>10</b>	<b>Confidence Intervals</b>	<b>49</b>
10.1	Confidence Intervals . . . . .	49
10.2	Example . . . . .	50
10.3	Example . . . . .	50
<b>11</b>	<b>Hypothesis Testing &amp; Statistical Significance</b>	<b>53</b>
11.1	Hypothesis Testing . . . . .	53
11.2	Statistical Significance . . . . .	54

# Statistics Primer

---

## 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

- Statistics
  - Introduction
  - Data Types and Levels of Measurement
  - Describing Data
  - Comparing Data
  - Data Visualisation
  - Correlation
  - Samples and Sampling
  - Confidence Intervals
  - Hypothesis Testing & Statistical Significance

## Summary

The **introduction** describes the main differences between inferential and descriptive statistics.

**Data Types and Levels of Measurement** provides an overview of the different types of data statisticians encounter and how they decide on which descriptive and inferential statistics to use.

**Describing Data** provides definitions of the measures of central tendency and dispersion as well as worked examples.

**Comparing Data** provides definitions and worked examples for percentage difference and percentage change.

**Data Visualisation** outlines how statisticians approach visualising different types of data.

**Correlation** describes the relationship (or association) between variables and how it is measured.

**Samples** provides an introduction to sampling.

**Confidence Intervals** explains what confidence intervals are and how they are calculated with worked examples.

**Hypothesis Testing & Statistical Significance** describes how statisticians and researchers use hypothesis testing and outlines how statistical significance is interpreted.

# 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).

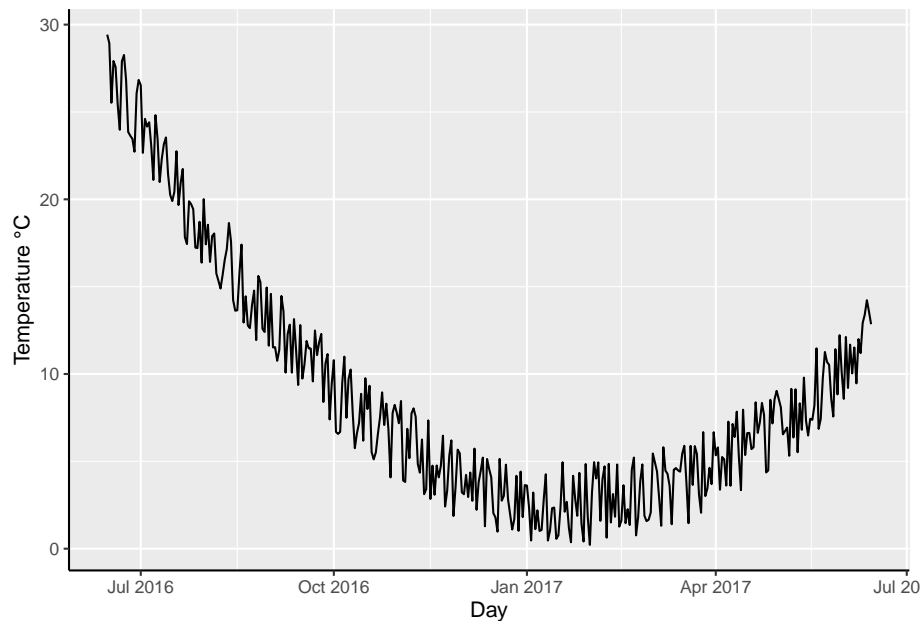


Figure 2.3: Temperature in Celsius is interval data. The values are ordered and separated by an equal interval. The distance between  $0^{\circ}\text{C}$  and  $1^{\circ}\text{C}$  is the same as the distance between  $2^{\circ}\text{C}$  and  $3^{\circ}\text{C}$ .

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  $0^{\circ}\text{C}$  was selected to match the boiling point of water and a value of  $100^{\circ}\text{C}$  was the freezing point of water. The scale was later reversed. Thermometers measure heat and at  $0^{\circ}\text{C}$  there is still heat, maybe not a great deal of it but heat is still measurable meaning  $0^{\circ}\text{C}$  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.

