Chapter 1 – Data Collection

- **Quantitative data** – data involving numerical observations (quantities). E.g. number of people, volume of water
- **Qualitative data** – data involving non-numerical observations (qualities). E.g. colour, name, nation.

- **Continuous variable** – data you can measure. E.g. seconds, height, weight, volume, speed.
- **Discrete variable** – data you can count. E.g. number of books, number of zebras, shoe size.
  *Both continuous and discrete variables produce quantitative data.

- Individual members of a population are known as **sampling units**.
- When the sampling units are named or numbered, they form a **sampling frame** (because you can order them).

	Sampling Method	Description	Advantages	Disadvantages
CENSUS	**Census**	Observes or measures every member of a population.	1. It should give a completely accurate result.	2. Time consuming and expensive 3. Cannot be used when testing destroys the item. 4. Hard to process large quantity of data.
RANDOM SAMPLING	**Simple Random Sampling**	A selection of observations from the population. Each sampling unit has an equal and random chance of being selected. *Random numbers are generated with a calculator or computer.	1. Free of bias (Eaxch sampling unit has a kown and equal chance of selection) 2. Easy and cheap to implement for small populations and small samples.	1.Not suitable when the population size is large (may not represent each element of the pop.) 2. A sampling frame is needed.
	Systematic Sampling	Make a list. Choose every nth thing in the list. (The first person is not necessarily 1st in the list, they're chosen at random).	1. Simple and quick to use. 2. Suitable for large samples and large populations.	3. A sampling frame is needed. 4. It can introduce bias if the sampling frame is not random.
	Stratified Sampling	Separate the population into mutually exclusive strata (categories like male/female, age etc.) and take a random (and proportional) sample within each.	1. Sample accurately reflects the population structure. 2. Guarantees proportional representation of groups within a population.	3. Population must be clearly classified into distinct strata. 4. Selectin within each stratum suffers from the same disadvantages as simple random sampling.
NON-RANDOM SAMPLING	**Quota Sampling**	An interviewer selects a sample which reflects the characteristics of the whole population.	1. Allows even a small sample to be representative of the whole population (unlikely with random sampling) 2. No sampling frame required. 3. Quick, easy, inexpensive. 4. Allows for easy comparison between groups in a population.	5. Non-random sampling can introduce bias. 6. Population must be divided into groups, which can be costly or inaccurate. 7. Non-responses are not recorded as such.
	Opportunity Sampling **(Convenience Sampling)**	Taking a sample of people who can be conveniently found.	1. Easy to carry out 2. Inexpensive	3. Unlikely to provide a representative sample 4. Highly dependent on individual researcher.

Chapter 2: Variance & Standard Deviation

sum of the variance from the mean (squared).

* If the data is in a frequency table, multiply each input in the sum by the frequency with which it occurs.

$$\text{Variance} = \sigma^2 = \frac{\sum_{r=1}^{n}(x_r-\bar{x})^2}{n} = \frac{\sum_{r=1}^{n} x_n^2}{n} - \left(\frac{\sum_{r=1}^{n} x_n}{n}\right)^2$$

So square root of variance!

