HIGHER APPLCIATIONS OF MATHEMATICS

Unit 1: Statistics

Contents

[Introduction 3](#_Toc104458953)

[Probability 3](#_Toc104458954)

[Venn Diagrams 15](#_Toc104458955)

[Types of Data 22](#_Toc104458956)

[Numerical Representation 26](#_Toc104458957)

[Graphical Representations 28](#_Toc104458958)

[Histograms 32](#_Toc104458959)

[Normal Distribution 37](#_Toc104458960)

[Scatterplots and Correlation 43](#_Toc104458961)

[Calculating Correlation 45](#_Toc104458962)

[Linear Regression Analysis 51](#_Toc104458963)

[Removing Outliers 56](#_Toc104458964)

[Hypothesis Tests 60](#_Toc104458965)

[Types of Hypothesis Test 65](#_Toc104458966)

Acknowledgements

Material has been extracted from the following places. Thanks to the authors for their contribution.

Bryn Jones - Bryn.Jones@\_fe.gov.uk

Heriot Watt Scholar – Higher Applications of Maths Course

Mr Barton Maths

Montana Reid

Carluke High School

# Introduction

The field of **statistics** is concerned with collecting, analysing, interpreting, and presenting data.

As technology becomes more present in our daily lives, more data is being generated and collected now than ever before in human history.

Statistics is the field that can help us understand how to use this data to do the following things:

* Gain a better understanding of the world around us.
* Make decisions using data.
* Make predictions about the future using data.

Lifesaving drugs and vaccines for example cannot be used unless they have been through rigorous tests and analysis to make sure that they are safe enough for use and any side effects.

Jobs which being proficient in statistics include; Meteorologist, Market Analyst, Actuary, Data Scientist, Teacher/Government worker, Engineer, Police Officer, Political Analyst, Zoologist and many more.

It can be argued that Statistics is the most important part of the Higher Applications course based on the many jobs that is can relate to. It is likely you have encountered some statistics before in school. The big difference is about how you need to understand why different graphs are used, what the benefits are and analyse the information they tell us with regards to the context of a questions.

Another big difference will be the introduction of statistical software. You will be asked to input data into spreadsheets and then analyse them by creating graphs and perform statistical tests.

# Probability

**Trial –** This is the probability experiment you are looking at. E.g. rolling the dice, selecting a product from a larger group.

**Outcomes –** These are all the possible results of the probability experiment. If it is rolling a 6 sided dice there would be 6 possible outcomes. It can also be referred to as the **sample space**.

**Event –** This is the particular outcome you are interest in, it is also known as a **success**.

Calculating the probability is done as follows

Where r is the number of results you would class as a success (the event you are looking at) and n is the total number of outcomes.

Probabilities are generally written as a percentage or fraction.

When p = 1, this means

|  |
| --- |
|  |

When p = 0, this means

|  |
| --- |
|  |

Example: A farmer has a several fields of cows and sheep. There are 15 male cows, 12 female cows, 20 male sheep and 30 female sheep.

If the trial was to select an animal at random and check with animal and gender it was.

(a) How many difference outcomes are there?

|  |
| --- |
|  |

(b) Give an example of a possible outcome you could be looking for

|  |
| --- |
|  |

(c) What is the probability of an animal being picked at random being a Male Sheep?

|  |
| --- |
|  |

(d) What is the probability that a cow picked at random is a Female?

|  |
| --- |
|  |

Example 2: A coin is thrown three times and the number of heads is recorded.

(a) What is the sample space of this trial?

|  |
| --- |
|  |

(b) What is the probability of each of the outcomes?

|  |
| --- |
|  |

When looking at all the outcomes you will find that when you add up all the possible outcomes…

|  |
| --- |
|  |

**Exercise 1: Outcomes, Events, Sample Spaces and Basic Probability.**

1. Using a standard pack of card (52 cards, 13 of each suit). Looking at the suits and trying to get a diamond.

(a) What is the sample space?

(b) What is the event?

(c) What is the probability of a success?

(d) What is the probability of it not being a success?

2. In a bag of sweets there are 5 lollipops, 3 packets of crisps and 4 chocolate bars.

The trial is to pick one at random.

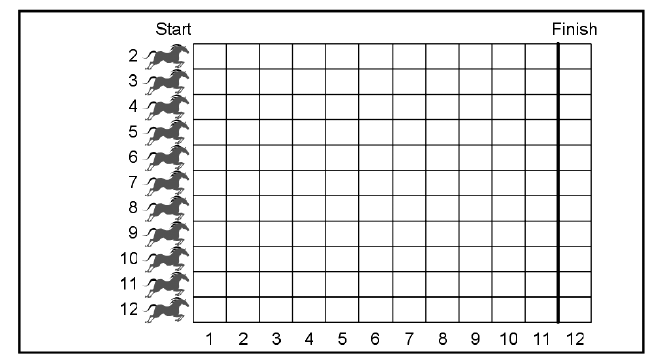
(a) What is the sample space?

(b) What is the probability of picking each sweet?

(c) Ryan picks a lollipop, Alice goes after him. How will the odds for her change?

3. What is the probability of rolling a number greater than 7 when you roll 2 dice?

4. There is a game that Maths Teachers play with S1s when they introduce probability called Dice Races.



The Game involves rolling two dice and moving the horse that corresponds to the number on both the dice added together (i.e. if 3 and 8 are rolled, 3+8 = 11, so horse 11 moves).

Can you explain why this game is unfair? Can you think of a way to make the game fairer?

5. What is the probability, if you flip a coin and roll a dice, that you get a heads on the coin and roll a prime number?

6. The table below shows the usual adult population of York in 2011. (Usual means only includes permanent residents so excludes uni students).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | | | | | | | | | |
| Sex | 16-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65-74 | 75-84 | 85+ | Total |
| All persons | 23 937 | 26 393 | 26 231 | 25 705 | 22 368 | 16 926 | 11 630 | 4 856 | 158 046 |
| Males | 12 039 | 13 321 | 13 004 | 12 680 | 10 901 | 7 842 | 4 927 | 1 639 | 76 353 |
| Females | 11 898 | 13 072 | 13 227 | 13 025 | 11 467 | 9 084 | 6 703 | 3 217 | 81 693 |

A person is chosen at random with is the probability they are

(a) Male

(b) Someone between ages of 25 and 34

(c) Someone over 64

(d) A female below the age of 35

**Types of Events**

**Independent –** Two events are independent if they do not affect each other. For example rolling a dice over and over are independent from each other. Rolling a 6 does not lessen the chance of getting a 6 on the next.

**Dependent –** This means it is affected by the previous event. For example removing cards from a deck. So if I draw the King of Diamonds, this lowers the chances of getting a King or a Diamond on the next card.

**Mutually Exclusive –** Two outcomes are mutually exclusive if they cannot occur at the same time. For example we cannot get the outcome of the flip of a coin to be heads and tails so they are mutually exclusive. Same with drawing cards if it was a draw a spade and a club, they cannot happen at the same time. However drawing an Eight and a Spade are **not** mutually exclusive because they can happen at the same time.

Notation: For the following you will see statements like P (Getting heads when flipping a coin), this just means the probability of the event of flipping a coin and it landing on heads.

**Exercise 2: Mutually Exclusive**

State whether the following are mutually exclusive Yes/No

1. P (The weather being nice) and P (your football team winning)

2. P (A shop having the shoes you want) and P (Shop not having the shoes you want)

3. P (Winning the lottery) and P (Losing your phone)

4. P (Rolling an even number on a dice) and P (Rolling a prime number on a dice)

5. P (Running your personal best in 100m) and P (Winning the race)

6. A clinic collects data about blood pressure of a random sample of adults. The results are shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Normal Blood Pressure | High Blood Pressure | Total |
| Smoking | 10 | 20 | 30 |
| Non smoker | 80 | 40 | 120 |
| Total | 90 | 60 | 150 |

Based on the information is high blood pressure independent from smoking? Show this by calculating the probabilities

P (Smoker with High Blood Pressure) and P (Non-smoker with High Blood Pressure)

7. A random survey is carried out about playing computer games and sleeping

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sleeps Well | Difficulty Sleeping | Total |
| Plays computer games often | 82 | 18 | 100 |
| Rarely or never plays computer games | 123 | 27 | 150 |
| Total | 205 | 45 | 250 |

Based on the information in this survey, is sleeping well independent of frequency of playing computer games.

**Expected Probability**

There is also something called expected probability. This is using the probability of something to predict how many times you expect something to happen. To use a very basic example.

**Example 1.** If you flipped a coin 40 times, how many times would you expect to get a tails?

|  |
| --- |
|  |

This is a very basic example that doesn’t show how important expected probability can be. A lot of businesses use expected probability in order to give a strong indication of how a company will do this year so they can adjust their finances accordingly.

In the example of the coin, if you were to flip a coin 40 times you would obviously not guarantee the result would be as predicted but it is about preparing for what we think might happen.

**Example 2:** Davy runs an ice cream truck. In the month of July there is a 0.05 probability that it will rain. When it rains Davy is unable to sell any ice cream.

(a) How many days in July should Davy expect it not to rain and therefore he can sell ice cream?

|  |
| --- |
|  |

(b) Davy needs to ensure he has enough ice cream to sell. On a day when it is not raining Davy will sell 200 ice cream cones.

Davy buys the ice cream mixture in packs that contain enough for 50 ice cream cones in each.

Each pack costs £6.50.

How much is Davy going to need to spend on packs to cover the month of July?

|  |
| --- |
|  |

(c) Since ice cream cones do not go out of date very often, Davy can bulk buy all the ice cream cones he needs for £50

Davy sells the ice cream for £0.90 each.

Calculate how much profit Davy can expect to make in the month of July?

|  |
| --- |
|  |

(d) This is Davy’s expected profit. Why might it be useful for Davy to calculate his expected profit rather than just see how much he earns?

|  |
| --- |
|  |

This question began with Davy using expected probability. It is a very basic example compared to what companies do in order to predict their own expected profit. These predictions can then be used to make a variety of decisions.

**Example 3:**

**Probability Trees and computing multiple probabilities**

When calculating the probability of an event happening you might want to draw a table out.

So when calculating the probability of flipping two coins and them both being heads you could do the following

|  |  |  |
| --- | --- | --- |
|  | Coin 1: Heads | Coin 1: Tails |
| Coin 2: Heads | HH = ¼ | TH = ¼ |
| Coin 2 Tails | HT = ¼ | TT =¼ |

From the table is easy to see that the chances of getting 2 heads is 1 in 4 (i.e. 0.25) because there are only 4 different outcomes. But what if we wanted to look at flipping more coins our using dice with many more sides?

**Calculating multiple probabilities –** All you need to do is calculate the probability of each event and multiply them together

**Example 1:** Show that the probability of getting two heads is ¼ or 0.25 and hence find the probability of getting a heads 3 times in a row.

|  |
| --- |
|  |

**Example 2:** The probability it will rain tomorrow is 0.55 and the probability that my bus will be late is 0.4. What is the probability that it will rain tomorrow and my bus being late?

|  |
| --- |
|  |

**Exercise: Multiple Probabilities**

1. Gary is playing cricket.

When attempting to catch the ball, the probability Gary is successful is 3/4. During the game Gary attempts two catches.

Find the probability Gary is successful with both catches.

1. Helen is taking part in a quiz on TV/

The probability she answers a question is

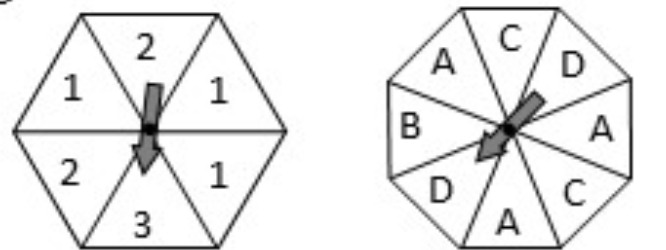
Helen is asked 3 questions.

Calculate the probability she answers both questions correctly.

1. A fair 6 sided dice is rolled 4 times.
2. Calculate the probability of getting a six all four times
3. Find the probability of getting no sixes.
4. A normal pack of cards has 52 cards in it, with 13 of each suit.
5. What is the probability of getting a club? (There are 13 clubs in the beginning)
6. What is the probability of getting a club 2 times in a row?
7. What is the probability of not getting a club when you draw 2 cards?
8. In a bag there are 5 blue marbles, 6 red marbles, 8 yellow marbles and 3 white marbles.

What are the odds of getting 2 blue marbles in a row when picking from the bag?