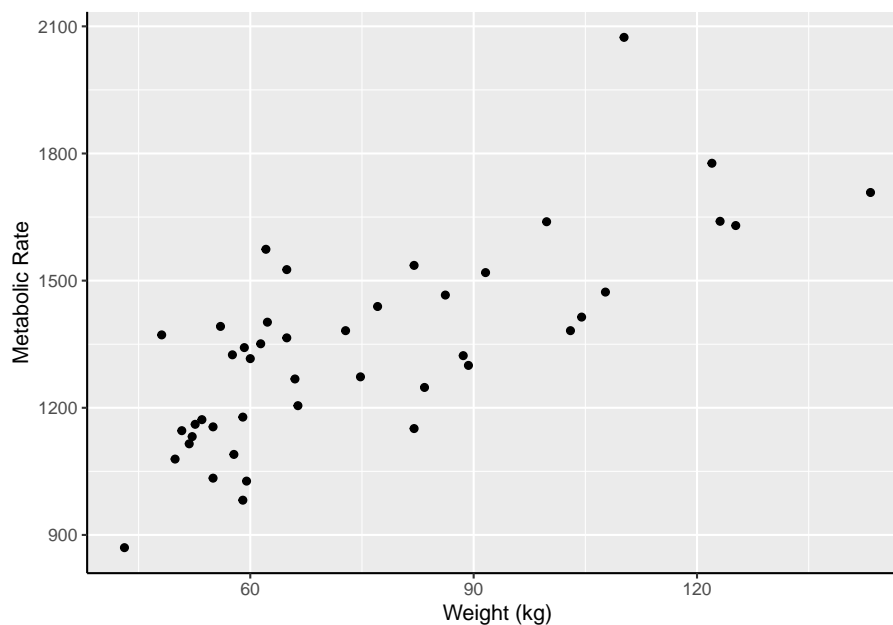


Figure 2.4: The scatterplot above shows metabolic rate plotted against body weight (kg). These are examples of ratio data. This data comes from the Introduction to Statistics with R (ISwR) library for R.

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:  $100\% \times 0.16 = 16\%$ .

## 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:

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

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

### 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

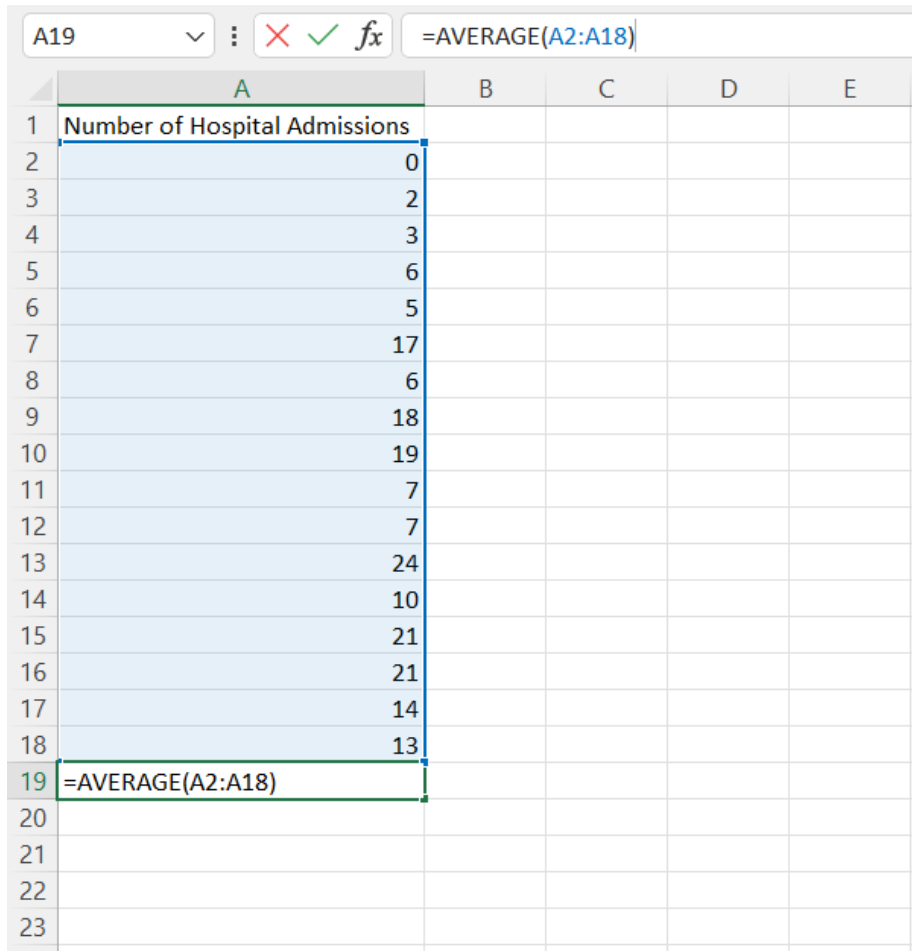
In the example below, the variable name is in cell A1 and the values are in cells A2 to A18. To calculate the average, type “=AVERAGE(A2:A18)” in another cell. It doesn’t matter which cell but in this example A19 has been used. Press enter to return the value.