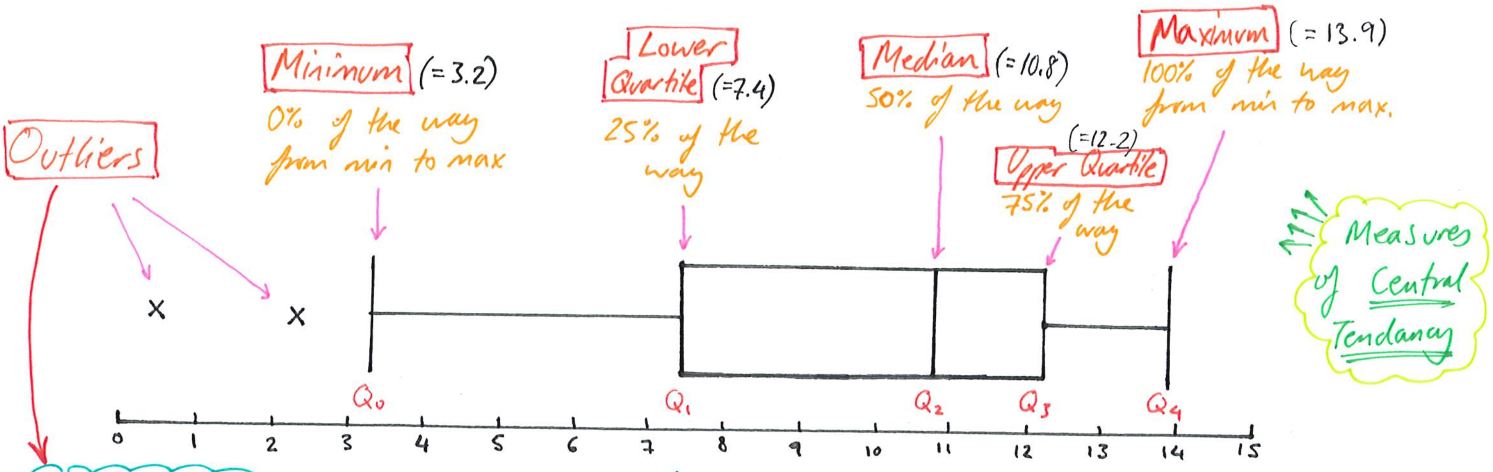
$$\text{Standard Deviation} = \sigma$$

* If data is grouped in class intervals like $15 \leq x < 18$, then use the midpoint (16.5) for all the values in the interval.

$$\sigma^2 = \frac{\sum_{r=1}^{n} f_r(x_r-\bar{x})^2}{n} = \frac{\sum_{r=1}^{n} f_r x_n^2}{\sum_{r=1}^{n} f_r} - \left(\frac{\sum_{r=1}^{n} f_r x_r}{\sum_{r=1}^{n} f_r}\right)^2$$

σ = standard deviation is the square root again!

Chapter 2/3 - Representing Data



Interquartile Range

UQ − LQ = 12.2 − 7.4 = **4.8**

Measures of Spread

Range Maximum − Minimum = 13.9 − 3.2 = **10.7**

Outliers are defined (usually) as less than $Q_1 - 1.5(IQR)$ or more than $Q_3 + 1.5(IQR)$ though the 1.5 can change. They'll tell you in the question if it does.

SHEET 3 - PROBABILITY & VENNS

CHAPTER 5 (Y1) - PROBABILITY + CHAPTER 2 (Y2) CONDITIONAL PROBABILITY.

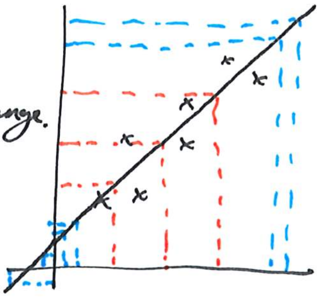
(also includes some basic tree diagrams, which are used to analyse repeated events).

DEFINITIONS

- An EXPERIMENT is a repeatable process that gives rise to a number of OUTCOMES
- An EVENT is a collection of one or more outcomes.
- A SAMPLE SPACE is the set of all possible outcomes.

e.g. If I roll 2 fair dice, then the experiment is "rolling the 2 dice", an outcomes is "1, 1" "3, 4" or "4, 3". An event is "we get a 4, then a 3" ($\frac{1}{36}$ chance), or maybe "we get the same number twice", or "we get 2 evens." The sample space is all 36 possible combinations.

- EXTRAPOLATION is extending what the data shows, outside of the range. INTERPOLATION is inside the range.



- Black is 'real' data
- Blue is extrapolation
- Red is interpolation

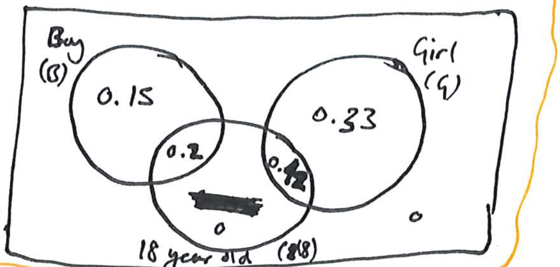
- MUTUALLY EXCLUSIVE events cannot happen at the same time (a randomly selected student cannot be "a boy" and "a girl")

- INDEPENDENT EVENTS have no effect on eachothers probabilitys (e.g. if a randomly selected student is "a girl" and is "18 years old").