1. These two spinners are both spun



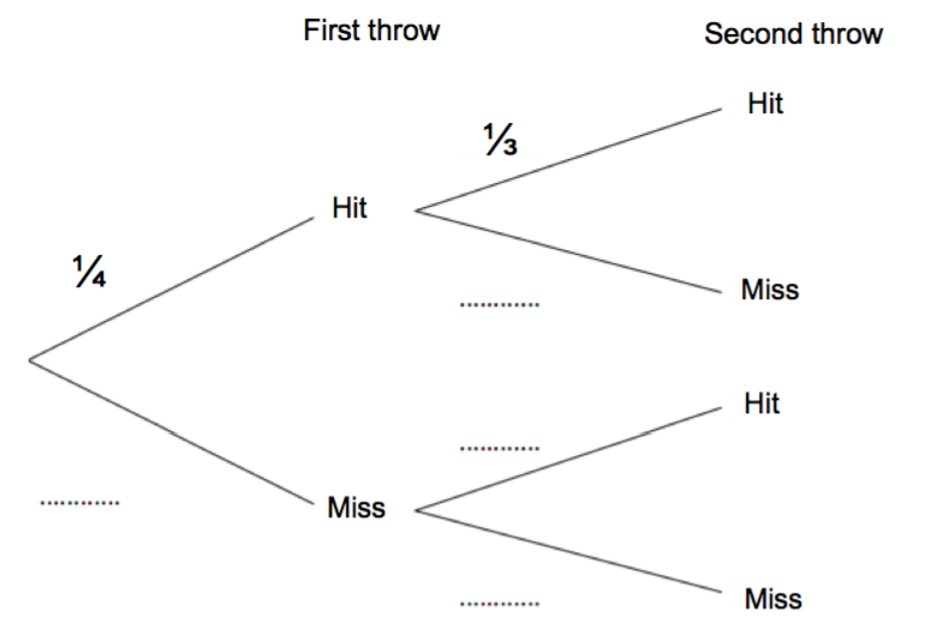
What are the odds of getting a 1 and an A when both are spun simultaneously?

**Probability Trees**

Probability trees are a visual representation of how probabilities can be set out.

**Example:** Jennifer was throwing darts at a bullseye. Her first throw she has a ¼ chance of hitting it, while her second throw has a 1/3 chance of hitting a bullseye.

Fill in the probability tree below accordingly.



State the probability of missing both times.

|  |
| --- |
|  |

State the probability that Jennifer only hits once for both her throws.

|  |
| --- |
|  |

**Example 2:** Draw a probability tree for taking a card from a pack 3 times and if the card **is a diamond** or is **not a diamond**. Remember the card is removed every time.

|  |
| --- |
|  |

**Example 3.** Tracy is always running late. Every day there is a chance that she will miss her train.

Draw a probability tree to show Tracy’s chances for 3 days.

|  |
| --- |
|  |

1. What is the probability Tracy is not late every day this week.

|  |
| --- |
|  |

1. What is the probability Tracy makes her train exactly 2 times?

|  |
| --- |
|  |

1. What is the probability at least once?

|  |
| --- |
|  |

**Exercise: Probability Trees**

1. A bag contains 4 red balls and 5 blue balls.

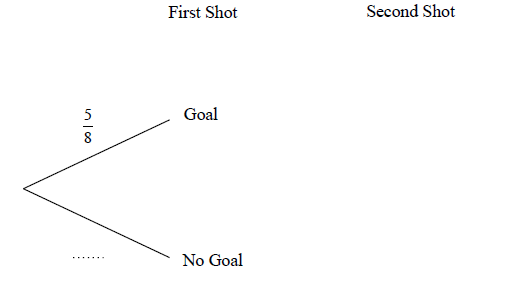
Raheem picks 2 balls at random.

Calculate the probability that he selects the same coloured ball each time, given that after each time a ball is selected, it is replaced.

Construct a probability tree to show this and calculate the probability that Raheem picks 2 red balls.

2. Sam is playing football where he plays as a forward. Every time he takes a shot he has a chance of scoring a goal. Sam take’s two shots.

(a) Complete the probability tree



(b) (i) What is the probability that Sam scores no goals

(ii) What is the probability he scores exactly 1 goal

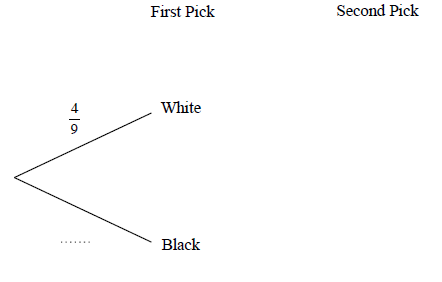
(iii) What is the probability he scores at least 1

3. A bag contains 4 white beads and 5 black beads.

Amy picks at random a bead from the bag and replaces it.

The beads are mixed and she then picks at random bead from the bag.

1. Complete the probability tree below



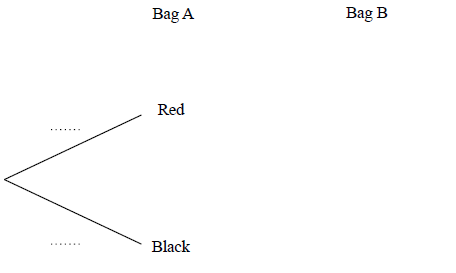
1. What is the probability that Amy picks
2. Two black beads
3. A black bead in his second draw
4. Beads of different colours

4. Bag A contains 10 marbles of which 2 are red and 8 are black.

Bag B contains 12 marbles of which 4 are red and 8 are black.

Imogen chooses a marble at random from each bag.

1. Complete the probability tree



1. Find the probability that Imogen chooses.
2. Two red marbles
3. Two black marbles
4. One black and one red marble

5. Hannah goes to an arcade

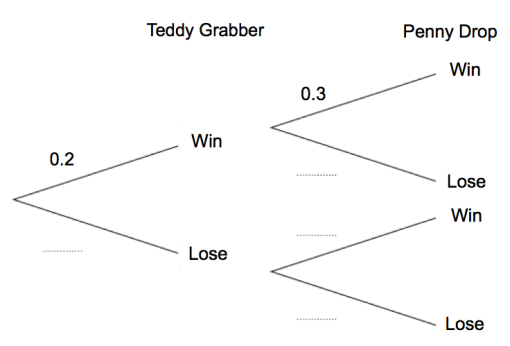
She goes on the Teddy Grabber once.

Then goes on the Penny Drop once.

The probability of winning the Teddy Grabber is 0.2.

The probability of winning the Penny Drop is 0.3.

1. Complete the tree diagram.



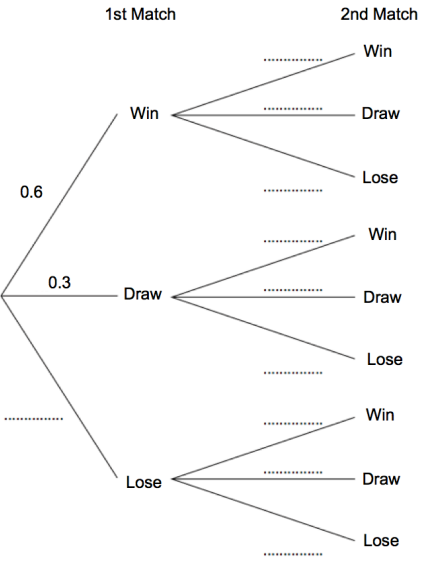
1. Work out the probability that Hannah wins on the Teddy Grabber and she also wins on the Penny Drop.
2. What is the probability that Hannah wins at least once?

6. A hockey team has two matches to play before the end of the season.

The probability that the team wins is 0.6

The probability that the team draws is 0.3

(a) Complete the tree diagram



(b) Work out the probability that the team will lose both matches?

A win gets the team 3 points and a draw is 1 point. A lose is 0 points.

The team needs at least 4 points to win the championship

(c) What is the probability that the team wins the championship?

7. Shown is a spinner.



The probability of a 1 is 2x

The probability of a 2 is x

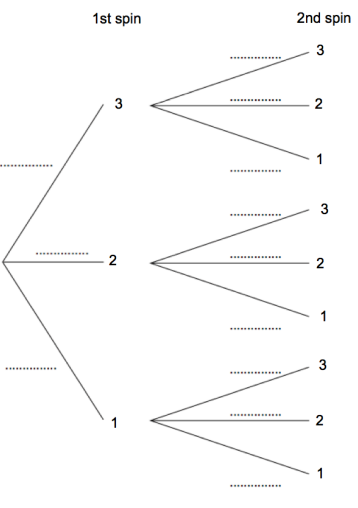
The probability of a 3 is 2x

(a) Calculate the value of x

The spinner is spun twice and the scores are added together.

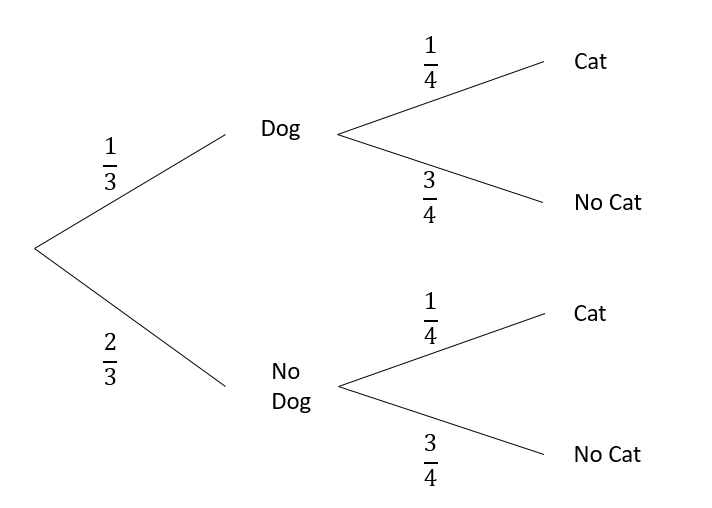
(b) Work out the probability of the final score being 4.

You may use the tree diagram to help you.



8. Owen asks 36 students at his school if they have at home:

* A dog
* A cat



(a) Owen claims that, for these students, “Having a dog” and “Having a cat” are independent from each other.

Explain why Owen is correct?

(b) In Owen’s school there are 1500 pupils.

(i) Estimate the number of students in Owen’s school who have neither a smart TV nor a dishwasher at home.

(ii) State what you have assumed in order to complete question (b) (i)

# Venn Diagrams

Venn Diagrams are another visual representation of data that can be used to calculate probabilities. Let’s look at a basic example.

We will look at the set of numbers from 1-12.

We will look at 2 categories

A: Even numbers

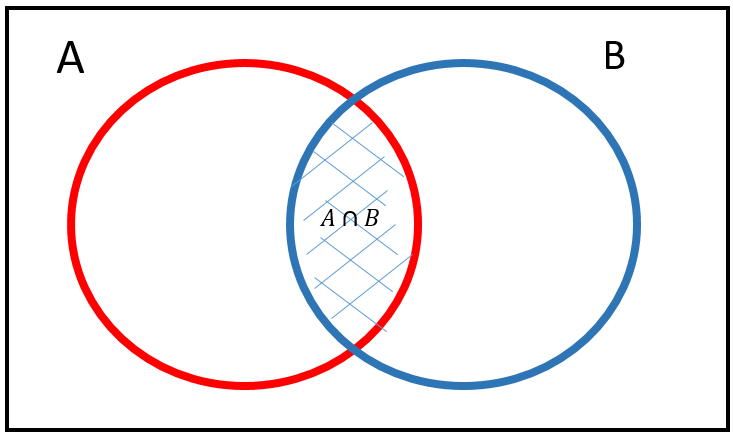
B: Multiples of 3

Produce the Venn Diagram below.

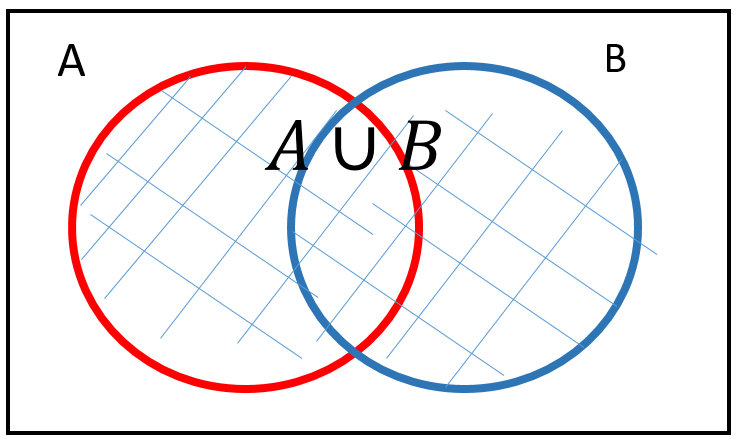
|  |
| --- |
|  |

There is some notation that we need to be familiar with

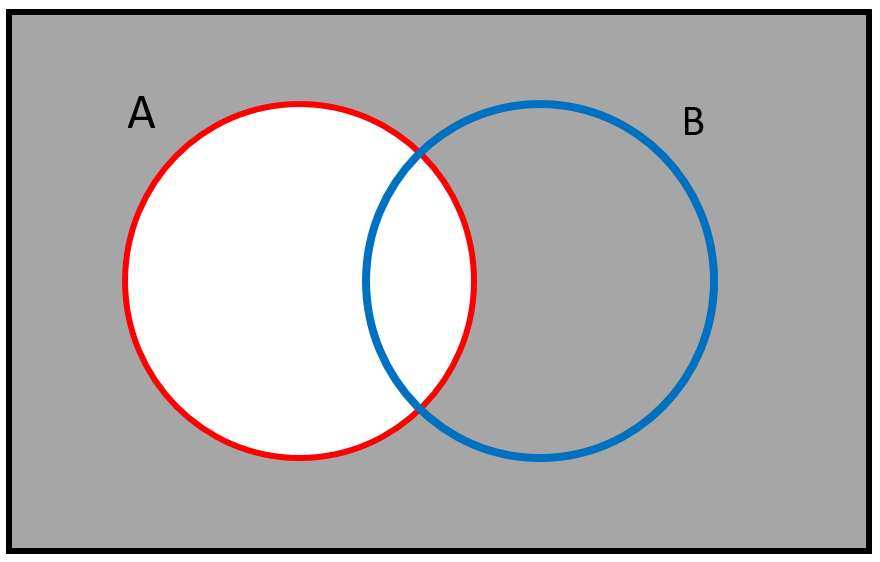
– This is the i**n**tersection of Events A and B. For the example above it refers to the numbers \_\_\_ and \_\_\_\_\_ because they are both even and multiples of 3.



