Transfinite Research

Transfinite Top Tips for Edexcel A-Level Maths

Free Sample

There are two kinds of people.
1) Those who can extrapolate from incomplete data

Fourth edition, October 2019

Data Collection

POPULATION, CENSUS, SAMPLE

A **population** is the **whole set of items** that are under consideration, for example:

- All the students at this school
- All the Dark Chocolate Hobnobs produced by McVities in 2017

There are two main ways to find something out about a population.

1) Carry out a **census** which observes or measures **every single member** of a population.

2) Take a **sample** (which is just **part of the whole population**) and measure or observe only the members of the sample.

	Advantages	Disadvantages
Census	• The result has the potential to be completely **accurate** (apart from people giving blank or false responses)	• **Expensive** to carry out • Takes a **long time** to collect and process all the data • **Can't be used** if the test **destroys** the item – you'd lose everything
Sample	• **Cheaper** to carry out • **Quicker** to collect and process data • Enables **more questions** to be asked / **more characteristics** to be tested (since there are fewer items in the sample). • **Okay** if the test is **destructive** (bring on the Hobnobs…)	• **Less accurate** than a census, especially as the sample gets smaller • The sample may **not represent** small **subgroups** of the population

If we are going to take a sample of a population, we usually need a **complete list of the population** (e.g. in a spreadsheet) so we can then select our sample from it.

This list is called the **sampling frame**.
[A database of all students; a list of laptop serial numbers.]

The **individual units of a population** (the items that are in our sample frame) are called **sampling units**. [Each pupil, each meerkat that you are going to compare, etc.]

TYPES OF SAMPLING

There are many different methods for choosing a sample from the population. Some are quick, easy and not very good while others are much fairer but more complicated to do.

Many sampling methods involve **generating a random number between 1 and N**.
This can be done using a **calculator**, a **computer** or a **table of random numbers** (simply ignore any numbers in the table that are too big).

Simple Random Sampling
Each element has an equal chance of being selected

- Get a **sampling frame** and **number every item** uniquely from 1 to N
- Generate a **random number** between 1 and N
- **Select** the corresponding item as long as it has **not already been picked**
- **Repeat** the previous two steps until you have **picked enough items** for your sample

*Alternatively, **draw numbers from a 'hat'** – this automatically prevents repeats*

Systematic Sampling
Elements are chosen at regular spaced intervals starting at a random position to avoid selecting clusters, e.g. two children from the same family

- Get a **sampling frame** and **number every item** uniquely from 1 to N
- **Divide** the **population size** N **by** the required **sample size** to get the **spacing** between selected samples – call this D
- Generate a **random number** between 1 and D as the **starting** point, and pick this item
- **Count down** by D each time to **select** the **next item** until you reach the end of the list

Stratified Sampling
A random sample is taken from various sub-sections (strata) of the population to ensure that they are all fairly represented

- **Sort** the **sampling frame** into non-overlapping groups (**strata**), e.g. different age groups
- **Number** the **items** in each stratum
- Take a **random sample** (see above) from each stratum so that the **sample size is proportional to the stratum size**:
  samples per stratum = (number in stratum/number in population) × overall sample size

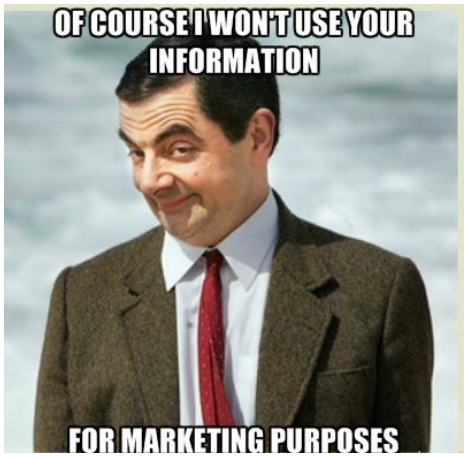
Opportunity (or Convenience) Sampling
Sample the first people/items you come across that meet your criteria, up to some limit