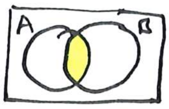
$P(B \cup G) = (0.15+0.2)+(0.33+0.42)$ because

for mutually exclusive events $P(A \cup B) = P(A) + P(B)$

for independent events $P(A \cap B) = P(A) \times P(B)$

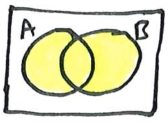


VENN DIAGRAMS & SET NOTATION



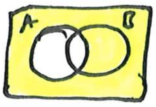
$P(A \cap B)$ A and B

"A intersection B"



$P(A \cup B)$ A or B, or both!

"A union B"



$P(A')$ Not A

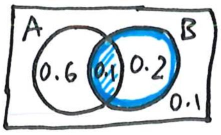
"A complement"

For rolling a fair dice once,

$n(\text{even}) = 3$, because 3 of the 6 outcomes are even numbers.

$P(\text{even}) = \frac{1}{2}$ because its a probability (which is always between 0 and 1).

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.2+0.1} = \frac{1}{3}$$

"A, given that we're already in B"

So $P(A \cap B) = P(A|B) \times P(B)$

You can use this and the above "$P(A \cap B) = P(A)P(B)$" formulae to test for independence.

If A and B are independent, then $P(A|B) = P(A|B') = P(A)$

$P(B|A) = P(B|A') = P(B)$

SHEET 4 - DISTRIBUTIONS

DEFINITIONS

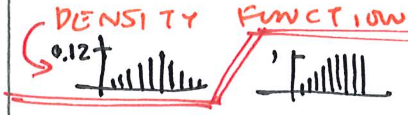
A RANDOM VARIABLE is a variable X, Y, A, (always a capital letter) which can take a number of possible values (the outcomes), but we don't which one until we've done the experiment and observed the outcome. These outcomes have probabilities which can be found using formulas:

- $\binom{n}{r} = \frac{n!}{r!(n-r)!}$
- CONTINUOUS DATA
- DISCRETE DATA

~~PROBABILITY DENSITY FUNCT~~

Individual probabilities of $P(X=x)$ is:

- PROBABILITY DISTRIBUTION / DENSITY FUNCTION (PDF)

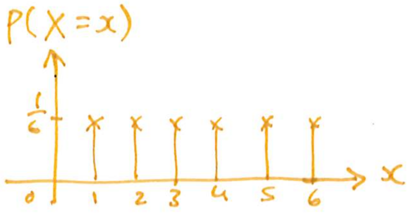


- CUMULATIVE DISTRIBUTION / DENSITY FUNCTION is all the individual ones added up.

DISCRETE UNIFORM DISTRIBUTION

~~Thing~~ Things where eney outcome has the same likelihood. E.g. flipping a coin, rolling a die, picking a boy or girls.

e.g. Rolling a fair die:



$P(X=x) = \frac{1}{6}$ for all x in the sample space.

Generally

$P(X=x) = \frac{1}{n}$ for n possible outcomes.

BINOMIAL DISTRIBUTION (DISCRETE DATA)

- $X \sim B(n, p)$ no. of trials; probability of "success"
- $P(X=r) = \binom{n}{r} p^r (1-p)^{n-r}$
- For cumulative distributions, see the table.

(rember that it's discrete, so $P(X \leq 8) = 1 - P(X \geq 9) = 1 - P(X > 8)$, because if $X > 8$, then $X \geq 9$, as X cannot be 8.5, between the 2 values).

When can you model a random variable X with the ~~discrete~~ Binomial distribution? If:

1. There are a fixed no. of trials n.
2. There are 2 possible outcomes (success and failure)
3. There is a fixed probability of success for each trial (p).
4. The trials are all independent.