– This is the union of Events A and B. So that at least one of them happens. So this includes \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ because they are even, a multiple of 3 or both.



The Complement of A or A’, this is the event that is not A. So A’ for the example above would be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.



– This means ‘what is the probability of A given that B is correct’. The formula is as follows

for the example above is the following

|  |
| --- |
|  |

**Example 2:** A Venn Diagram is drawn to show the different colours used in flags.

Diagram, venn diagram

Description automatically generated

(a) What is the probability that a flag chosen at random is blue?

|  |
| --- |
|  |

(b) Calculate

|  |
| --- |
|  |

(c) Calculate

|  |
| --- |
|  |

(d) Calculate the probability that the flag has no red in it.

|  |
| --- |
|  |

**Example 3:** A survey of 100 University students asked which newspapers they read.

40 read ‘The Times’

30 read ‘Daily Mail’

25 read ‘The Herald’

15 read ‘The Times’ and ‘Daily Mail’

12 read ‘The Times’ and ‘The Herald’

10 read ‘Daily Mail’ and ‘The Herald’

4 read all three

(a) Draw the Venn Diagram

|  |
| --- |
|  |

(b) How many read at least one newspaper?

|  |
| --- |
|  |

(c) What is the probability that a student chosen at random reads The Herald or The Times?

|  |
| --- |
|  |

(d) What is the probability that a student chosen at random reads the Daily Mail, given that that read The Times.

|  |
| --- |
|  |

(e) Calculate the percentage of students that read The Times and The Herald

|  |
| --- |
|  |

(f) Calculate the percentage of students that don’t read the daily mail

|  |
| --- |
|  |

**Exercise: Venn Diagrams**

1. Fill in the Venn diagram below using the set

S = {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20}

Diagram, venn diagram

Description automatically generated

A random number is selected. Calculate

(a) P(Prime)

(b) P(Even)

(c) P(Even ∩ Prime)

(d) P(Prime | Even)

2. An S2 class are put into groups.

Diagram, venn diagram

Description automatically generated

If one classmate is chosen at random, calculate

(a) P(S ∪ V)

(b) P(T’)

(c) P(S | T)

(d) What percentage of pupils are not in S or T?

(e) P((S ∩ T) ∪ V)

3. A casino records what 160 customers do at their casino.

60 Play blackjack

80 Play roulette

75 Play slot machines

20 play blackjack and roulette

77 play slot machines and roulette

12 play blackjack and slot machines

5 play all 3

(a) Complete the Venn Diagram below

Diagram, engineering drawing, venn diagram

Description automatically generated

(b) Estimate if 2000 people visit in a day what percentage play either at least one of Roulette or Blackjack

(c) Given that a customer played the slots, what is the probability they also played roulette?

4. 10000 Songs from the past 20 years are split into different genres but often the lines between each genre is unclear.

3000 songs were dance

500 were dance and rock

750 were pop and rock

Diagram, venn diagram

Description automatically generated

(a) Complete the Venn Diagram above.

(b) What is the probability a song selected at random is a Rock song.

(c) What is the probability a song selected at random is a Dance but not a Rock song.

(d) What is the probability a song chosen at random is a Pop song, given it is a Dance song.

(e) What percentage of songs were Rock and Pop songs

(f) What percentage of songs were not Rock, Pop or Dance?

5. A person may suffer from one or more of the linked medical conditions asthma, eczema and hay fever.

The Venn diagram shows the percentages of people in the UK who suffer from these conditions.

Diagram, venn diagram

Description automatically generated

(a) Calculate the probability that a person who suffers from eczema also suffers from hay fever

A new brand of indigestion medication is made.

The indigestion cannot be taken with hay fever medication.

(b) Use the Venn diagram to calculate the probability that a person in the UK chosen at random can take the indigestion medication

(c) Why might the actual probability be higher for someone being able to take the medication.

6. A gym runs two fitness classes, spinning and circuits.

On Saturday 100 people visited the gym

18 people attended the spinning class.

10 people attended both classes

56 people did not attend either class

(a) Write this as a Venn Diagram

Find the probability that a random person

(a) Attended only circuits

(b) Attended exactly 1 class

(c) Attended spinning given they attended circuits

7. A group of friends have been surveyed.

38% have been to Canada

80% have been to France

11% have been to neither Canada or France.

Find the percentage group that have been to Canada and France.

8. 100 pupils were asked if they like Maths, Science or Social Subjects. Everyone answered at least one.

56 said Maths

43 said Science

35 Social Studies

18 said Maths and Science

10 said Science and Social Studies

12 said Maths and Social Studies

6 said all 3.

(a) Create a Venn Diagram

(b) What percentage said Social Studies or Maths?

(c) What percentage said Science or Social Studies but didn’t say Maths

# Types of Data

Many companies and organisations collect information to improve their products or services. They can do this through online surveys, via telephone, market research, census, etc. Have you been asked to provide feedback, e.g. rate the quality of a WhatsApp call? Then you have taken part in a survey!

The information stored from these surveys is called data and it is used to make analysis and judgements about what to do next.

There are two types of data **numerical** and **categorical**.

**Numerical (quantitative) –**

|  |
| --- |
|  |

**Categorical (qualitative) –**

|  |
| --- |
|  |

Each of these can be classified in two different ways.

**Categorical** can be **Nominal** or **Ordinal.**

**Nominal**

|  |
| --- |
|  |

**Ordinal**

|  |
| --- |
|  |

**Numerical** can be **Discrete** or **Continuous**.

**Discrete**

|  |
| --- |
|  |

**Continuous**

|  |
| --- |
|  |

**Example:** The following data is regarding a house. Decide what kind of data it is.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Categorical | | Numerical | |
| House | Nominal | Ordinal | Discrete | Continuous |
| Type of house |  |  |  |  |
| Number of rooms |  |  |  |  |
| Height of House |  |  |  |  |
| Number of Windows |  |  |  |  |
| Colour of front door |  |  |  |  |
| What people think of the house (1 – very bad…. 5 very good) |  |  |  |  |

**Population, Sampling and Bias**

**Population –** The whole set of individuals, items or data from which a statistical sample is drawn. It does not just have to involve people.

**Sample** – A small section of the population.

So a School may want to improve the quality of its food at lunch so they would take a survey from the pupils. The **population** in the example is all the pupils. But they wouldn’t ask all the pupils they would ask a select number who are the **sample**.

**Trials and Sampling**

These are the different ways of choosing the sample from the population.

**Random Sample –** Pick randomly from the sample.

**Systematic Sample –** Set it up so only every 4th person is task for example

**Stratified Sample –** This is random but in ratio of group size. So if a gym has 400 members, 300 are men and 100 are women. If they were doing a survey they would take a larger sample from men than women.

**Cluster Sample –** Whole group chosen at random.

**Example:** A company runs 3 restaurants in Glasgow

Each restaurant has 60 employees.

The owner of the company wants to conduct face to face interviews with its employees.

The owner wants to use a cluster sample

(a) Is this a suitable sampling method to use?

|  |
| --- |
|  |

The owner then decides to try a different method that takes into account that of the 180 employees 72 are Male and 108 are female.

(b) What kind of sampling method should they use that takes this into account?

|  |
| --- |
|  |

(c) The owner requires a sample of 50 members. How many more Females than Males should be in the sample?

|  |
| --- |
|  |

**Example 2:** Look at the list of people below.

Mr Adams Miss Evans Mr Johnson Mr Rashid

Miss Booth Mrs Grant Miss Klein Miss Ryan

Mrs Carlisle Mr Graham Mr Khan Mr Thomson

Mr Carter Mr Henry Miss MacDonald Miss Violet

Miss Davidson Mr Hutton Mrs McCulloch Mr Zacharias

Give a way you could create a random sample using the names above.

|  |
| --- |
|  |

**Exercise: Data and Sampling**

1. What type of data is each of the following?

a) Number of texts sent

b) Colour of flowers

c) Weight of a dog

d) Exam Grade

e) Time asleep

f) Population of a country

g) Area of a floor

h) Ticket prices

I) favourite football team

j) Nationality of football players

h) Number of trees in a forest

I) weight of a whale

j) Breeds of bird

k) Number of eggs sold by a shop

2. For each of the following what kind of sample would you do if you were trying to survey the following?

Random, Cluster or Stratified

a) Glasgow city is going to go ask people on the high street if they should add another subway stop. (Think about if they should ask anyone)

b) Newton Mearns local council is trying to find out what people think about Wi-Fi in the area.

c) Hamilton FC are taking a survey on the toilet facilities at their football stadium (What is the issue if they did a random sample and only asked 10 randomly selected fans? 0

d) A cinema is going to do a survey on what kind of films their customers prefer to watch.

3. A dentist wants to ask a random sample of patients how satisfied they are from the service they receive from her. Describe how she could select patients at random and what problems might be involved with doing this? (Potential bias)

4. A gardener wants to study the distribution of plants in a random sample of 20 square meters of land from a rectangular site that is 50m wide and 40m long. How could he select a random sample?

5. A factory has the following employees

|  |  |  |
| --- | --- | --- |
|  | Full Time | Part Time |
| Male | 80 | 36 |
| Female | 24 | 44 |

If the boss wanted to do a random survey of 40 employees, how many of each subgroup should they select if they wanted to do a random sample?

6. A shop looks at its customers and how much they spend per visit

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| Under £20 | 400 | 850 |
| £20+ | 700 | 750 |

If the shop wanted to do a survey of 200 random customers, then how many of each sub group should they select?

7. The table gives the number of entrants to higher education in one year.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | 18 | 19 | 20 | Other | Total |
| Female | 94 775 | 42 640 | 11 950 | 33 980 | 183 345 |
| Male | 79 020 | 40 625 | 12 075 | 26 230 | 157 950 |
| Total | 173 795 | 83 265 | 24 025 | 60 210 | 341 295 |

a) A polling organisation wants to invite 10000 of the students to take part in a survey about their experiences of higher education. How many students should they use from each sub category?

b) List other characteristics that might be relevant.

c) Are there any issues with this poll?

8. A sample is a small survey. A census is a collection of the entire population.

An estate agency wants to know the following should they use a census or a sample.

a) The number of houses in England with an empty room.

b) How long an advert should appear on a website?

c) The number of people who offer less than the asking price of a house.

9. In each of the following situations state which kind of sample should be used.

(a) A large airline wants to know what passengers think of the service on their flights

(b) A newsworthy event has occurred near a town centre. A TV news crew wants to include some representative local comments in its report.

(c) A town council wants to know how the various groups living there view the quality of its nursery provision.

(d) A scientist wants to investigate any damage that an invasive species may be causing to wildlife on our rivers.

(e) A clothing manufacturer is considering developing a new range and wants to assess how popular it would be with both females and males across a range of ages.

(f) An election is coming up and the local branch of the political party want asses local opinion on a number of issues, they can only afford to interview 5% of the electorate.

10. We want to survey 300 people in the USE. Stratified sampling is chosen. In the table below,

Calculate how many people of each age group would need to be asked.

Table

Description automatically generated

11. ‘Do premiership footballers speak more than 1 language’ Think of 3 ways you could ask 20 different footballers how many languages they speak.

# Numerical Representation

From National 5 Applications of Mathematics you will have encountered numerical statistics before. The main numerical statistics you have encountered so far should be

Measures of average – mean, mode and median.

Measures of spread – standard deviation and SIQR.