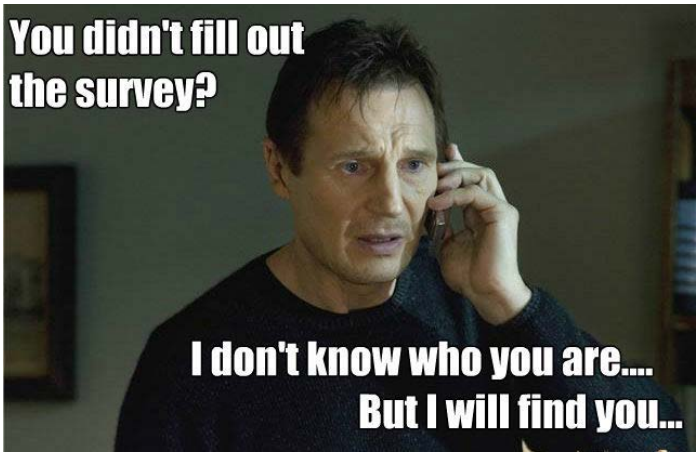
- Ask the **first person who comes along** who belongs to the population you are studying (e.g. supermarket shoppers)
- Ask the **next person**, and the next, until you have **reached** your **sample size**

Quota Sampling
Split the population into strata and then use opportunity sampling on each stratum

- **Split** the **population** according to some **criteria** and choose sample sizes (**quotas**) which are **proportional** to the **strata sizes**
- **Test the items** as they come in until you have **filled** your **quota** for each **stratum**

	Advantages	Disadvantages
Simple Random Sampling	• **Easy** and **cheap** to carry out • Suitable for **small samples** and small **populations** • Free from bias	• **Not suitable** for a **large** population • A **sampling frame** is needed
Systematic Sampling	• **Easy** and **cheap** to carry out • Only **one random number** needed and no checking for duplicates • Suitable for **large samples** and large **populations** • **Avoids** picking **adjacent clusters** by accident	• **Errors** if there is an **underlying pattern** in the data (e.g. sample every 12th month) • A **sampling frame** is needed
Stratified Sampling	• The sample **more accurately** reflects the population as a whole • Subgroups (strata) are **fairly represented** • Can **compare** different **groups** within a population	• The sampling frame must be **sorted** into strata • Selection within each stratum has the **same drawbacks** as for Simple Random Sampling
Opportunity Sampling	• **Easy** to carry out • **Quick** and **cheap** • **No sampling frame** required – perhaps you don't have access to one	• **Non-random** sampling is susceptible to **bias** (e.g. location, time of day) • **Depends** on the individual **researcher**
Quota Sampling	• **Easy** to carry out • **Fairly quick** and fairly **cheap** • **No sampling frame** required • A **small** sample can more **fairly represent** the **population** • Can **compare** different **groups** within a population	• **Non-random** sampling is still susceptible to **bias** within each stratum • The population must be **divided** into **groups**, which is more costly or inaccurate • **Non-responses** don't show up





Representing Data

OUTLIERS

An **outlier** is an **extreme data value** that lies outside the overall pattern of the data.
There is no official, fixed, standard way of defining what is and what isn't an outlier.

We usually say that a data value is an outlier if it is

- more than k interquartile ranges beyond either quartile, where k is typically 1.5.
  E.g. a value greater than $Q_3 + 1.5\times(Q_3 - Q_1)$ or less than $Q_1 - 1.5\times(Q_3 - Q_1)$.
- more than k standard deviations away from the mean, where k is typically 2.
  E.g. a value greater than $\bar{x} + 2\sigma$ or less than $\bar{x} - 2\sigma$.

The question will always tell you what value of k to use…

It's usually easiest to **work out** the **lowest** and **highest values** that **aren't outliers**, and then see if any of your data points lie outside this range – if so, they are outliers.

Example:
Find any outliers in the list 3, 5, 6, 13, 14, 15, 31.
(Here, an outlier falls more than 1.5 times the interquartile range above the upper quartile or below the lower quartile.)