NORMAL DISTRIBUTION (CONTINUOUS DATA)

$X \sim N(\mu, \sigma^2)$

mean

variance $= \sigma^2$

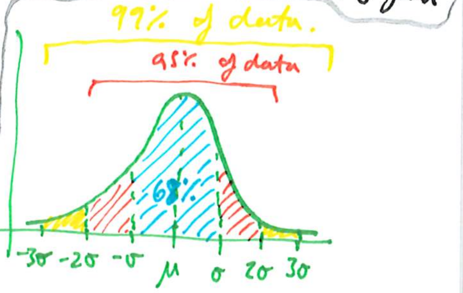
(σ = standard deviation)

FACTS

- It is symmetrical
- μ = mean = median = mode
- It is a bell shape with an asymptote horizontally at $y=0$
- Total area under the curve is 1.
- $P(X<3) = P(X\leq3)$ always.

- Because its continuous data, probabilities are represented by areas. So to say $P(X=3)$ is silly, because the infinitely ~~long~~ thin bar will have area 0. We can only find things like $P(3<X<7)$ which is $P(X<7) - P(X<3)$

These are found in the tables.

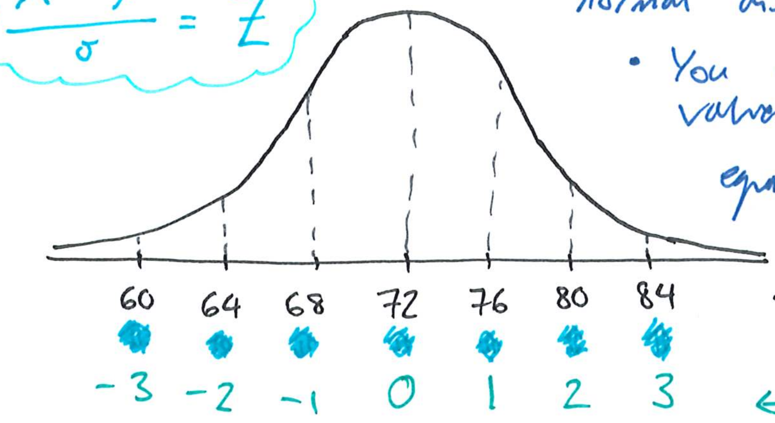


THE STANDARD NORMAL CURVE

Convert between standard normal Z and your observed normal X:

$$\frac{X-\mu}{\sigma} = Z$$

- You can transform any normal curve into the 'standard normal' and not affect any of the probabilities. In this way its easier to compare 2 different normal distributions (heights of men vs. heights of women).
- You can also use this to find unknown values of μ and σ by using simultaneous equations, and the formula to the left.



← height (inches) of UK citizens. $X \sim N(72, 4^2)$

← standard normal $Z \sim N(0, 1^2)$

APPROXIMATING THE (DISCRETE) BINOMIAL DISTRIBUTION WITH THE (CONTINUOUS) NORMAL DISTRIBUTION

What is a "continuity correction"

- Because a normal is continuous, all the outcomes from 69.5 - 70.5 represent '70' on the (discrete) binomial bar chart, which only has bars at 69, 70, 71, etc. (70.5-71.5 represents 71 and so on). Therefore:

$$P(X \leq 70) \approx P(Y < 70.5)$$
$$P(\underset{\text{Binomial}}{X} \geq 70) \approx P(\underset{\text{Normal}}{Y} > 69.5)$$

When can I approximate a Binomial with a Normal?

If: ① n is large (>20)
② p is close to 0.5

In this case, the approximation of $X \sim B(n,p)$ is

$$\underset{\text{The approximation of X is Y.}}{Y} \sim N(\underset{\mu = np}{np}, \underset{\sigma = \sqrt{np(1-p)}}{np(1-p)})$$

HYPOTHESIS TESTING ON ~~THE~~ THE SAMPLE MEAN

Generally: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Suppose that I alone know the mean age of students at our school. I ask you to investigate it. The 10 of you each take 5 samples of 40 kids. You then have 50 samples (each of 40 kids), and 50 '~~means~~' "sample means". If the kids' ages are distributed $X \sim (\mu, \sigma^2)$ then the sample means are distributed $\bar{X} \sim N(\mu, \frac{\sigma^2}{40})$

The "mean of the means" (the mean of $\bar{X}$) ~~the sample means the approximate~~ gives an estimate for the real "population mean" that only I know.

- If I told you that last year the population mean was 14.71 years of age; You could do a Hyp. test to see if your 'mean of means' is extreme enough to say the pop. mean has changed.