**Recap:**

**Example 1:** Calculate the mean and median of the following test scores from Mr Williams’ class.

45 81 29 10 26 38 40 28 77 63 90

|  |
| --- |
|  |

A pupil was absent from the class. What would happen to the average if the person scored 52, would the average go up or down?

|  |
| --- |
|  |

Mrs Davis’s class has an mean score of 62. Make a comparison between the two classes.

|  |
| --- |
|  |

A picture containing text, watch

Description automatically generated

**Example 2.** Calculate the mean and standard deviation of the following

24 62 36 48 51 37

|  |
| --- |
|  |

**Example 3:** Calculate the median and SIQR of the following

19 38 23 20 36 45 50 20 24 37 54

|  |
| --- |
|  |

**Example 4:** Look at table below showing marks obtained in a test for 88 pupils.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Marks |  |  |  |  |  |
| Frequency | 6 | 16 | 24 | 25 | 17 |

Find an estimate for the mean and standard deviation

|  |
| --- |
|  |

**Exercise: Standard Deviation and Interquartile Ranges**

1. Alex records his times in minutes for a 10k race below.

44 49 50 57

(a) Calculate the mean and standard deviation.

(b) Katrina’s times had an average of 53 and a standard deviation of 2.3. Compare the times of Alex and Katrina.

2. Mr Thomas records the score of his class below

88 80 76 52 64 67 70

(a) Calculate the mean and standard deviation.

(b) Mr Thomas realises there is an issue with one of the questions. He needs to increase everyone’s score by 4. How will this affect the mean and standard deviation?

3. Freya looks at the price of her weekly shopping for the last few weeks

32 40 27 35 26 20 34 32

Calculate the median and SIQR

4. Graham records how long it takes for him to get to work every day.

11 14 13 12 38 18 12 10 12

(a) Calculate the mean and median

(b) Which is the better average?

(c) Calculate the IQR

5. Look at the interval table below where 40 results have been recorded.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Height () |  |  |  |  |  |
| Frequency |  |  |  |  |  |

(a) Which interval does the median fall between?

(b) Estimate the mean

The grouped frequency table shows the length of service in years employees who have been working for a company for at least ten years.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Length of service |  |  |  |  |  |  |
| Frequency | 30 | 42 | 23 | 13 | 8 | 4 |

Calculate an estimate of the standard deviation of the result.

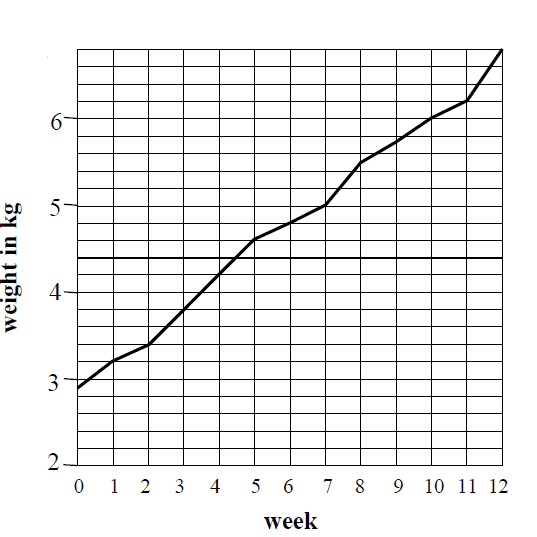
# Graphical Representations

You will have seen different graphs and charts before. You should be familiar with the creating and interpreting the following.

1. Bar Chart
2. Line Graph
3. Pie Chart
4. Stem and Leaf Diagram
5. Scatter graph and line of best fit
6. Boxplot

In the following series of lessons you will also be introduced to new graphs such as **Histograms**, along with new concepts such as **Correlation** and **Normality**.

**Example 1:** The following graph shows a baby’s weight throughout the year.



(a) What was the baby’s weight at birth?

|  |
| --- |
|  |

(b) What did the baby weigh after 5,9 and 12 weeks?

|  |
| --- |
|  |

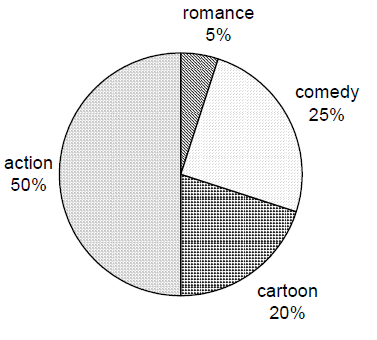
(c) How much weight did the baby put on between 3 and 7 weeks?

|  |
| --- |
|  |

(d) Between which two consecutive weeks was the greatest increase in weight?

|  |
| --- |
|  |

**Example 2:** The following pie chart was made to show pupil’s favourite genre of film.



(a)What fraction of pupils chose each type of film?

|  |
| --- |
|  |

(b) If 90 pupils said cartoons, how many pupils were asked in total?

|  |
| --- |
|  |

For Higher Applications of Mathematics Excel and R Studio will primarily be used to create these charts and graphs however the important is being able to decide which type of graph is the most suitable in order to make conclusions on questions.

**Example 3:** A poll is taken of different votes in an area.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Blue | Red | Green | Yellow | Orange | Purple |
| 3200 | 5000 | 2000 | 5800 | 1800 | 300 |

Input the table into Excel and decide what kind of charts are suitable and which are not.

State which kind of table you found suitable.

|  |
| --- |
|  |

**Example 4**. The following shows the average attendance of a cinema in Glasgow throughout the first 12 weeks of the year which the number indicating the percentage of tickets sold in the median showing.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 56 | 54 | 40 | 79 | 73 | 63 | 55 | 88 | 62 | 60 | 51 | 83 |

(a) Using Excel decide what type of graph is suitable, state which graph you have chosen and why?

|  |
| --- |
|  |

(b) What does the information tell you about the attendance of cinema’s why do you think that is?

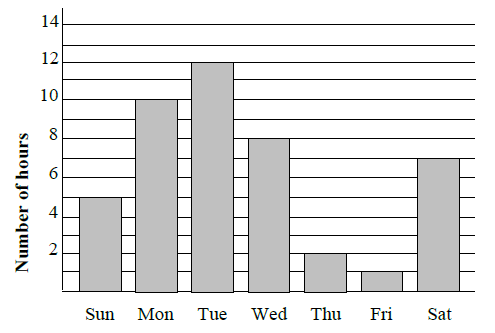
|  |
| --- |
|  |

(c) What kind of graph would be suitable to show the spread of the information?

|  |
| --- |
|  |

**Exercise: Drawing and Interpreting Graphs**

1. The following bar chart shows the number of hours of sunshine for a week in April.



(a) Which day was the sunniest?

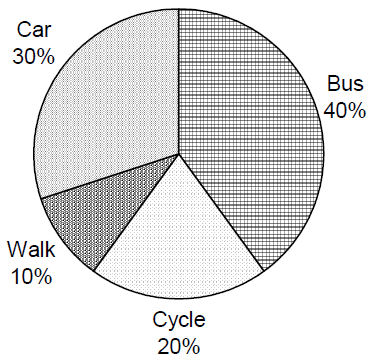
(b) Which day had 8 hours of sunshine?

(c) What was the total number of hours of sunshine over the weekend(Sat & Sun)

(d) Which of the following would not be suitable?

Pie Chart, Line Graph, Box Plot, Stem and Leaf Diagram.

2. Students were asked how they got to school. Their answers were made into a pie chart.

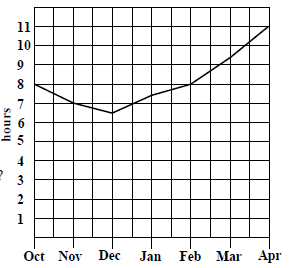


(a) What fraction said each of the answers?

(b) What was the least popular method?

(c) It turns out that the information for Train wasn’t included. 90 pupils said train. Given that 30 pupils said Walk, what should the percentages of all the methods of travel be?

3. The line graph shows the average daily hours of sunshine in a holiday resort in low season.



(a) Which month has the least hours of sunshine?

(b) What is the average daily hours of sunshine in April and December?

(c) How many more hours of sunshine are there in March than November?

4. The table below shows the destination of a class of pupils going on holiday.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Country | Scotland | England | Spain | France | Italy | USA |
| Number of pupils | 3 | 5 | 12 | 4 | 2 | 4 |

Create a graph that makes it easy to compare between the choices.

5. The table shows a patient’s temperature in oC, taken a 2 hourly intervals for a 24 hours period.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Time | 0000 | 0200 | 0400 | 0600 | 0800 | 1000 | 1200 | 1400 | 1600 | 1800 | 2000 | 2200 |
| Temp | 38.0 | 38.2 | 37.8 | 37.8 | 37.5 | 37.4 | 37.4 | 37.6 | 36.8 | 37.0 | 37.1 | 37.0 |

Create a graph that shows how the temperature changes with time.

*Extension: You can input the data into an excel spreadsheet (remember to use a csv file) or type the data straight into R Studio. In R studio complete questions 4 onwards).*

# Histograms

A histogram is a type of bar chart used for **continuous** data. The data must first be grouped, and do not all need to be the same size.

A histogram features a title, labelled axes and bars of (potentially) variable width that touch, e.g.

Chart, histogram

Description automatically generated

**Example 1:** Draw a histogram of the following.

|  |  |  |  |
| --- | --- | --- | --- |
| Test Score | Frequency | Interval | Frequency Density |
|  | 15 |  |  |
|  | 22 |  |  |
|  | 28 |  |  |
|  | 30 |  |  |
|  | 9 |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

The shape of the histogram illustrates the distribution (the shape of the data).

Data can be skewed, meaning it tends to have a long tail on one side and not the other.

We are going to look at each of these illustrations in terms of averages (mean, median) and spread (standard deviation and interquartile range).

**Normally Distributed**

**Chart, histogram

Description automatically generated**

Normally distributed is the most important kind and the type that we will look and test for most frequently. Most of the results must sit in the middle of the graph with the least results at either side. When the line is draw it referred to as bell shaped.

**Skewed to the Right**

Chart, histogram

Description automatically generated

(Note how it goes the opposite way of what you would expect)

This is because the numbers on the right actually skew the mean and median to a higher number. This is why it is also called positively skewed.

**Skewed to the left**

Chart, histogram

Description automatically generated

This is called skewed to the left because the data on the lower end of the graph will make the mean and median appear lower, hence it is also called negatively skewed.

**Exercise:** Draw the following histograms and decide if they are normal, skewed to the right or skewed to the left?

1.

Table

Description automatically generated

2.

|  |  |
| --- | --- |
| Scores | Frequency |
|  | 4 |
|  | 8 |
|  | 10 |
|  | 10 |
|  | 15 |

3.

|  |  |
| --- | --- |
| Attendance | Frequency |
|  | 12 |
|  | 40 |
|  | 90 |
|  | 96 |
|  | 65 |
|  | 36 |
|  | 60 |

4.

Table

Description automatically generated

5. Sam asks some students how long they took to finish their science homework. The table and histogram show some of this information.

Table

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Complete the information in the table and histogram.

6. The table and histogram below give some information about how far some teachers travel to school.

Table

Description automatically generated

Chart, histogram

Description automatically generated

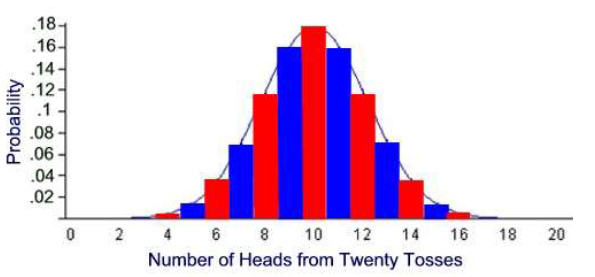
(a) Use the histogram to complete the table   
(b) Use the table to complete the histogram

# Normal Distribution

The Normal Distribution is a very frequently occurring continuous variable distribution.

In models, for example, the heights of people, the weights of similar animals, and measurements on machine produced items are all normally distributed.

For example, if a coin is thrown 20 times the most likely number of Heads is 10, followed by 9or 11 Heads, then 8 or 12 Heads and so on. Getting something like 5 or 15 Heads is relatively unlikely. The full bar chart for throwing a coin 20 times is shown below, and is overlaid with the *normal curve*.



The distribution is symmetrical about the mean, which is also the median and the mode.

* The curve is symmetrical about the mean
* 50% of values are below the mean
* 50% of values are above the mean
* 68% of values fall within 1 standard deviation
* 95% of values fall within 2 standard deviations
* 99.5% of values fall within 3 standard deviations

Chart, line chart, histogram

Description automatically generated

**Example:** The height of women in the UK follows a normal distribution. The mean is 161cm and the standard deviation is 6cm.

Complete the graph below showing the heights for 68%, 95% and 99.5% of the data.

|  |
| --- |
|  |

**Example:** The life of a food mixer is normally distributed with mean 90 months and standard deviation of 15 months.

(a) What proportion of mixers last less than 90 months

(b) What proportion of mixers last between 75 and 105 months

(c) What proportion of mixers last less than 60 months?

|  |
| --- |
|  |

**Exercise:** Normal Distribution

Sketch a normal curve for each distribution. Label the x-axis at one, two, and three standard deviations from the mean.