IQR = 15 – 5 = 10. **1.5 × IQR** = **15**.
Lower limit = 5 – 15 = **–10**,
Upper limit = 15 + 15 = **30**.

So the only outlier is **31**.

Why might we want to **identify outliers** in a data set?

- They might be **incorrect readings** from an experiment (an **anomaly**).
- They might be **false/mistaken replies** given by people in a survey (an **anomaly**).
- They might be **unusual** but **valid results** that need acting upon, e.g. a pupil who got a very high or low score in an exam and is in the wrong set.

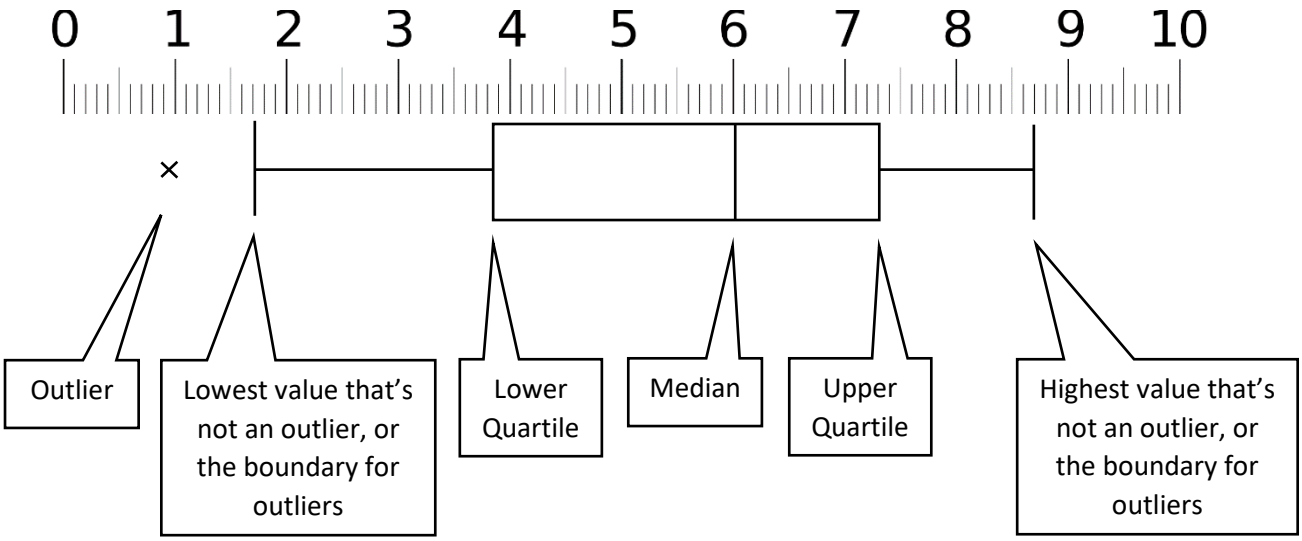
What can you do with the outliers you find?

- **Remove** them from the data set if they are **anomalies** – this is **cleaning the data**.
  You may then need to **recalculate** the mean and standard deviation with the cleaned data.
- **Keep them** in the data set if they are **valid results**.

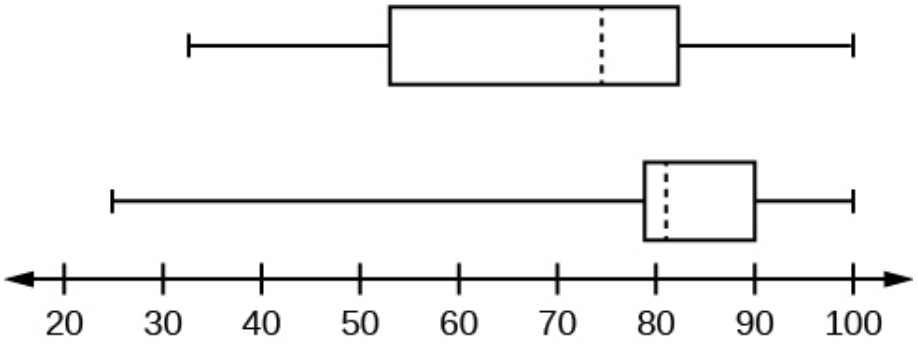
BOX PLOTS

A **box plot** shows the **important features** of a data set in a **simple graphical form** on a numerical scale.