## 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.



The image shows an Excel spreadsheet with a list of hospital admissions in column A and an average calculation in cell A19. The formula bar at the top shows the formula `=AVERAGE(A2:A18)`. The data in column A is as follows:

	A	B	C	D	E
1	Number of Hospital Admissions				
2	0				
3	2				
4	3				
5	6				
6	5				
7	17				
8	6				
9	18				
10	19				
11	7				
12	7				
13	24				
14	10				
15	21				
16	21				
17	14				
18	13				
19	=AVERAGE(A2:A18)				
20					
21					
22					
23					

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.

Range =  $75 - 19$ ,

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 is 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:

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

#### 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.

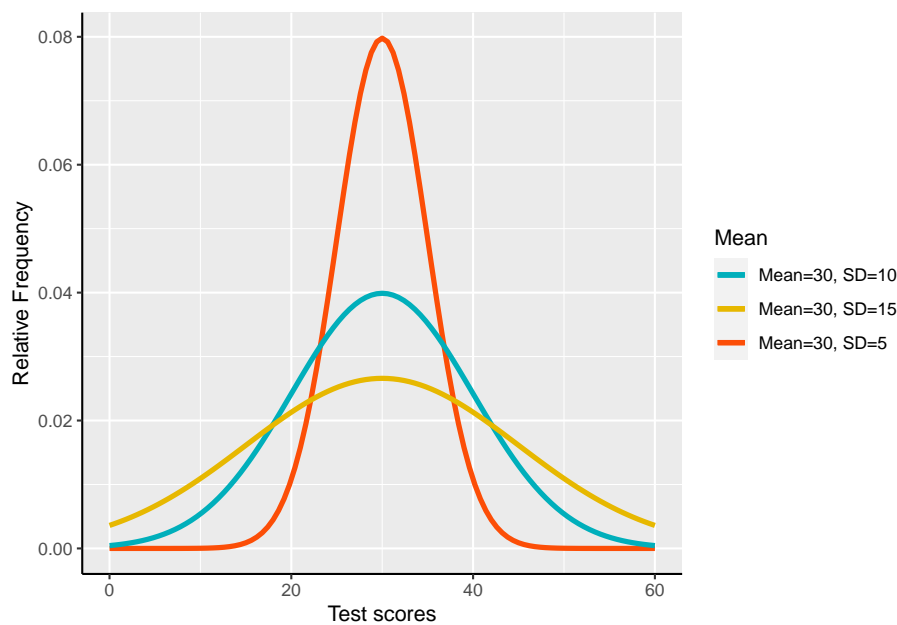


Figure 4.3: Plot showing several distributions of simulated test score data the same means but differing standard deviations.

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 statistics 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(start: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

# Data Visualisation

---

### 7.1 Data Visualisation

Data visualisation is formally defined as the encoding of data using visual cues such as variations in the size, shape and colour of geometric objects (points, lines, bars). The encoding is generally informed by the relationships within the data.

In the bar chart example below the frequencies of different eye colours have been mapped to the heights of the bars.

### 7.2 Visual Cues

Whether data is visualised using points, lines, bars or something else entirely is largely determined by the relationships within the data. Some of the visual cues and relationships used to inform data visualisation are shown below.

The illustration above shows some of the visual cues used to encode data. Magnitudes are typically mapped to sizes of objects. Colour is often used to represent quantities or highlight data. Shapes can be used to represent qualitative data.

### 7.3 Relationships in Data

The Government Statistical Service has produced guidance on the relationships in data and how they inform chart choices. The guidance can be useful and some of the key points are summarised below.