1. Mean = 95; standard deviation = 12

2. Mean = 100; standard deviation = 15

3. Mean = 60; standard deviation = 6

4. Mean = 23.8; standard deviation = 5.2

5. A set of data has a normal distribution with a mean of 5.1 and a standard deviation of 0.9.

(a) Sketch a normal curve for the distribution.

Find the percent of data within each interval.

(b) Between 6.0 and 6.9

(c) Greater than 6.9

(d) Between 4.2 and 6.0

(e) Less than 4.2

6. Test scores are normally distributed with a mean of 76 and a standard deviation of 10.

a. In a group of 230 tests, how many students score above 96?

b. In a group of 230 tests, how many students score below 66?

c. In a group of 230 tests, how many students score within one standard deviation of the mean?

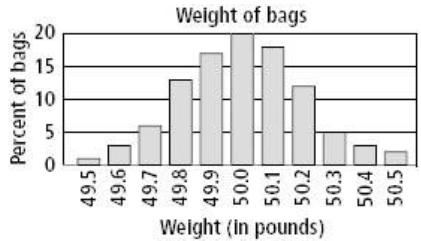
7. The number of nails of a given length is normally distributed with a mean length of 5.00 in. and a standard deviation of 0.03 in.

a. Find the number of nails in a bag of 120 that are less than 4.94 in. long.

b. Find the number of nails in a bag of 120 that are between 4.97 and 5.03 in. long.

c. Find the number of nails in a bag of 120 that are over 5.03 in. long.

8. The actual weights of bags of pet food are normally distributed.



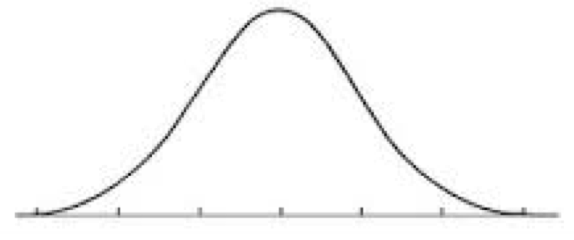
The mean of the weights is 50.0 lb, with a standard deviation of 0.2 lb. Use the graph for a – c.

a. About what percent of bags of pet food weigh less than 49.8 lb?

b. In a group of 250 bags, how many would you expect to weigh more than 50.4 lb?

c. In a group of 50 bags, how many would you expect to be within 1.5 standard deviations of the mean?

9. A machine is used to put bolts into boxes. It does so such that the actual number of bolts in a box is normally distributed with a mean of 106 and a standard deviation of 2.



a) Draw and label the Normal curve from the information.

b) What percentage of boxes contain more than 104 bolts?

c) What percentage of boxes contain more than 110 bolts?

d) What percentage of boxes contain less than 108 bolts?

e) What percentage of boxes contain less than 100 bolts?

f) What percentage of boxes contain between 102 and 112 bolts?

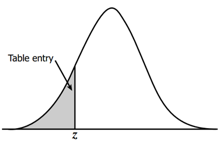
g) What percentage of boxes contain between 100 and 106 bolts?

**Normal Distribution Z-Table**

Normal Distribution tables use the standard normal distribution with a mean of 0 and a standard deviation of 1.

The table shows the values of z which measures the distance from the mean in standard deviations.

The shaded area, A, gives the probability of data that is less than the given value z.



*For the following you will need the use of a z-table*

**Example:** Calculate P(Z<0.24)

This is the percentage of data that is below 0.24 standard deviations away from the mean.

|  |
| --- |
|  |

**Example:** Calculate P(Z<-1.3)

|  |
| --- |
|  |

**Example:** Calculate P(Z>0.75)

|  |
| --- |
|  |

**Example:** Calculate P(Z > -2.34)

|  |
| --- |
|  |

**Exercise: Z Tables**

1. Find the following probabilities as a percentage.

(a) P(Z < -0.01)

(b) P(Z < -0.13)

(c) P(Z < – 0.54)

(d) P(Z< -0.72)

(e) P(Z< -0.83)

(f) P(Z < -1.02)

2. Find the following probabilities as a percentage.

(a) P(Z > 0.03)

(b) P(Z > 0.10)

(c) P(Z > 0.22)

(d) P(Z > 0.33)

(e) P(Z > 0.52)

(f) P(Z > 0.94)

**Z Table Exam Questions.**

The formula to calculate the Z number is as follows

**Example:** Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council’s GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the GMAT?

|  |
| --- |
|  |

(b) How high must an individual score on the GMAT in order to score in the highest 5%?

|  |
| --- |
|  |

**Example:** The average number of acres burned by forest and range fires in a large New Mexico county is 4,300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2,500 and 4,200 acres will be burned in any given year?

|  |
| --- |
|  |

**Exercise: Z Table Exam Questions**

1. The weights of bags of gravel may be modelled by normal distribution with mean of 25.8kg and standard deviation 0.5kg.

Determine the percentage of a bags that weight

(a) Less than 25kg

(b) Between 25.5kg and 26.5kg.

(c) Determine to two decimal place the weight exceeded by 75% of bags.

2. The time taken for a company to install a satellite dish is normally distributed, with mean 134 minutes and standard deviation 16.

(a) Determine P(X < 150)

(b) Determine to one decimal place the time exceeded by 10% of the installations

3. The length of eels in a river is assumed to be normally distributed with mean 48 and standard deviation 8.

Determine the probability that the length of an eel selected at random is;

(a) Less than 60cm.

(b) Within 5% of the mean length.

# Scatterplots and Correlation

Scatter plots are used to investigate the relationship between two variables. Corresponding values are plotted on a graph in the same way as coordinates are plotted. When changes in one variable match changes in the other, there may be a cause-and-effect relationship between the two. This is called correlation (connection).

Correlation can be described in three ways;

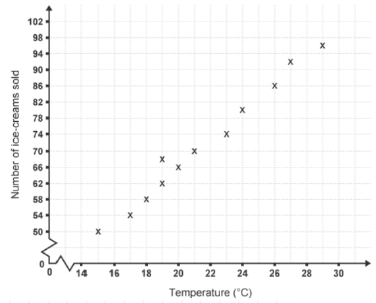
● Positive correlation

● Negative correlation

● No correlation

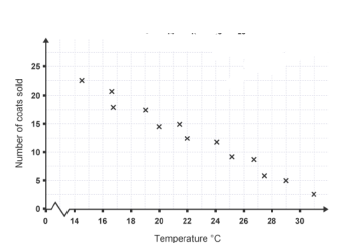
**Positive Correlation**

If both variables are increasing you get a positive correlation. For example, you would expect that as the temperature increases, the number of ice creams sold also increases, so there is a positive correlation between height and weight.



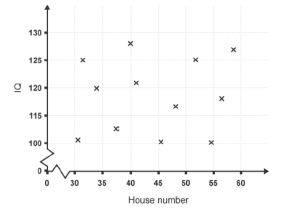
**Negative Correlation**

If one variable increases as the other decreases, you get a negative correlation. For example, you would expect that, as the temperature increases, the number of coats sold decreases, so there is a negative correlation between temperature and sales of coats.



**No Correlation**

If there is no relationship between the variables, there is no correlation. For example, there is no correlation between a person’s house number and their IQ.



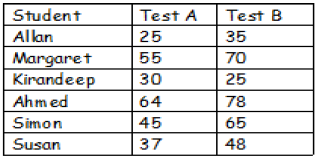
When asked to draw “a line of best fit” you should:-

● Go through as many points as possible

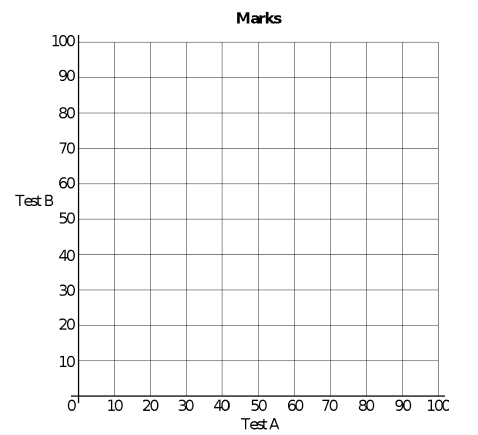
● Split the group so there are roughly as many points above the line as there are below it.

We can also use the line to make further predictions or estimates.

**Example:** A class of students sat two tests in the same subject. Here are the marks for six of the students who sat the tests:



(a) Draw a scatter graph of these marks on the grid below.



(b) Draw a best fitting straight line for this scatter graph.

(c) Using your line of best fit, if a student scored 40 in test A, give them a likely prediction for test B.

|  |
| --- |
|  |

**Exercise: N5 Recap – Scatter Plots**

1. The height of a plant is measured over five days as shown below.

Table

Description automatically generated with medium confidence

(a) Plot the points and draw the best fitting straight line through them.

(b) Work out the equation of the line.

(c) Use your line to estimate the height after 1.5 days.

2. The result below show the length of a spring when a force is applied.

Table

Description automatically generated

(a) Plot the points and draw the best fitting straight line through them.

(b) Find the equation of the line.

(c) Use your graph to estimate the length when a force of 4.5 is applied.

# Calculating Correlation

**Regression Analysis** - Looking at the relationship independent variables have with one or more other variables.

Correlations are useful for describing simple relationships among data. For example, imagine that you are looking at a dataset of campsites in a mountain park.

You want to know whether there is a relationship between the elevation of the campsite (how high up the mountain it is), and the average high temperature in the summer.

Correlation is a measure of the degree of **linear association** between two *numerical* variables

•the correlation is said to be *positive* if ‘large’ values of both variables occur together and *negative* if ‘large’ values of one variable tend to occur with ‘small’ values of the other

•the range of possible values is from -1 to +1

•the correlation is high if observations lie close to a straight line (i.e. values close to +1 or -1) and low if observations are widely scattered (correlation value close to 0)

•it does not indicate a causal effect between the variables

For the following state something that would have a correlation to the following.

(a) As the temperature increases.

(b) Height of people.

(c) Money spent by a premier league football team.

(d) How much time someone spends studying.

**Types of Correlation**

|  |
| --- |
|  |

Chart, bubble chart

Description automatically generated

|  |
| --- |
|  |

Chart, bubble chart

Description automatically generated

|  |
| --- |
|  |

Chart, scatter chart, bubble chart

Description automatically generated

|  |
| --- |
|  |

Chart, bubble chart

Description automatically generated

|  |
| --- |
|  |

Chart, scatter chart, bubble chart

Description automatically generated

**Calculating Correlation (Formula)**

Correlation can be measured using Pearson's coefficient *r*. (A number from -1 to 1)

A picture containing text, clock, watch

Description automatically generated

**NOTE**: If the linear correlation coefficient is close to 0 it does not always mean the variables have no correlation; it could be that they have a more complicated relationship than a line.

**Example:** Calculate and interpret the correlation coefficient of the two variables below.

|  |  |  |
| --- | --- | --- |
| Person | Hand | Height |
| A | 17 | 150 |
| B | 15 | 154 |
| C | 19 | 169 |
| D | 17 | 172 |
| E | 21 | 175 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | Person | Hand | Height |  |  |  |  |  | | A | 17 | 150 |  |  |  |  |  | | B | 15 | 154 |  |  |  |  |  | | C | 19 | 169 |  |  |  |  |  | | D | 17 | 172 |  |  |  |  |  | | E | 21 | 175 |  |  |  |  |  | |  |  |  |  |  |  |  |  | |

**Exercise: Calculating Coefficient (Formula)**

1. Calculate the Pearson’s correlation coefficient and interpret what that means.

|  |  |  |
| --- | --- | --- |
| Person | Weight | Blood Pressure |
| A | 150 | 125 |
| B | 170 | 130 |
| C | 175 | 160 |
| D | 180 | 170 |
| E | 200 | 150 |

2. Calculate the Pearson’s correlation coefficient and interpret what that means

|  |  |  |
| --- | --- | --- |
| Person | IQ | Height |
| A | 122 | 170 |
| B | 94 | 196 |
| C | 100 | 182 |
| D | 114 | 175 |
| E | 120 | 187 |

3. From the following scatter plot, the relationship can be described as

Chart, scatter chart

Description automatically generated

(a) Strong positive

(b) Strong negative

(c) Perfect negative

(d) Perfect positive

4. Which of the following values cannot represent a correlation coefficient?

(a) r = 1.08

(b) r = 0.95

(c) r = 0

(d) r = - 1.0

5. The range of the correlation coefficient (r) is …….

(a) -1 < r < 1

(b) 0 ≤ r ≤ 1

(c) -1 ≤ r ≤ 0

(d) -1 ≤ r ≤ 1

6. Which of the following is the most appropriate interpretation of a product-moment correlation coefficient of 0.8?

(a) a strong negative linear correlation

(b) a moderate negative linear correlation

(c) a moderate positive linear correlation

(d) a strong positive linear correlation

(e) no correlation

7. What is the most likely value of the product-moment correlation coefficient for the data shown in the diagram?

Chart, scatter chart

Description automatically generated

(a) -0.24

(b) 0

(c) 0.37

(d) -0.89

(e) 0.05

8. If all points on a scatter diagram lie directly on a straight line of positive slope, what is the value of the product-moment correlation coefficient for this data set?