2 ways to improve the accuracy of the sample means:

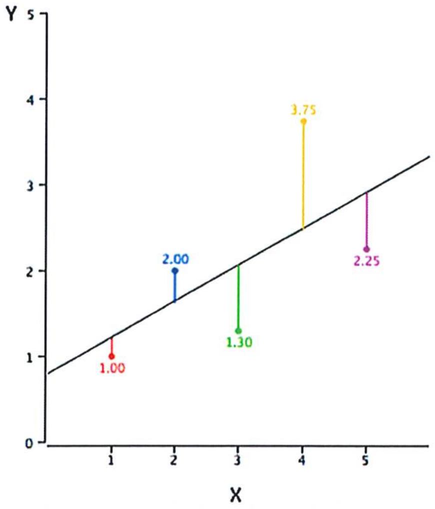
① ↑ sample size n

② ↑ number of samples taken.

Linear Regression

Chapter 4 (Y1) + Chapter 1 (Y2)

- **Bivariate data** has pairs of values for 2 variables (e.g. English and Maths scores for Y6 children)
- **Correlation** describes the nature of the linear relationship between the 2 variables. It is often measured by calculating the **Product Moment Correlation Coefficient (PMCC):**
    - r = –1 is perfectly negative
    - r = 0 is no correlation (a random scattering)
    - r = 1 is perfectly positively correlated.
        - Anything over 0.7 is "strong". Anything less that about 0.3 is "weak".
        - It is VERY common to ask if a linear regression is a suitable model for a given situation. The closer the PMCC is to 1 or -1, the more suitable/sustainable the model is.
- **Independent variables** (explanatory variables) are usually plotted on the x-axis
- **Dependent variables** (response variables) are usually plotted on the y-axis.
    - E.g. amount eaten on the x-axis, weight on the y-axis
    - E.g. child's age is independent, child's height is dependent (on age)
- **A least squares regression line** is a type of 'line of best fit'. It tries to minimise the vertical distances between the data points and the regression line:



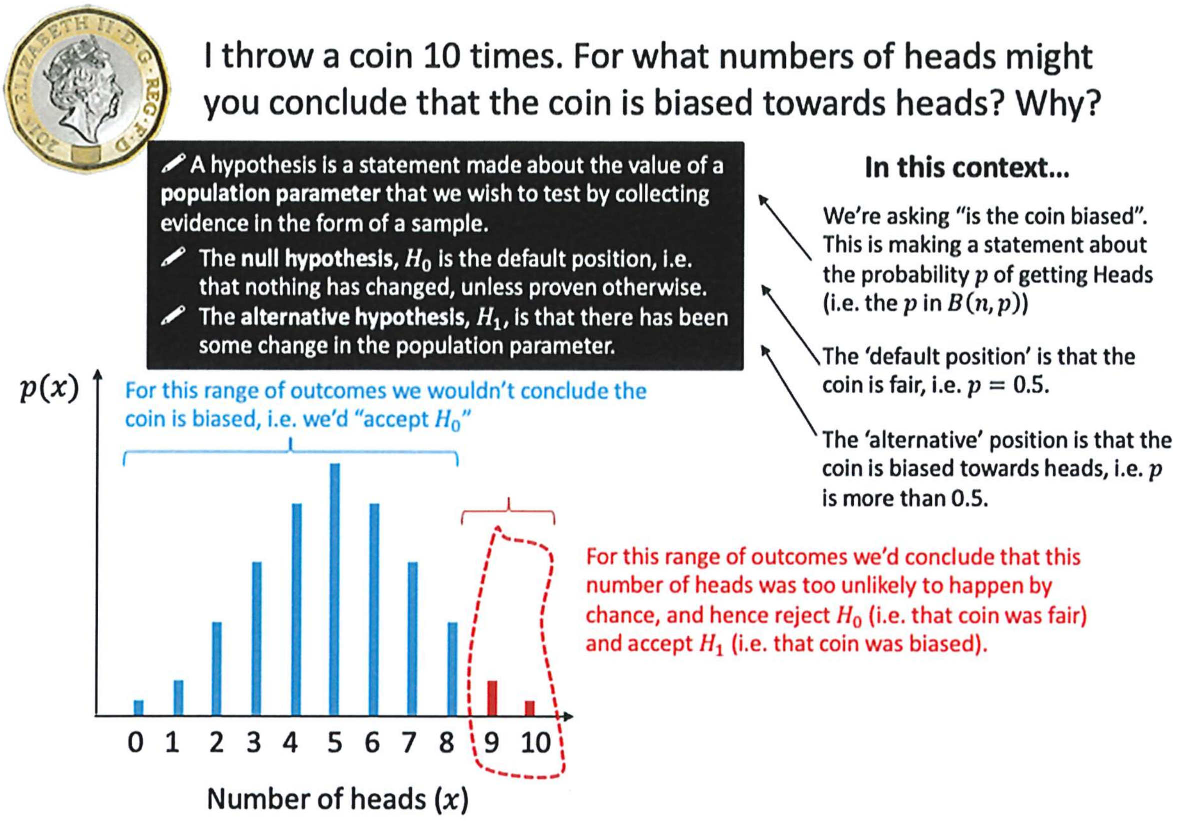
In this case the number that needs to be minimised is:
(blue + yellow) – (red + green + pink)