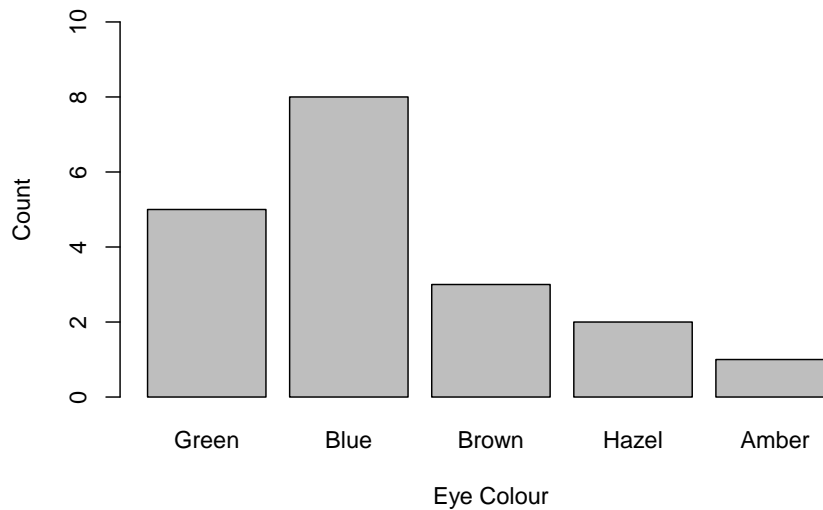


Figure 7.1: The frequency of each observation has been mapped to the height of the bars in the visual.

### 7.3.1 Frequency Distributions

Histograms and bar charts are useful for showing category frequencies. Population by age band for instance could be visualised using a histogram or bar chart. A boxplot can also be useful in visualising additional descriptive statistics such as the mean, median, quartiles, outliers and the range.

### 7.3.2 Time Series

A line chart is often used to demonstrate the trend of a variable over some time period. For instance, temperature over time can be visualised with a line chart.

### 7.3.3 Rankings

Data that is ranked usually consists of categories presented in ascending or descending order. A bar chart may be used to show the comparisons between the different categories. Sometimes, change in ranking over time is shown through slope charts but usually only when comparing a start date and an end date without consideration for the time period in between.

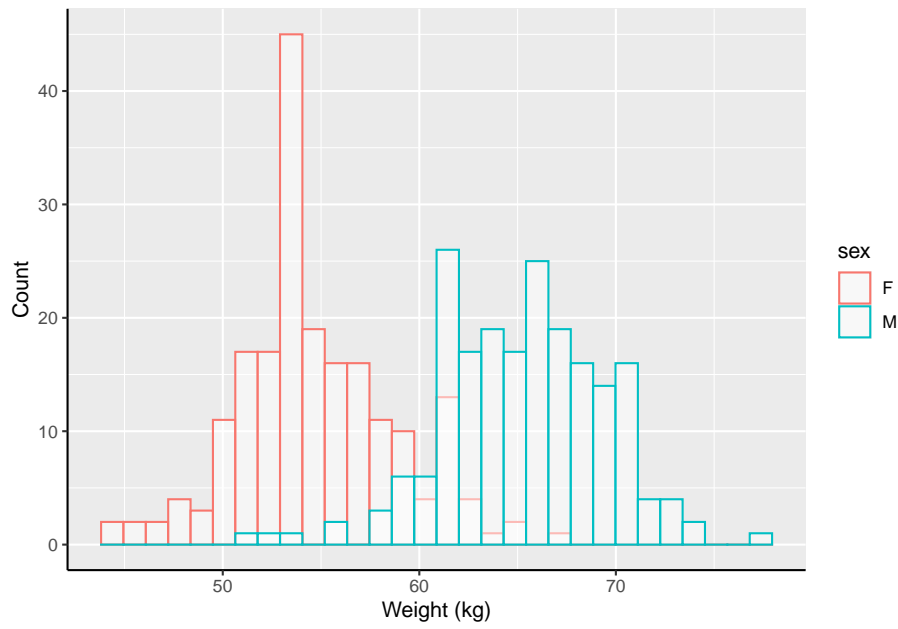


Figure 7.2: Simulated weight data illustrating the use of a histogram.