It shows the **quartiles**, the **median**, the **lowest** and **highest values** and any **outliers**.



This is a **useful format** for **comparing two distributions** easily and clearly.



If you are comparing two distributions, you will often be asked to **write a comment** about the '**location**' or **average value** (median or mean) and the **spread** (IQR or standard deviation).

You should refer to the context of the question when you do this, for instance

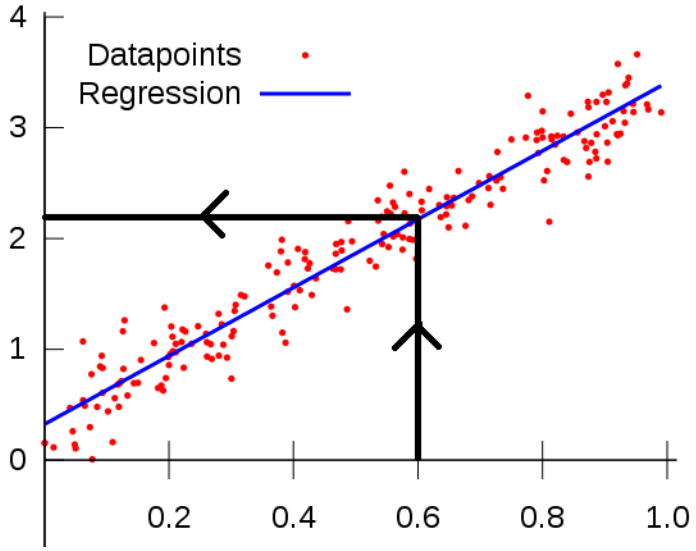
The median weight of Guatemalan weasels is slightly lower than the median weight of Bolivian weasels. While the weights of Guatemalan weasels have a much smaller interquartile range than their Bolivian counterparts, they also have a greater maximum weight owing to some significant outliers at the top end of the scale.

(This corroborates recent reports of a growing obesity problem in the Guatemalan weasel population because of excessive banana consumption.)

MAKING PREDICTIONS

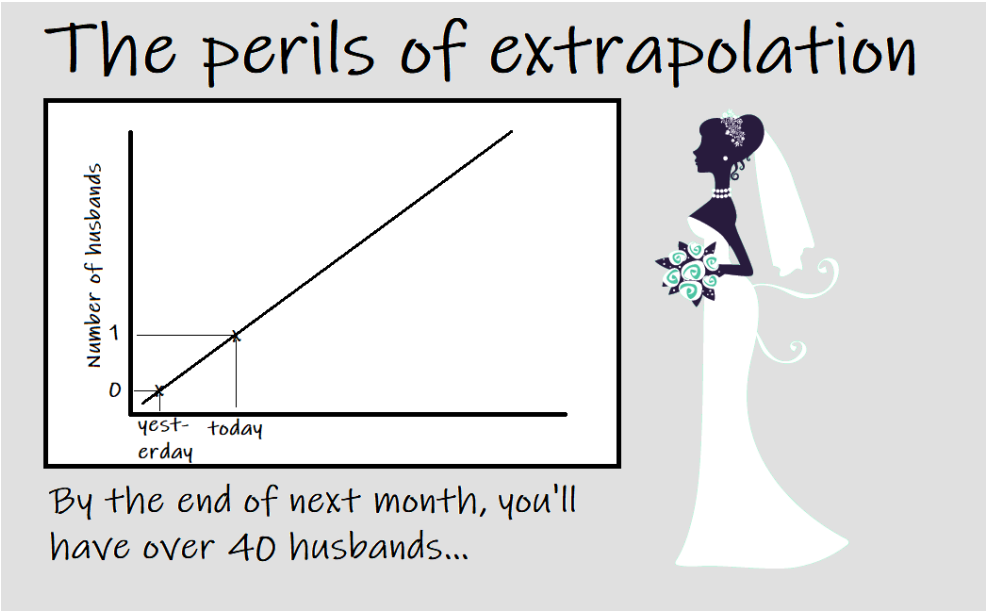
You can use the regression line to make predictions about the data, as long as:

- You are starting with the independent variable and **predicting** the **dependent variable** (finding *y* from *x*, not *x* from *y*).
- You are **predicting within the range of the given data** – you are **interpolating** and not **extrapolating**. Extrapolating is where you extend the regression line beyond the data points – this will tend to be inaccurate because you don't know that the relation is still true for larger or smaller values than you actually have.



Here we can predict that if $x = 0.6$, $y \approx 2.2$.

This is valid because we are **predicting *y* from *x*** and we are **within the data range**.



If you *did* want to **predict the independent variable** instead, you would need to use the **regression line of *x* on *y***.

FINDING BINOMIAL VALUES ON A CALCULATOR

There are various types of question we get involving binomial distributions $X \sim B(N, p)$.

- **Finding $\boldsymbol{P(X = x)}$**

  On your ClassWiz calculator, **press** [MENU] [7] [4] [2] to select Distribution – Binomial PD (Probability Distribution) – Variable.

  Enter the values of x, N and p needed in your question, **pressing** [=] after each number. **Press** [=] **again** to see the value of $\boldsymbol{P(X = x)}$.

- **Finding $\boldsymbol{P(X \leq x)}$**

  Press [MENU] [7] [CursorDown] [1] [2] to select Distribution – Binomial CD (Cumulative Distribution) – Variable.

  Enter the values of x, N and p, pressing [=] after each number.

  Press [=] **again** to see the value of $\boldsymbol{P(X \leq x)}$.

- **Finding $\boldsymbol{P(X \geq x)}$** = $1 - P(X \leq x - 1)$
  [*'4 and above' is the opposite of '3 and below'*]

  Press [MENU] [7] [CursorDown] [1] [2] to select Distribution – Binomial CD (Cumulative Distribution) – Variable.

  Enter the values of $(x - 1)$, N and p, pressing [=] after each number.

  Press [=] **again** to see the value of $\boldsymbol{P(X \leq x - 1)}$ and now work out $\boldsymbol{P(X \geq x)} = 1 - P(X \leq x - 1)$.

- **Finding $\boldsymbol{P(x \leq X \leq y)}$** = $P(X \leq y) - P(X \leq (x - 1))$
  [*'between 4 and 7' is the same as '7 and below' – '3 and below'*]

  Press [MENU] [7] [CursorDown] [1] [2] to select Distribution – Binomial CD (Cumulative Distribution) – Variable.

  Enter the values of y, N and p given in your question to find $\boldsymbol{P(X \leq y)}$.

  Then **enter the values of $(x - 1)$, N and p** to find $\boldsymbol{P(X \leq (x - 1))}$.

  Finally work out $\boldsymbol{P(x \leq X \leq y)} = P(X \leq y) - P(X \leq (x - 1))$.

  If they ask for $P(x < X < y)$, you find $P(X \leq (y - 1)) - P(X \leq x)$ instead…

Binomial Hypothesis Testing

NULL AND ALTERNATIVE HYPOTHESES

Hypothesis testing involves making a **hypothesis** (an unverified statement) about the value of a **population parameter**.

In this section, the **parameter** is always the **probability *p*** in a binomial distribution B(N, *p*).

You can then **test** your hypothesis by **carrying out an experiment** or **taking a sample** from the population.

The result of the experiment or the statistic calculated from the sample is the **test statistic**.

To do this, you need:

- The **null hypothesis H_0**, which is the hypothesis you assume to be correct.
  H_0 is the boring option where nothing unusual, dodgy or amazing is going on…
  Here, this is always $p = k$ (a fixed value given in the question).
- The **alternative hypothesis H_1**, which tells you about the parameter if the assumption is wrong.
  H_1 is the new, improved, save-the-world (or decidedly dodgy) option where a new washing powder really gets clothes cleaner or a new medicine cures more patients or a set of dice are actually weighted to land on a six more often…
  Here, this is either $p > k$, $p < k$ or $p \neq k$.