9. Input the following into an excel spreadsheet and save as a csv file.

|  |  |
| --- | --- |
| Long Jump | High Jump |
| 5.51 | 1.65 |
| 5.72 | 1.77 |
| 5.81 | 1.83 |
| 5.88 | 1.77 |
| 5.91 | 1.75 |
| 6.05 | 1.77 |
| 6.08 | 1.82 |

Load the dataset into *R Studio*

(a) Create a scatter plot of the data and write what you think the correlation will be.

(b) Calculate the correlation using *cor.test(X,Y)*.

**10.** Shown in the table opposite is the number of ice cream sales from one shop over the last 12 days.

Table

Description automatically generated

(a) Type into an excel spreadsheet and save as a csv file.

(b) Import the data into R studio.

(c) Create a scatter plot.

(d) Calculate the Pearson’s correlation coefficient.

(e) What does this result tell us?

11. Input the following data into R Studio and calculate the simple correlation coefficient between wing length & tail length of the following 12 birds of a particular species.

Table, calendar

Description automatically generated

# Linear Regression Analysis

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line. In previous learning, you have placed the line of best fit by eye. Each student would have produced different best fitting lines, therefore different results would have been given for predictions. For better accuracy, there is a more mathematical way to establish this line.

We will use computer software to calculate this line of best fit; the Least Squares Regression Line.

The general equation of the line is 𝑦=𝑎+𝑏𝑥 where 𝑏 is the slope of the line and 𝑎 is the y-intercept (where the line crosses the *y*-axis.

The equation of the Least Squares Regression line can be produced in a calculator but for the purpose of this work, we will use computer software. i.e. EXCEL.

**Example 1:** The following table shows the average temperature in July every year.

|  |  |
| --- | --- |
| Year | Average Temp |
| 1990 | 20.8 |
| 1992 | 20.9 |
| 1994 | 20.9 |
| 1996 | 21.1 |
| 1998 | 21.2 |
| 2000 | 21.3 |
| 2002 | 21.3 |
| 2004 | 21.5 |
| 2006 | 21.9 |
| 2008 | 21.8 |
| 2010 | 22.4 |

\*enter the years into excel as 0, 2 etc…

(a) Type the data into Excel and save as a csv file.

(b) Import the data into R Studio.

(c) Draw a Scatter Plot in R Studio

(d) Create a linear model and write the equation in the form y = mx + c where x is the year and y is the average temp. Write it below

|  |
| --- |
|  |

(e) Use your linear model to predict what average temp will be in 2018.

**Example 2.** The table below gives the amount of Krabby Patties made by Spongebob for each year he’s worked. Graph the data on a scatter plot, find the line of best fit, and write the equation for the line you draw.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Years worked** | 1 | 2 | 3 | 4 | 5 | 6 |
| **Patties made** | 6,500 | 7,805 | 10,835 | 11,230 | 15,870 | 16,387 |

(a) Using statistical software write the linear regression equation.

|  |
| --- |
|  |

(b) Calculate the Correlation Coefficient (r).

|  |
| --- |
|  |

Coefficient of Determination

The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.

With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.

An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.

An R2 of 1 means the dependent variable can be predicted without error from the independent variable.

An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R2 of 0.20 means that 20 percent is predictable; and so on.

(c) Calculate the coefficient of determination.

|  |
| --- |
|  |

(d) Using the linear regression equation predict how many Krabby Patties he will make after working 10 years.

|  |
| --- |
|  |

**Exercise: Linear Regression Analysis**

1. The table below gives the amount of time students in a class studied for a test and their test scores. Graph the data on a scatter plot, find the line of best fit, and write the equation for the line you draw.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hours Studied** | 1 | 0 | 3 | 1.5 | 2.75 | 1 | 0.5 | 2 |
| **Test Score** | 78 | 75 | 90 | 89 | 97 | 85 | 81 | 80 |

(a) Using statistical software to create a linear model, compute the linear regression equation.

(b) Compute the correlation coefficient.

(c) Compute the coefficient of determination.

(d) Using the linear regression equation predict a student’s test score if they studied for 4 hours.

2. The table below gives the estimated world population (in billions) for various years.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | 1980 | 1990 | 1997 | 2000 | 2005 | 2011 |
| **Population** | 4400 | 5100 | 5852 | 6080 | 6450 | 7000 |

(a) Using statistical software to create a linear model, compute the linear regression equation.

(b) Compute the correlation coefficient.

(c) Compute the coefficient of determination.

(d) Using the linear regression equation predict the world population in the year 2015.

3. The table below shows the income for an employee over his first 8 years of work. Use this to estimate his income for his 15th year of work.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Years** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| **Income** | 45,000 | 46,814 | 48,212 | 52,870 | 54,125 | 58,532 | 61,075 | 62,785 |

(a) Using statistical software to create a linear model, compute the linear regression equation.

(b) Compute the correlation coefficient.

(c) Compute the coefficient of determination.

(d) Using the linear regression equation predict his income for his 15th year of work.

4. For a piece of coursework, Tom investigates cars passing their school from 3pm to 4pm.

Tom records the make and model of each in his sample.

For each car, he found the figures for the fuel economy in miles per gallon (F) and the CO2 emissions in grams per kilometre (C).

He plotted F against C on this scatter diagram.

*Chart

Description automatically generated*

(a) Calculate the equation of the regression line of F on C. (Using statistical software)

(b) Draw the regression line.

(c) Calculate the correlation.

(d) You have to pay vehicle tax if you own a car.

The vehicle tax depends on CO2 emission of your car as shown in the table.

Table

Description automatically generated

You want to buy a car for which tax is not more than £30 per year.

Using your graph or the equation of the regression line of F on C, estimate the minimum fuel economy of a car you might buy.

5. Jamie and Lily are investigating different types of correlation

(a) Match each scatter diagram below to the most appropriate type of correlation.

Diagram

Description automatically generated

Jamir and Lily each wear a special bank that measures the number of steps walked in each day (S)

The number of calories burned each day (C).

The tables below show Jamir’s data and Lily’s data for the last eight days.

Table

Description automatically generated

(b) Jamir and Lily want to know if it is justified to use S to estimate C. Calculate the correlation coefficient between S and C for Jamir’s data.

(c) Calculate the product moment for Lily’s data. Hence explain why Jamir’s estimate of X is likely to be more accurate than Lily’s estimate of for any given value S.

(d) Complete the scatter diagram of C against S for Jamir’s data on the grid below. The table with Jamir’s data is repeated below.

Chart, scatter chart

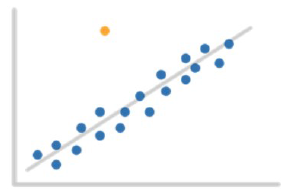
Description automatically generated

(d) Calculate the equation of linear regression of C on S for Jamir’s data.

(e) Jamir wants to burn at least 21000 each week. Work out how many steps he should aim to do each day.

# Removing Outliers

An outlier is a data value that is out of keeping with the other values. This could either be caused by a measurement or recording error, e.g. recording a long jump distance as 67.2 metres instead of 6.72 metres, or a genuine freak result, e.g. a long jump distance of 8.95 metres which stood as the World Record for 30 years.



Informally, outliers can be determined by eye. More formally, a statistical test can determine if a data value is an outlier. It is important to identify outliers, and if it is appropriate, remove them from the data, as they can affect any conclusions drawn.

**Example:** The following shows the height and weight of dogs in a kennel.

|  |  |
| --- | --- |
| Height | Weight |
| 53 | 2.5 |
| 55 | 2.7 |
| 56 | 3.9 |
| 58 | 2.8 |
| 59 | 3 |
| 62 | 2.2 |
| 65 | 3.3 |
| 66 | 3.9 |
| 67 | 8.1 |
| 70 | 4 |
| 75 | 4.3 |

(a) Type the information into excel.

(b) Create another table with a **Before** column and calculate the following;

Mean height, mean weight, standard deviation of height and weight, correlation and coefficient of determination.

(c) Create a scatter plot to show the information and remove any outliers.

(d) Repeat part b and create a column in the table labelled **After**.

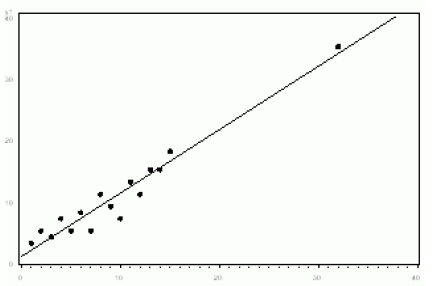
(e) What numerical statistics do you see change and what does this tell us?

|  |
| --- |
|  |

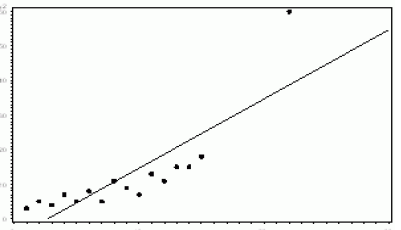
**To drop or not to drop?**

1. If it is obvious that the outlier is due to incorrectly entered or measured data you should drop the outlier, e.g. if you are analysing the weight of a sample of women and the answer is 19 lbs, then you know this is physically impossible and may be a mistype.

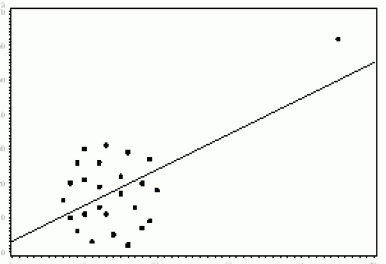
2. If the outlier does not change the results but does affect some of the assumptions then you may drop the outlier, e.g. neither the presence of absence of this outlier would change the outcome.



3. More commonly, the outlier affects both results and assumptions. In this situation, it is *not* legitimate to drop the outlier.



4. If the outlier *creates* a significant association, you should drop the outlier and not report any significance. From the diagram below you can see that the linear relationship is only created due to the outlier.



**Example:** A group of patients are all suffering from the same condition in a hospital ward. Their temperature is in degrees Celsius are recorded as follows.

37.2 36.8 40.0 38.2 38.5 63.9 39.6 38.0 39.3 38.0

(a) Identify the outlier in these figures and explain why it should be discarded.

|  |
| --- |
|  |

(b) Using Excel, Calculate the difference in the mean before and after removing the outlier.

|  |
| --- |
|  |

(c) Without calculating what would happen to the standard deviation and SIQR of the information above if you removed the outlier?

|  |
| --- |
|  |

With regards to linear regression, if you were to remove the outliers how might that be a benefit?

|  |
| --- |
|  |

**Exercise: Outliers**

1. The following data shows the amount of times a website is used every month.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Times used (Thousands) | 1.4 | 1.6 | 1.5 | 1.9 | 0.25 | 2.3 | 2.4 | 2.8 | 2.6 | 2.7 | 5.06 | 2.9 |

a) Type the following into an Excel Spreadsheet, create a scatterplot and calculate the following put your answers into the ‘Before’ column.

|  |  |  |
| --- | --- | --- |
|  | Before | After |
| Mean Times Used |  |  |
| Standard Deviation Times Used |  |  |
| Linear Expression |  |  |
| Correlation |  |  |
| R2 (Coefficient of determination) |  |  |

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.

2. The following data shows the amount of trees being cut down in a rain forest each year.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Amount of Tree(Ten Thousand) | 12 | 11.6 | 10.9 | 10.8 | 28 | 42 | 10.5 | 10.2 | 9.9 | 9.6 | 9.3 | 9.2 | 9.5 | 8.5 |

a) Type the following into an Excel Spreadsheet, repeat the same process you did for question 1.

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.

3. The following data shows the amount of rainfall each year in the month of November

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Average Rainfall cm | 0.04 | 4.8 | 5.2 | 5 | 5.5 | 5.8 | 1.7 | 6.7 | 6.5 | 6.9 | 11.8 | 6.8 | 7.1 | 7.5 |

a) Type the following into an Excel Spreadsheet, repeat the same process you did for question 1.

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.

4. The following shows a cars average speed vs its fuel consumption

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Average Speed | 28 | 29 | 33 | 36 | 37 | 48 | 49 | 50 | 50 | 52 | 55 | 56 | 57 | 61 | 70 | 75 | 81 |
| Fuel Consumption | 78 | 76 | 24 | 72 | 70 | 65 | 68 | 62 | 63 | 58 | 26 | 54 | 53 | 52 | 50 | 48 | 19 |

a) Type the following into an Excel Spreadsheet, repeat the same process you did for question 1.

b) Copy and paste the table in Excel to create another one. Remove any that you consider to be ‘outliers’. Once this is done, compute the same calculations and put your answers in the ‘After’ column.