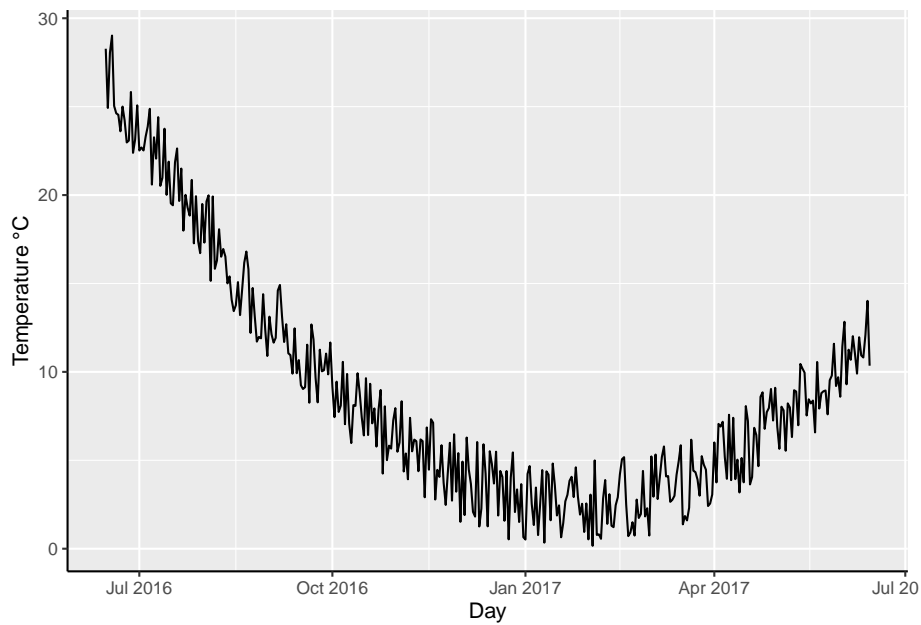
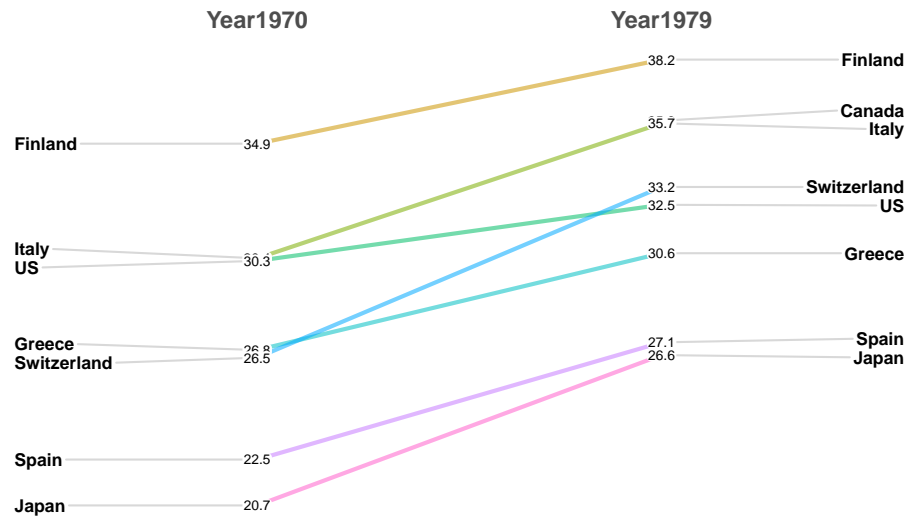


Figure 7.3: Simulated temperature data.

**GDP evolution**

1970–1979



Produced using the R Charts library

Figure 7.4: Slope chart showing changes in GDP between 1970 and 1979.

**7.3.4 Deviation**

Deviation from a reference value can be shown through bar charts.

**7.3.5 Correlation**

Correlation is usually visualised using scatterplots. Scatterplots are a good way to show comparisons between observations of two variables to determine if there is some correlation because it quickly becomes apparent if there is correlation between the variables or not.

**7.3.6 Magnitude**

Comparing differences in the magnitudes of values often relies on bar charts. Comparing the total number of research papers by journal for instance.

**7.3.7 Spatial**

Cartograms and heat mapping are common ways to show differences between geographical regions.