Basically, you assume that the new washing powder **isn't** any better (in other words, stick with H_0), and see if your improved result could have easily happened by chance anyway.

Set a **significance level** of 10% or 5% or 1%, etc. which means '*how much of a fluke does my test statistic result have to be before I suspect that the null hypothesis might not be true?*'

If I roll 2 dice and get 2 sixes, that's not very unusual (1 in 36) and no cause for suspicion.

But if I roll 10 dice and get 10 sixes, that is roughly a 1 in 60,000,000 chance, so I'm pretty sure that I have dodgy dice.

In practice we find a **critical region** which is the set of results that are **sufficiently unusual** for us to suspect that H_0 isn't true (at the given significance level).

The **first value** which is **just inside** the critical region is the **critical value**.

The region where we **accept H_0** is the **acceptance region** - the opposite of the critical region.

We now need to tidy this up and do the maths properly…

IT'S COMPLICATED: DEALING WITH NONLINEAR RELATIONSHIPS

In the real world, not everything fits a straight-line formula.
Even so, we can sometimes **code** our data to get a linear relation out of it.

Two quantities might be related by a **power law**, for example

- Sphere volume: $V = \frac{4}{3}\pi r^3$
- Inverse square law in Physics: $F = k/r^2$

In general, we can say that $\boxed{y = ax^n}$

If we take logs on both sides, we get $\log y = \log a + n \log x$

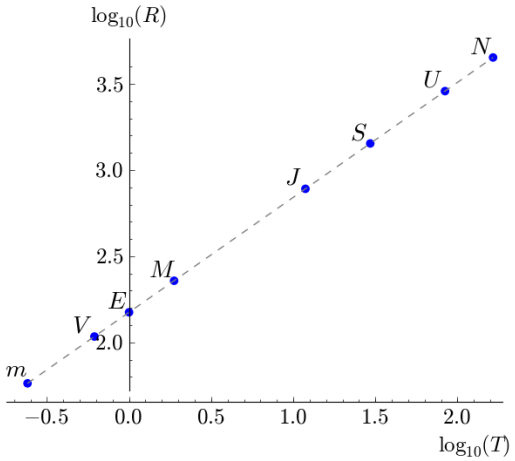
If we now **code** $\boxed{X = \log x}$ and $\boxed{Y = \log y}$ we get $\boxed{Y = \log a + nX}$

which is a linear relation with **gradient *n*** and ***Y*-intercept log *a***.

If we compare the orbit times T of the planets with the radius R of their orbits, plotting log R against log T gives a nice straight-line graph, showing that R and T are connected by a power law.

We could then use the gradient and R-intercept to get the full equation connecting R and T.

$$T = kR^{3/2}$$



Two quantities might alternatively be related by an **exponential law**, for example

- Compound interest: $Money = £1000 \times 1.03^{Years}$
- Radioactive decay: $N = N_0 e^{-kt}$

In general, we can say that $\boxed{y = k\,b^x}$

If we take logs on both sides, we get $\log y = \log k + x \log b$

If we now **code** $\boxed{X = x}$ and $\boxed{Y = \log y}$ we get $\boxed{Y = \log k + X \log b}$

which is a linear relation with **gradient log *b*** and ***Y*-intercept log *k***.

After coding, we may need to find the **pmcc** of the coded values. This value then tells us how good a fit our data points are to the original **power law** or **exponential law**.

HYPOTHESIS TESTING FOR ZERO CORRELATION

We saw hypothesis testing in the Year 1 course, and it appears several times more in Year 2.

As before, the **null hypothesis H_0** is the **boring**, '**nothing-to-see-here**' situation where in the **population as a whole**, two quantities have **zero correlation**.

$H_0: \rho = 0$

The **alternative hypothesis H_1** is that there **is** some correlation in the population:
maybe just negative or just positive [one-tailed] or maybe anything non-zero [two-tailed].