- If the equation of the regression line is $y = a + bx$, then b is the gradient, and it is the rate at which the dependent variable (y-axis) responds to a change of 1 in the independent variable (x-axis).

- Use your common sense to decide how to answer 1 markers in the exam:
- Things to consider:
    - Correlation between 2 variables does not necessarily mean causation.
    - You must include the numbers when describing what 'b' means in the regression line.
    - If there is a positive correlation, then b is positive. If there's a negative correlation, then b is negative.

Hypothesis Testing

Chapter 7 (Y1) + Chapter 1 (Y2)



I throw a coin 10 times. For what numbers of heads might you conclude that the coin is biased towards heads? Why?

- A hypothesis is a statement made about the value of a **population parameter** that we wish to test by collecting evidence in the form of a sample.
- The **null hypothesis**, H_0 is the default position, i.e. that nothing has changed, unless proven otherwise.
- The **alternative hypothesis**, H_1, is that there has been some change in the population parameter.

In this context...

We're asking "is the coin biased". This is making a statement about the probability p of getting Heads (i.e. the p in $B(n,p)$)

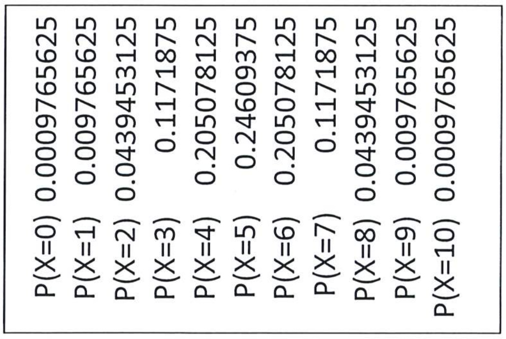
The 'default position' is that the coin is fair, i.e. $p = 0.5$.

The 'alternative' position is that the coin is biased towards heads, i.e. p is more than 0.5.

For this range of outcomes we wouldn't conclude the coin is biased, i.e. we'd "accept H_0"

For this range of outcomes we'd conclude that this number of heads was too unlikely to happen by chance, and hence reject H_0 (i.e. that coin was fair) and accept H_1 (i.e. that coin was biased).

$p(x)$
0
1
2
3
4
5
6
7
8
9
10
Number of heads (x)



P(X=0)	0.0009765625
P(X=1)	0.009765625
P(X=2)	0.0439453125
P(X=3)	0.1171875
P(X=4)	0.205078125
P(X=5)	0.24609375
P(X=6)	0.205078125
P(X=7)	0.1171875
P(X=8)	0.0439453125
P(X=9)	0.009765625
P(X=10)	0.0009765625

In this case including the P(X=8) would take us over the 5% mark, so it is not in the critical region.