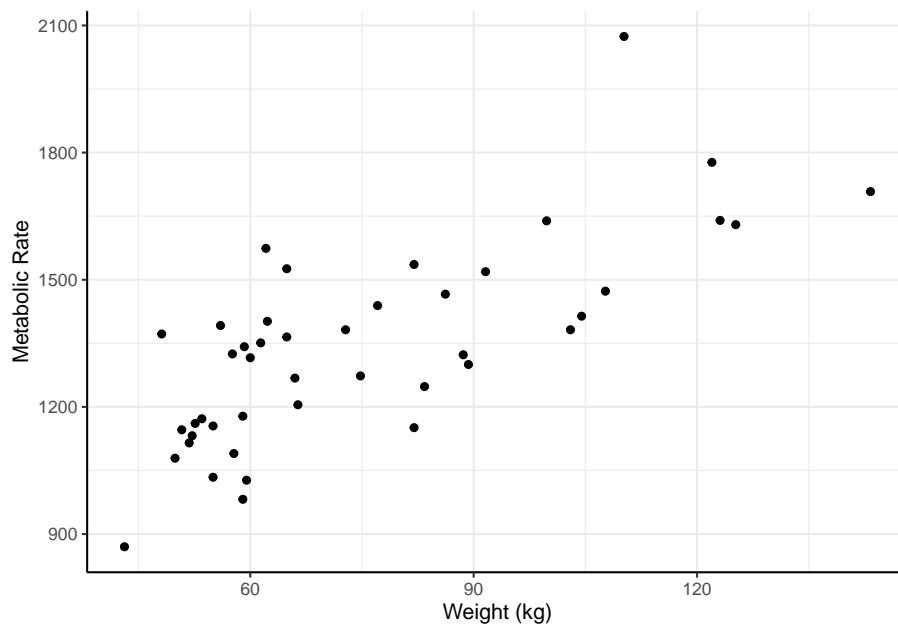


Figure 7.5: The scatterplot above shows metabolic rate plotted against body weight (kg). This is an example of correlation. This data comes from the Introduction to Statistics with R (ISwR) library for R.

## 7.4 Why Visualise Data?

In general, people are better at recognising differences in shapes, colours and sizes than they are at identifying the number of times a value occurs or the differences between values in a large excel spread sheet. For this reason data visualisation can be used to find errors in data quickly. It's much easier to recognise an anomalous value on a bar chart than in an Excel spread sheet. Data visualisation can also be used to see patterns that are difficult to determine by looking at raw data. Data visualisation can also be used to:

- Answer research questions.
- Discover new research questions.
- Explain complex relationships in data visually.
- Aid in decision making.
- Engage and inform.

## 7.5 Data Visualisation Tools

New programming languages and software products have made data analysis and visualisation vastly more accessible. In addition, many of these facilitate dynamic or interactable visualisations. There is an ever expanding ecosystem of data visualisation tools (many of which have been used in this document) including:

- **Excel** and **SPSS** produce high quality visualisations and while dynamic visuals are not their focus they are often the simplest and most time efficient option for visualising data.
- **Genially** is an online tool for creating interactive and animated content that is particularly effective for presentations.
- **Tableau** and Power BI are visual analytics platforms which are well suited to the development of dashboards to visualise complex interconnected data sets.
- **Flourish** can be used to produce interactive visuals although its functionality is more limited than Power BI or Tableau.
- **Javascript** facilitates data visualisation through its D3 library. D3 has a steep learning curve as it requires JavaScript skills to use it effectively however it offers a greater degree of customisation and a broader spectrum of visualisation options as a result.
- **Python** libraries such as Matplotlib, Seaborn and Plotly can also be used to visualise data. The learning curve is steep as it requires programming skills to use Python effectively however Python offers customisation options that are not available in Excel or Power BI.

- **R** is another useful tool with libraries such as `ggplot2` which can be used to visualise data. This is the programming language used to develop this resource and many of the visualisations throughout.

## 7.6 Dynamic Visualisations (Dashboards)

There are a number of considerations when developing dynamic data visualisations (sometimes called dashboards) as not all data visualisations need to be dynamic.

Considering the audience, objectives and what visuals will be most appropriate to communicate data can help in determining whether a dynamic or interactive visualisation is needed.

Dashboard style visualisations are best suited to data reporting where there is a need to repeatedly produce the same visuals or reports either daily, monthly, quarterly or annually.

Power BI is well designed for these types of visualisation requirements as it offers automation options enabling data sets to be refreshed at regular time intervals. Automation can be as simple as setting a refresh time in the Power BI dashboard and manually updating the excel file it stores in memory or it can be more complex and involve using programming languages to make API calls and perform automated calculations.