# Hypothesis Tests

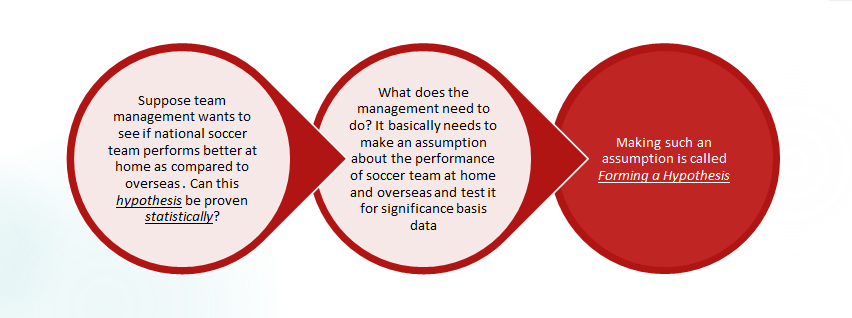
**What is a hypothesis?**

A hypothesis is a statement which might be true e.g. Sam has a hypothesis that large dogs are better at catching balls than small dogs.

A hypothesis is a prediction that is supported with an explanation. The explanation might use previous knowledge or scientific evidence, e.g. if a plant is given unlimited amount of sunlight then the plant will grow to its largest possible size; if I eat more vegetables then I will lose weight; if I brush my teeth every day then I will not need fillings.

**What is a hypothesis test?**

A hypothesis test is a statistical test to ‘prove’ a hypothesis held by a researcher. A hypothesis test compares two datasets, or a sample from a dataset. Afterwards, a conclusion will be made about the characteristics of the comparison.



**Why do we need one?**

Hypothesis testing is the process used to evaluate the strength of evidence from the sample. This is often called a research question and is basically an idea that must be put to the test. Hypothesis testing is very important in the scientific community and is necessary for advancing theories and ideas, e.g. a hypothesis test is used when you are wishing to test a new product e.g. a new drug for treatment of a disease, a claim that a new shampoo works, a new vaccine is safe, etc.

For these statistical tests, when a researcher attempts to prove some hypothesis of interest e.g. that a new treatment is more effective than the existing treatment. Initially they assume the contrary view (i.e. the new treatment is *not* more effective than the existing one) and only comes down in support of the hypothesis of interest if the data gathered is sufficiently unlikely to have been generated by the contrary view.

The contrary view is known as the **Null Hypothesis** and it is written as Ho.

What we are testing for is referred to as the **Alternative Hypothesis** and is written as H1.

**Example:** A study was conducted to compare the effect of two different pain killers on blood glucose levels. Fifteen subjects were given painkiller *A* and 12 were given painkiller *B* and the blood glucose levels recorded in mg/kg as shown in the table below. The objective of the study is to determine if the blood glucose levels are higher with one or other painkiller.

|  |  |
| --- | --- |
| Painkiller A | 44, 69, 51, 71, 52, 71, 55, 76, 60, 82, 62, 91, 66, 108, 68 |
| Painkiller B | 52, 95, 64, 97, 68, 107, 77, 116, 79, 83, 84, 88 |

(a) State the Null Hypothesis of test above

|  |
| --- |
| Ho = |

(b) State the Alternative Hypothesis

|  |
| --- |
| H1 = |

For some statistical tests, once you have decided what your Null and Alternative Hypothesis are, you need to see if you have normally distributed data.

This is done by the following 3 ways

* Check that the mean and median are close
* A Histogram of each to show a bell shaped distribution
* A Shapiro-Wilks tests where the resultant p-value is >0.05 (covered later on)

For now look at the first two points.

(c) On R studio input the data and show the mean and median of Painkiller A and B along with a Histogram to show if the distribution is Normal

Write below what you found.

|  |
| --- |
|  |

**P value** is the probability value is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true)

𝝆−𝒗𝒂𝒍𝒖𝒆𝒔

The level of statistical significance is often expressed as a 𝜌−𝑣𝑎𝑙𝑢𝑒 between 0 and 1. The smaller the 𝜌−𝑣𝑎𝑙𝑢𝑒, the stronger the evidence that you should reject the null hypothesis. In simpler terms, the 𝜌−𝑣𝑎𝑙𝑢𝑒 **does not** provide any evidence in support of 𝐻0.

𝜌<0.001 indicating very strong evidence against 𝐻0.

𝜌<0.01 indicating strong evidence against 𝐻0.

𝜌<005 indicating moderate evidence against 𝐻0.

𝜌<0.1 indicating weak evidence or a trend.

𝜌≥0.1 indicating insufficient evidence.

A p-value of less than 0.05 is statistically significant. It indicates strong evidence against the null hypothesis as there is a 5% probability the null is correct. If the p-value is < 0.05 it means we reject the Null Hypothesis Ho and accept the Alternative Hypothesis H1.

Going back to our example about painkillers. Perform a Hypothesis test on the information above using R studio and the command *t.test(X,Y)*.

(d) State the p-value given and write what that means with relation to the context of the question.

|  |
| --- |
|  |

(e) What are the positive aspects of performing these Hypothesis tests, how can the information be used?

|  |
| --- |
|  |

We will return to most of these aspects later on.

**Example 2:** In Santa Clara more than 30% of people voted in the primaries. Using Ho, H1 notation write the Null and Alternative Hypothesis.

|  |
| --- |
|  |

**Exercise: Null and Alternate Hypothesis**

For the following state the null and alternative hypothesis, label them as Ho and H1.

1. Are people who live in country A heavier than those that live in country B?

2. A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

3. We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0).

4. We want to test if college students take less than five years to graduate from college, on the average.

5. In an issue of U.S. News and World Report, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%.

**Confidence Intervals**

Because the datasets we use are typically samples of a larger population, the mean that we compute for the sample is only an estimate of the population mean.

For a given sample, we can produce a range of values (an interval) in which we are fairly certain the true population mean lies.

Usually we quote these “confidence intervals” at a level of 95% certainty.

This means there is only a 5% chance that the true population mean lies outside this range.

**Example:** Using the dataset *Cholesterol\_R.csv*

Input the data into R studio

(a) Calculate the mean of *before*, write it below

|  |
| --- |
|  |

(b) Calculate the mean of *After 8 Weeks*

|  |
| --- |
|  |

(c) Use the command *t.test*, to perform a hypothesis test and hence show the maximum and minimum difference in the two means.

|  |
| --- |
|  |

This means that there 95% chance that the difference in the two means will be between \_\_\_\_\_ and \_\_\_\_\_.

To calculate the confidence interval yourself you need the following

* Mean
* Sample Size
* Standard Deviation
* Critical value for 95%
* The data follows a normal distribution

The formula is as follows

**Example 2:** A hardware manufacturer produces bolts used to assemble various machines. Assume that the diameter of bolts produced by this manufacturer has an unknown population mean 𝝁 and the standard deviation is 0.1 mm, suppose the average diameter of a simple random sample of 50 bolts is 5.11 mm. Calculate the 95% confidence interval.

|  |
| --- |
|  |

**Exercise: Confidence Intervals:**

1. You want to rent an unfurnished one-bedroom apartment in Boston next year. The mean monthly rent for a simple random sample of 32 apartments advertised in the local newspaper is $1,400.

Assume that the standard deviation is known to be $220.

Find a 95% confidence interval for the mean monthly rent for unfurnished one-bedroom apartments available for rent in this community.

2. We have IQ test scores of 31 seventh-grade girls in a Midwest school district.

We have calculated that sample mean is 105.84 and the standard deviation is 14.27.

Give a 95% confidence interval for the average score in the population. What is the margin of error?

3. Determine the sample size needed to estimate the average weight of all second-grade boys if we want to be within 1 pound with 95% confidence.

Assume we know that the standard deviation of such weights is 3 pounds.

4. How big must the sample size be if the margin of error on the waiting times of a hospital is 2 minutes? If the standard deviation is 15 minutes and the confidence interval is 95%.

# Types of Hypothesis Test

**T-test**

This is the test you have previously used but you might not know what it does. A t-test looks at the means of data sets and tests for the likelihood there is any real difference between them.

To perform a t-test the data should be

* Numerical data (either continuous or ordinal)
* Data should be normally distributed
  + Check that the mean and median are close
  + A Histogram of each to show a bell shaped distribution
  + A Shapiro-Wilks tests where the resultant p-value is >0.05
* Ideally the data should be randomly selected

There are also different types of T-Test

**Two Sample T-Tests**

This is done on two independent samples of data. Two-sample t-Tests compare the means of precisely two groups, no more and no less. Typically you perform this test to determine whether two **population** means are different.

For example,

1. Do students who learn using Method A have a different mean score than those who learn using Method B?

2. Testing out to see if medication has an effect on patients resting heart rate if you were comparing patients who took the medications against patients who didn’t take the medication.

The standard form tests the following hypotheses;

Null hypothesis: the two population means are equal.

Alternative hypothesis: the two population means are not equal.

If the 𝜌−𝑣𝑎𝑙𝑢𝑒 is less than the *significance* level (e.g. 0.05), you can reject the null hypothesis. The difference between the two means is statistically significant. The sample provides strong enough evidence to conclude that the two population means are different.

A p-value of 0.05 means there is a 5% chance that there is no difference between the true means.

**Example:** For the following use the data set *Two Sample T Test.csv*.

Upload the data into R Studio.

(a) Look at the data and make sure it is numerical

(b) Check that the data is normally distributed, write what you found below.

|  |
| --- |
|  |

You will now perform a Hypothesis test.

(c) What is the null hypothesis of the test?

|  |
| --- |
|  |

(d) Using the command *t.test(x,y)* perform a hypothesis test. State the p-value and the result below.

|  |
| --- |
|  |

**Example 2:** Group Body Fat Percentages  
The sample data is from a group of men and women who worked out at the gym three times a week for a year. Their percentage of body fat is recorded in the table below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Men | 13.3 | 6.0 | 20.0 | 8.0 | 14.0 | 19.0 | 18.0 | 25.0 | 16.0 | 24 | 15.0 | 1.0 | 15.0 |
| Women | 22 | 16 | 21.7 | 21 | 30 | 26 | 12 | 21.2 | 28.0 | 23 |  |  |  |

Record the data into an excel spreadsheet and save as a csv file

Upload the data into R studio.

(a) Using R studio, show if the data is normally distributed. Report your findings below.

|  |
| --- |
|  |

(b) State the null and alternative hypothesis referring to them as Ho and H1.

|  |
| --- |
|  |

(c) Perform a Hypothesis test. State the p-value and confidence interval and interpret what that means.

|  |
| --- |
|  |

**Paired Sample T-Test**

A paired t-Test is also called a correlated pairs t-Test or dependent samples t-Test. Dependent samples are essentially connected, they are tests on the same person or thing say, one year apart. For example, tests on the same group of people before and after a fitness regime; or a comparison of two different measurements/treatments are applied to the same subjects.

To apply the paired t-test to test for differences between paired measurements, the following assumptions need to hold;

* Subjects must be independent.
* Each pair must be obtained from the same source (i.e. a person’s weight before and after.
* The data must be normally distributed.

**Example** 1.

Calculate a paired t-Test for the following:

Pupils in a Maths class sit a test and their scores are noted. After 6 weeks of supported study, there is another test and, again, their scores are noted.

|  |  |  |
| --- | --- | --- |
| Student | Before | After |
| 1 | 3 | 20 |
| 2 | 3 | 13 |
| 3 | 3 | 13 |
| 4 | 12 | 20 |
| 5 | 15 | 29 |
| 6 | 16 | 32 |
| 7 | 17 | 23 |
| 8 | 19 | 29 |
| 9 | 23 | 25 |
| 10 | 24 | 15 |
| 11 | 32 | 30 |

(a) State the null and alternative hypothesis

|  |
| --- |
|  |

We will use the paired t-Test to see if the supported study sessions affected the test scores.

Input the data into R studio.

Using the command *t.test(x,y, paired = “TRUE”)* perform a hypothesis test.

State what the p-value and interpret what this means in the context of the data.

|  |
| --- |
|  |

What does this mean for our test? It means our sample data shows a higher mean test score after help than before help (supported study). The significant result suggests that we can conclude that the same is true at the population level. We can infer that the population mean of the test score after help is greater than the population mean of the test score before help.

**Example 2.** For the following use the *Crime* preinstalled data set from R studio.

Look at the variables *CrimeRate* and *CrimeRate10*.

These show the amount of offences for every million people.

(a) Test these for normal distribution.

|  |
| --- |
|  |

(b) State the null and alternative hypothesis for performing a statistical test.

|  |
| --- |
|  |

(c) Perform a t-test. State the p-value and confidence interval and interpret what this means.

|  |
| --- |
|  |

**Mann-Whitney/Wilcox Test**

An alternative test for paired samples that can be considered is called the Wilcoxon (signed-rank) test. This test can be useful as it **does not** require the data to be normally distributed or for the standard deviations (variances) to be similar.

If the data is normally distributed and the standard deviations close together then it is recommended that a t-test is used.