- This red region is the **critical region**. The **level of significance** is normally set at 5%.
- The **actual significance level** is when you add up the heights of the 2 red bars to find the real probability of being in the critical region. (In this case it is 0.009765625 + 0.0009765625 = 0.01074218 [just over 1% actual chance]).
- A **one-tailed test** explores only extreme highs, or extreme lows.
- A **two-tailed test** explores both extremes, but for a significance level of 5%, the critical region is 2.5% on each end of the values.
- You may be asked to test if there is a correlation between 2 variables, having been given the **PMCC** (e.g. r = 0.441) as an observed value. You then see if this 0.441 is in the critical region.
  - If we want to test if 2 variables are independent, then r=0 is the null hypothesis.
  - There will be a "True" underlying correlation between 2 variables (which described how dependent they are on another – how much they affect on

another), and your "observed PMCC" is very unlikely to be exactly the 'true', 'cosmic' value determined by the laws of the universe (and even if it was we wouldn't know). The point I'm coming to, is that the more pieces of data you collect when you calculate your 'observed PMCC', the closer you get to the 'true', underlying correlation between the 2 variables. Example: if 2 variables are truly independent, then the more data we collect, the closer the PMCC will get to 0.0000.

There are 4 steps to a successful Hypothesis Test:

1. Define test statistic X (stating its distribution), and the parameter p.
2. Write null and alternative hypotheses $H_0: p = 0.5, H_1: p > 0.5$
3. Determine if you're observed outcome is in the critical region or not.
    a. The 'p-value' is a way of doing this. It is when you find the probability of getting your observed value of 'more extreme' according to the distribution in step 1 (e.g. There is a 5.5% chance of getting '8 or more heads' in the example above. This is more than 5% so if we get 8 heads in our experiment, then we are not in the critical region, and we accept the null hypothesis).
4. Two-part conclusion:
    a. Do we reject H_0 or not?
    b. Put in context of original problem.

Calculating the PMCC of the orange bivariate data on your calculator:

x	y
1	3
2	6
3	5
4	8

6: Statistics

The following instructions are for the Casio ClassWiz. Press MODE then select 'Statistics'.

$y = a + bx$

We want to measure **linear** correlation, so select $y = a + bx$

Data Entry

Enter each of the x values in the table on the left, press = after each input. Use the arrow keys to get to the top of the y column.

PMCC

While entering data, press OPTN then choose "Regression Calc" to obtain r (i.e. the coefficients of your line of best fit and the PMCC). a and b would give you the y-intercept and gradient of the regression line (but not required in this chapter).

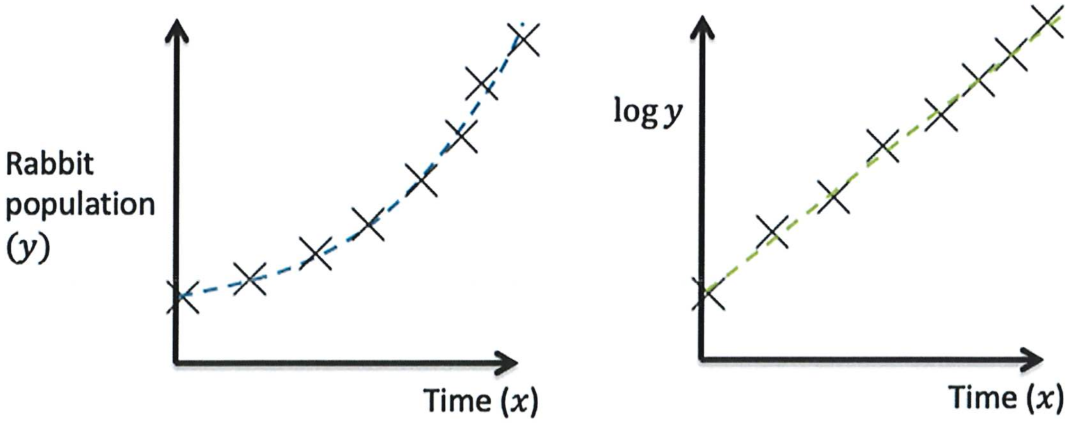
Pressing AC allows you to construct a statistical calculation yourself. In OPTN, there is an additional 'Regression' menu allowing you to insert r into your calculation.

You should obtain $r = 0.868$

Exponential modelling

You may be asked to deal with exponential models. You just log the dependent variable, then it forms a LINEAR regression line, and you deal with it as usual.

If $y = kb^x$ for constants k and b then $\log y = \log k + x \log b$



Comparing the equations, we can see that if we log the y values (although leave the x values), the data then forms a straight line, with y-intercept $\log k$ and gradient $\log b$.