Producing dynamic visualisations is often considerably more time expensive than producing static visuals and time constraints should be considered before developing a dashboard visualisation.

### 7.6.1 Best Practice

GSS have produced guidance on designing dashboards that covers most aspects of dashboard design. The content below summarises some of the key points in this guidance.

#### **Consider Audience and User Needs**

Consider the user needs and whether a dashboard is really needed. Often the simplest solution (bar charts drawn in Excel or SPSS) is the best. Consider the visuals used and whether they're the best way to communicate the data. Sometimes tables or even text can communicate data better than a visual.

#### **Guidance**

Providing guidance on how to use a dynamic visual or dashboard is important as many users will not be familiar with interactive dashboards. Guidance can be provided through supplementary documentation, blog text if the visual is

being embedded, or it can be provided through tool tips and information pages in the dashboard itself.

### **Streamline Content**

When adding any new data or visuals it is important to ask whether it adds value. Try to group related content and streamlining the content to guide the users through the data.

### **Automate**

Automation can be simple or complex, it can be achieved by setting a refresh date in a Power BI dashboard. It can also involve the use of programming languages to make API calls, web scrape data and perform calculations. Automation typically results in less manual updating and a reduced chance of error and can make the management of the product less resource intensive. It's important to note that automation does not necessarily mean less work, the scripts used to automate processes will need to be updated as languages are developed and updated over time.

### **Consider Design Principles**

Give your dashboard a header and dedicated areas for visuals. Consider other dashboards you have seen in the past and draw inspiration from web design. Most websites have a navigation bar at the top, lists with filters along the left or right hand side and content in the center of the page. Think about things like symmetry, flow and a consistent style or layout. Use white space where possible and try to avoid cluttered visualisations.

### **Ensure Accessibility**

Ensure your product is accessible by checking the colour contrast ratios of text and including alt text in your visualisations where possible. Ensure the fonts are large enough to read and avoid using multiple fonts.

# Chapter 8

## Correlation

---

### 8.1 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

### 8.2 Examples

#### 8.2.1 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.

#### 8.2.2 Negative Association

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

### 8.2.3 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.

## 8.3 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.

## Chapter 9

# Samples

---

### 9.1 What is a Sample?

A sample is a small subset of a larger set of data.

The purpose of using a sample is to draw inferences about a wider population. The voting intentions of 1,000 people (sample) might be used to predict the outcome of a general election (population).

Procedures for converting sample responses to population estimates are known as inferential statistics.

#### 9.1.1 Representative Sampling

To generalise the results from a sample to the full population, the sample must be representative of the population. If, in a project to gauge the view of the Northern Ireland population, only Antrim residents are surveyed then the sample results cannot be used to infer attitudes of all Northern Ireland residents.

#### 9.1.2 Sampling Error

Surveys are based on a sample rather than the whole population so they are subject to sampling error.

The sampling error is the difference between the sample estimate and the ‘true’ value (which would have been obtained if a census of the whole population were undertaken).

**9.1.2.1 Example**

If one measures heights of 1,000 individuals in Northern Ireland, the average height of the sample is typically not the same as the average height of all 1.8 million people in the country. The difference between the sample and the population values is considered a sampling error. If the sample mean is 176cm and the population mean is 178cm then the sampling error is 2cm.

The exact measurement of sampling error is generally not feasible, since the true population values are not known.

Sampling error however can be estimated by techniques such as the calculation of confidence intervals.



## Chapter 10

# Confidence Intervals

---

### 10.1 Confidence Intervals

A confidence interval (CI) is a range of values that would likely contain the true value.

Consider a study to estimate the mean weight of all 10 year old boys in Northern Ireland. It would be impractical to weigh them all so a sample of 16 might be taken. The mean weight of the sample might be 45kg. This is a point estimate of the population mean.

This point estimate has limited utility because it does not reveal uncertainty associated with the estimate. Is there confidence that the population mean is within 5kg of 45kg? It's not possible to know with this information.

That is why confidence intervals are calculated.

The calculation of a confidence interval is not a simple process. There are a number of things that are needed:

- The number of measurements recorded.
- The mean of those measurements and the standard deviation.
- A 'Z-score'.

Before continuing it might be useful to review variance and standard deviation.