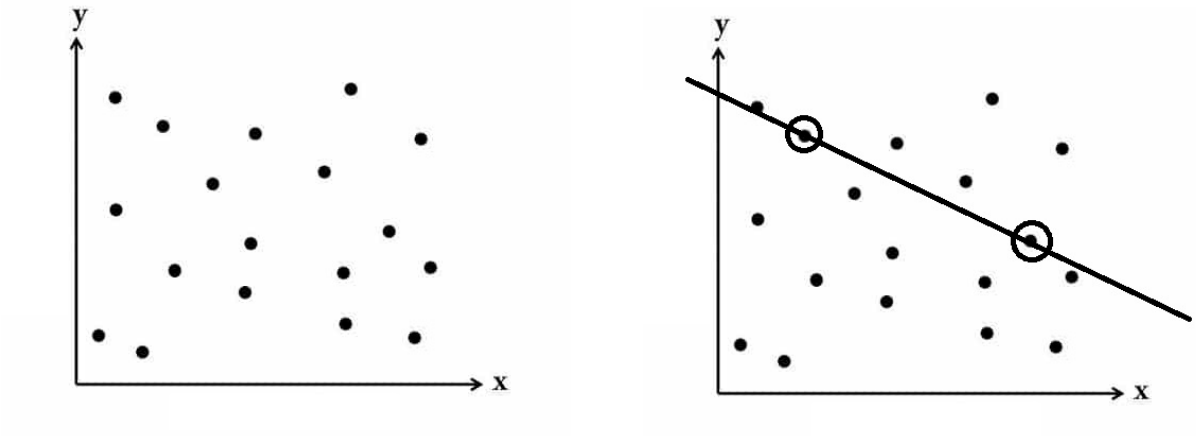
$H_1: \rho > 0$ **or** $H_1: \rho < 0$ [one-tailed]

$H_1: \rho \neq 0$ [two-tailed]

Remember that a hypothesis test is a way of drawing a conclusion about a **population parameter** (here it's ρ, the pmcc of the whole population) by analysing a **sample** drawn from it to obtain a **test statistic** (here it's the pmcc of the sample, called *r* – remember?)

The problem, of course, is that if your sample size is too small, you'll probably get a dodgy result accidentally.

Here is a population with no correlation at all, and yet my sample of two points seems to have perfect negative correlation!



But if we had picked **ten** random points and they all lay on a perfectly straight line we'd be utterly gobsmacked and flabbergasted – unless the null hypothesis wasn't actually true…

Remember that we **only reject** the **null hypothesis** if the chance of our sample pmcc being this large according to H_0 is simply **too freakishly unlikely** (at the 5% or 1% or whatever level we have chosen beforehand).

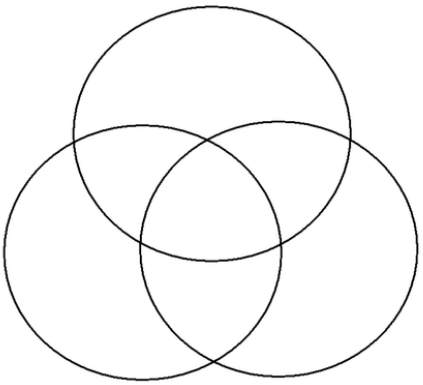
TWO-WAY TABLES VS. VENN DIAGRAMS. FIGHT!

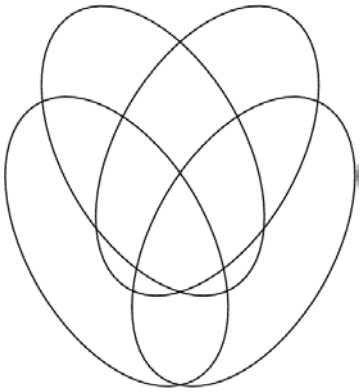
A **two-way table** is another, equivalent way of showing the number of outcomes or probabilities. Each region of the Venn diagram matches a box in the two-way table.