**Example:** A comparison of two methods A and B for measuring physical fitness, a random sample of eight people was assessed by both method. Their scores were recoded as follows.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Method A | 11.2 | 8.6 | 6.5 | 17.3 | 14.3 | 10.7 | 9.8 | 13.3 |
| Method B | 10.4 | 12.1 | 9.1 | 15.6 | 16.7 | 10.7 | 12.8 | 15.5 |

(a) Is the data suitable for a T test? State your reason below.

|  |
| --- |
|  |

(b) State the null and alternative hypothesis below

|  |
| --- |
|  |

(c) Perform a hypothesis test to show if there is any significant difference between the two methods of measuring fitness.

|  |
| --- |
|  |

**Exercise: T-testing**

1. A tennis coach wants to determine whether a new racquet improves the speed of his pupil’s serves (faster serves are considered better). He tests a group of 9 children to discover their service speed with their current racquet and with the new racquet. The results are shown in the table below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Child | A | B | C | D | E | F | G | H | I |
| Speed with current racquet | 58 | 68 | 49 | 71 | 80 | 57 | 46 | 57 | 66 |
| Speed with new racquet | 72 | 81 | 52 | 59 | 75 | 72 | 48 | 62 | 70 |

(a) State the null and alternative hypothesis.

(b) Perform a statistical analysis to see if buying a new racquet is guaranteed to increase your service speed. Make sure to confirm which type of test you will use and state the reasons for this.

2. Petra noticed that one of her apple trees grew in the shade and other did not. She wanted to find out if apples from the tree in the shade weighed less than those in the sun. She picked 9 random apples from each tree and weighted them.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tree in Shade | 75 | 82 | 93 | 77 | 85 | 78 | 91 | 83 | 92 |
| Tree Not in Shade | 74 | 81 | 95 | 79 | 95 | 82 | 93 | 88 | 90 |

(a) Is the data paired?

(b) Show if the data is normally distributed and if it is suitable for a t test.

(c) State the null and alternative hypothesis

(d) Perform a statistical test to show if there is any difference in the weight from trees that grow in shade and those that don’t.

3. In an attempt to determine if two competing brands of cold medicine contain, on the average, the same amount of acetaminophen, twelve different tablets from each of the two competing brands were randomly selected and tested for the amount of acetaminophen each contains. The results (in milligrams) follow. Use a significance level of 0.01.

Brand A Brand B

517, 495, 503, 491 493, 508, 513, 521

503, 493, 505, 495 541, 533, 500, 515

498, 481, 499, 494 536, 498, 515, 515

State and perform an appropriate hypothesis test.

4. In an investigation to compare the accuracy of Crackshot and Fastfire12-bore shotguns in clay pigeon shooting, ten competitors each fired 100 shots with each make of gun. Their scores are shown in the table below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Competitor | A | B | C | D | E | F | G | H | I | J |
| Crackshot | 93 | 99 | 90 | 86 | 85 | 94 | 87 | 91 | 96 | 79 |
| Fastfire | 87 | 91 | 86 | 87 | 78 | 95 | 89 | 84 | 88 | 74 |

Assume the data is normally distributed

(a) State the null and alternative hypothesis

(b) Examine the claim that the Crackshot shotgun is more accurate for clay pigeon shooting.

5. Use a Mann-Whitney U test to test if heart rate differs between men and women at the 95% level.

|  |  |
| --- | --- |
| Heart Rate Women | Heart Rate Men |
| 84 | 80 |
| 81 | 74 |
| 80 | 73 |
| 70 | 72 |
| 72 | 78 |
| 69 | 75 |
| 65 | 70 |
| 74 | 74 |
| 80 | 69 |

(a) State the null and alternative hypothesis

(b) Compare using a Wilcox text to determine if there is a significant difference.

**Chi-Square Test for Independence**

This test is used to determine if there is a relationship between two categorical variables. The chi-square test evaluates is there is a significant association between the categories of the two variables e.g. between gender and whether they smoke or not.

For the Chi-square test of independence we need two variables.

Our idea is that the variables are not related (null hypothesis);

Null hypothesis: the variables are not related

*Alternative hypothesis: the variables are related*

For a valid test, we need:

* Data values that are a simple random sample from the population of interest.
* Two categorical or nominal variables. (Nominal identifies the category, e.g. 1 for male, 2 for female).

Look at the following.

**Example:**

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four year college students, and nonstudents. In the table below a sample of volunteers and the number of hours they work per week is shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type of Volunteer | 1-3 Hours | 4-6 Hours | 7-9 Hours | Total |
| Community College Students | 111 | 96 | 48 | 225 |
| Four Year College Students | 96 | 133 | 61 | 290 |
| Non Students | 91 | 150 | 53 | 294 |
| Total | 298 | 379 | 162 | 839 |

(a) For a Chi Square test there needs to be two categorical variables, state the categorical variables below

|  |
| --- |
|  |

(b) Input the data into R Studio

*Extension: Do so directly and add Row and Column titles.*

(c) State the Null and Alternative Hypothesis in the box below.

|  |
| --- |
|  |

(d) Perform a Chi-Square test for independence to see if there is a relationship between the variables. State the p-value in the box below and interpret the result.

|  |
| --- |
|  |

**Example 2:** A cinema looks into if there is a relationship between if customers buy snacks and what kind of movie they attend.

(a) State the Null and Alternative Hypothesis

|  |
| --- |
|  |

The table is written below

|  |  |  |
| --- | --- | --- |
| Type of Movie | Snacks | No Snacks |
| Action | 50 | 75 |
| Comedy | 125 | 175 |
| Family | 90 | 30 |
| Horror | 45 | 10 |

(b) Input the data into R Studio

(c) State the two categorical variables

|  |
| --- |
|  |

(d) Perform a Chi Square Test and interpret the result below, remember to state the p-value and if it is a significant result.

|  |
| --- |
|  |

**Chi Square Test for Goodness of Fit**

Another use of the chi-squared test is for goodness-of-fit. This test is used to find out how the observed value of a given categorical outcome differs from the expected value. The term ‘goodness-of-fit’ is used to compare the observed sample distribution with the expected probability distribution.

For the goodness-of-fit test, we need one variable. We also need an idea (hypothesis) about how that variable is distributed. Here are a couple of examples.

**Example:** We have bags of sweets with five flavours in each bag. The bag should contain an equal number of pieces of each flavour. The idea we’d like to test is that the proportions of five flavours in each bag are the same.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Red | Yellow | Green | Blue | Orange |
| Observed | 7 | 11 | 16 | 7 | 9 |
| Expected | 10 | 10 | 10 | 10 | 10 |

A Chi Square test will look at these numbers and assuming that they should all appear the same what is the probability that this could happen if everything is fair i.e. there is an even chance of getting each sweet.

Null Hypothesis: There is no difference between the amounts of each sweet

Alternate Hypothesis: There is a difference between the amounts of each sweet.

We will look at a similar example later on.

**Example 1:** 200 rolls of a die result in the following distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 28 | 30 | 22 | 31 | 38 | 51 |

(a) How many of each should you have expected?

|  |
| --- |
|  |

(b) State the null and alternative hypothesis

|  |
| --- |
|  |

(c) Perform a Chi Square Test to see if we can conclude if the die is fair.

|  |
| --- |
|  |

**Example 2:** A college reports that 50% of the students in its statistics classes are freshmen, 30% are sophomores, 10% are juniors, and 10% are seniors. A simple random sample of 65 such students has the following breakdown.

|  |  |  |  |
| --- | --- | --- | --- |
| Freshman | Sophomore | Junior | Senior |
| 28 | 24 | 9 | 4 |

Using a statistical test show whether there is any reason to doubt the percentages recorded by the college.

|  |
| --- |
|  |

**Exercise: Chi Square Testing**

1. The members of a sports team are interested in whether the weather has an effect on results. They play 50 matches, with the following results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Weather** | | Total |
| Good | Bad |
| **Result** | Win | 12 | 4 | 16 |
| Draw | 5 | 8 | 13 |
| Lose | 7 | 14 | 21 |
| Total | | 24 | 26 | 50 |

Formulate suitable null and alternative hypotheses. Then perform a Hypothesis test to show if the weather has any effect on the result of the games.

2. A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck.

In a test, you counted 1600 cards, and observed the following:

Spades - 404

Hearts- 420

Diamonds -400

Clubs - 376

Could it be that the suits are equally likely? Or are these discrepancies too much to be random?

3. A staff member of an emergency medical service wishes to determine whether the number of accidents is equally distributed during the week. A week is selected at random, and the following data were obtained.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Day | Mon | Tues | Wed | Thurs | Fri | Sat | Sun |
| No. of Accidents | 28 | 32 | 15 | 14 | 38 | 43 | 19 |

Is there evidence to reject the hypothesis that the number of accidents is equally distributed throughout the week? State any null or alternative hypothesis you may have used and interpret the p-value.

4. The chair of the history department of a college hypothesizes that the final grades are distributed as 40% A’s, 30% B’s, 20% C’s, 5% D’s, 5% F’s. At the end of the semester the following grades were earning.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade | A | B | C | D | F |
| Number | 88 | 70 | 44 | 12 | 10 |

Perform a Hypothesis test to show is if the History is perform as expected or do some changes need to be made?

5. University students may be interested in whether or not their degree have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. Below are the data. Conduct a hypothesis test.

| **Degree** | **< $50,000** | **$50,000 - $68,999** | **$69,000 +** |
| --- | --- | --- | --- |
| English | 5 | 20 | 5 |
| Engineering | 10 | 30 | 60 |
| Nursing | 10 | 15 | 15 |
| Business | 10 | 20 | 30 |
| Psychology | 20 | 30 | 20 |

**Two Proportion Z-Test**

A **two proportion z-test** is used to test for a difference between two population proportions.

Suppose we want to know if there is a difference in the proportion of residents who support a certain law in county A compared to the proportion who support the law in county B.

Since there are thousands of residents in each county, it would take too long and be too costly to go around and survey every individual resident in each county.

Instead, we might take a simple random sample of residents from each county and use the proportion in favour of the law in each sample to estimate the true difference in proportions between the two counties:

**Example 1:**  
Let’s say we have two groups of student A and B. Group A with an early morning class of 400 students with 342 female students. Group B with a late class of 400 students with 290 female students. Use a 5% alpha level. We want to know, whether the proportions of females are the same in the two groups of the student? Here let’s use **prop.test()**.

|  |  |
| --- | --- |
| ***Syntax:*** *prop.test(x, n, p = NULL, alternative = “two.sided”, correct = TRUE)*  ***Parameters:******x*** *= number of successes and failures in data set.* ***n*** *= size of data set.* ***p*** *= probabilities of success. It must be in the range of 0 to 1.* ***alternative*** *= a character string specifying the alternative hypothesis.* ***correct*** *= a logical indicating whether Yates’ continuity correction should be applied where possible.*   |  | | --- | | prop Test in R  prop.test(x = c(342, 290),            n = c(400, 400)) |   **Output:**  2-sample test for equality of proportions with continuity correction  data: c(342, 290) out of c(400, 400)  X-squared = 19.598, df = 1, p-value = 9.559e-06  alternative hypothesis: two.sided  95 percent confidence interval:  0.07177443 0.18822557  sample estimates:  prop 1 prop 2  0.855 0.725 |

**Example 2:** It was conjectured that more women than men live to one hundred years old.

1. State a null and alternative hypothesis

|  |
| --- |
|  |

(b) An audit found 23 out of 9,500 women and 11 out of 10,000 men were centenarians.

In the *sample*, did more women live to 100?

|  |
| --- |
|  |

(c) What statistical test should be used to determine if there is a *significant* difference?

|  |
| --- |
|  |

1. Perform this test and state your findings below

|  |
| --- |
|  |

**Sampling error**

Statistics is the art of using a sample to try and understand a population. It should therefore be understood that any conclusions about the population drawn from the sample are necessarily approximate, and could be different if a different sample was chosen.

There are two types of error:

**Sampling error** occurs if the sample is not typical of the population, and accidentally introduces some bias (a problem with the design of the sample)

**Non-sampling error** occurs if data is incorrectly recorded, or there are gaps in the data as it was not possible to sample fully (a problem with the execution of the sample)

A common cause of sampling error is if the sample is performed all at one time.

Sampling error is minimised by taking a large sample, although this is time consuming so instead there is an attempt to try and describe the uncertainty, for example with confidence intervals.

**Question:** A survey at a local Post Office last Tuesday found that 85% of respondents agreed that

*“Young people have much less respect for their elders than they used to”.*

This was then used to claim that in the whole country people don’t respect their elders.

(a) State two possible causes of sampling error

(b) State two possible causes of non-sampling error

*(c) The sample is very limited. Because the sample was made at just one location it does not reflect the whole population well. The sample was also just made at just one time.*

*(d) Respondents may not have filled in the survey honestly. The figure of 85% may have been reached by a counting error. Some surveys may have got lost.*