## 10.2 Example

Table 10.1:

Group	Rate (%)	CI (%)	Lower Limit (%)	Upper Limit (%)
Protestant Males	53.3	+/-2.6	50.7	55.9
Roman Catholic Males	46.7	+/-2.6	44.1	49.3
Protestant Females	52.6	+/-2.7	49.9	55.3
Roman Catholic Females	47.4	+/-2.7	44.7	50.1
Protestant both sexes	53.0	+/-1.9	51.1	54.9
Roman Catholic both sexes	47.0	+/-1.9	45.1	48.9

Based on a sample, the table above shows that 52.6 % of Protestant females (C.I. = +/- 2.7) were estimated to be economically active in 2011.

This means that there is 95% confidence that the ‘true value’ lies somewhere between 49.9% and 55.3%.

## 10.3 Example

We measure the heights of ten people in the office and get a mean height of 172cm and a standard deviation of 15cm.

We need to decide the confidence interval we want. 95% is the most common

We need to know something called the Z value for that confidence interval.

Without getting too technical, the Z-score describes how far a value is from the mean in terms of standard deviation. A Z-score of zero would indicate that a value is identical to the mean value. A Z-score of 1 would indicate a distance of one standard deviation from the mean and a Z-score of 2 would indicate a distance of 2 standard deviations from the mean. It’s not important to remember these details however because the Z score for a confidence interval of 95% is 1.96.

For a 95% confidence interval the Z score is 1.96. The confidence interval is then given by:

$$CI = \pm Z \frac{\sigma}{\sqrt{n}},$$

where the Greek letter sigma is the standard deviation, n is the number of observations or measurements and Z is the Z-score.

The mean and its associated confidence interval is given by:

$$172 \pm 1.96 \frac{15}{\sqrt{10}},$$

$$172cm \pm 9.30cm.$$

In other words, the lower bound of the confidence interval is 162.7cm and the upper bound is 181.3cm. Our true mean is likely between these two values.

The confidence interval can be narrowed by either reducing the standard deviation in our measurements (not always possible but sometimes by improving measurement tools or techniques it can be reduced) or by increasing the number of measurements taken. With 100 measurements of height and the same mean and standard deviation the mean and its associated confidence interval would be stated as:

$$172 \pm 1.96 \frac{15}{\sqrt{100}},$$

$$172cm \pm 2.94cm.$$

This would make the range 169.1cm to 174.9cm.



# Chapter 11

## Hypothesis Testing & Statistical Significance

---

### 11.1 Hypothesis Testing

There are two types of hypothesis in experimental design, namely:

- Null hypothesis.
- Research hypothesis.

The null hypothesis states there is no difference between two variables (any observed differences occurred merely by chance).

The research hypothesis states that there is a true difference between the two variables in a study.

#### 11.1.1 Example

When considering a study to determine the impact of a drug on hypertension it would be typical to:

1. Formulate a null hypothesis: The drug has no effect.
2. Conduct a study and measure pre- and post- intervention blood pressure.
3. Calculate the probability ( $p$ ) of obtaining the observed difference in scores if the null hypothesis is true.

If the probability is small enough ( $p < 0.05$ ) then it suggests that the pattern of results is unlikely to have arisen by chance and the drug may have a genuine effect in the population.

## 11.2 Statistical Significance

Statistical significance refers to the probability ( $p$ ) of finding the pattern of results in a particular study, if the hypothesis is true.

If  $p$  is small ( $< 0.05$ ) then the effect is unlikely to have occurred by chance alone.

### 11.2.1 Odds against Chance Fallacy

The null hypothesis is a conditional probability and the obtained probability value is conditional on the null hypothesis being true. As it focuses on the null hypothesis it cannot be used to test the research hypothesis and the probability that the research hypothesis is correct is not known.

### 11.2.2 Statistical Significance Versus Importance

In spoken English, the word ‘significant’ is typically taken to mean ‘important’ but in statistics the word means ‘probably true’ (not due to chance).

A research finding might be true (significant) without being important or a research finding might be important without being true (significant).

This is important in medical research.

#### 11.2.2.1 Example

A study evaluated four asthma treatments: yoga, transcendental meditation, physiotherapy and a self-help group.

There was substantial improvement in patients taught on the yoga programme (lung function, medication use, well-being...etc) however, while the results were clinically significant, across-group comparison showed results as being not statistically significant.

### 11.2.3 In Brief

Do not confuse statistical significance with real world significance.

Just because an observed difference is statistically significant does not mean it is important.