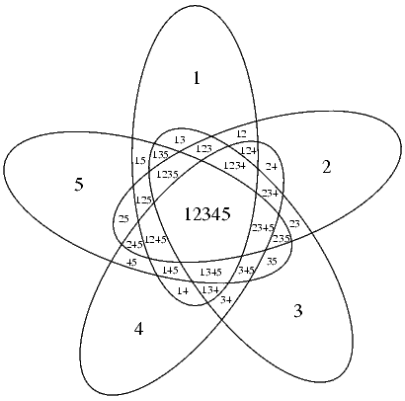
	Cats	No cats	Total
Dogs	5	7	12
No dogs	10	8	18
Total	15	15	30

Two-way tables are a little more powerful than Venn diagrams when an event has **three or more mutually-exclusive outcomes** (e.g. school years from Year 7 to Year 13), but they are **limited to** showing the outcome of **two events** (hence 'two-way').

Venn diagrams can have **three or more** overlapping shapes to show three or more events.

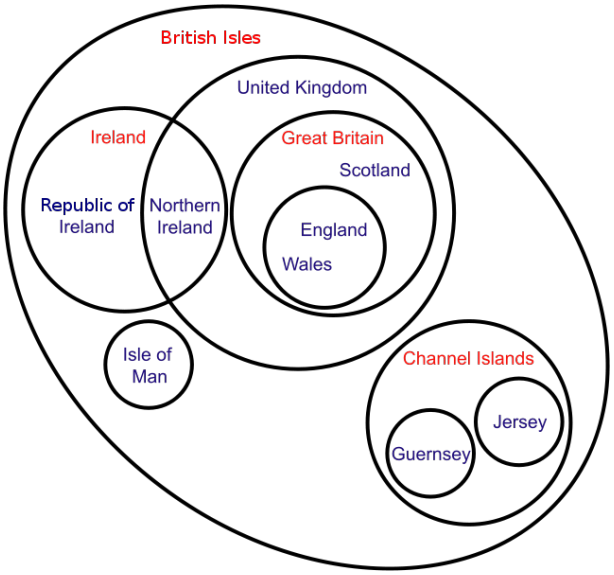


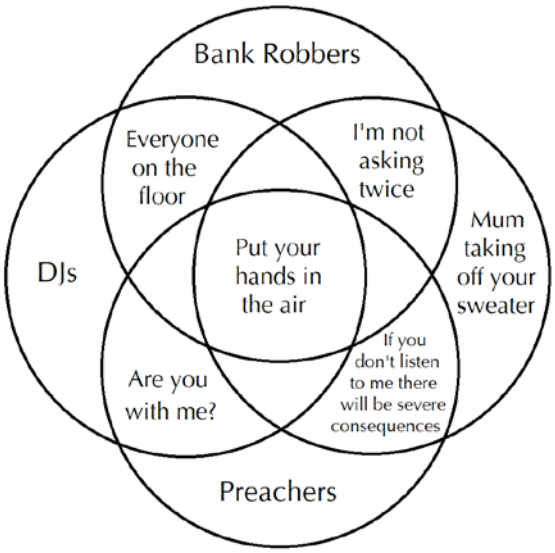




[Four overlapping *circles* do **not** make a Venn diagram: the diagram would contain only 14 of the 16 possible regions – see below.]

Then there's my usual rant: the following are **not** Venn diagrams; they are **Euler diagrams** (they don't contain every possible region).





But in GCSE and A-level examinations, we are meant to call them Venn diagrams. Grr…

CONDITIONAL PROBABILITY

The big new thing in the Year 2 course is that we need to find **conditional probabilities**.
In other words, the chance of B happening **given that** A has happened. (Or **given that** A has **not** happened.)

The good news is that you have already been dealing with conditional probability for years – in GCSE tree diagram questions.

Remember those questions about taking sweets out of a bag and **not replacing** them?
The probabilities would **change** for the second pick depending on which sweet you ate in the first pick. Those stage-two probabilities were conditional probabilities!

Some notation:

The probability of B happening if A has already happened is written $P(B \mid A)$.
You could read this as 'probability of B given A'.

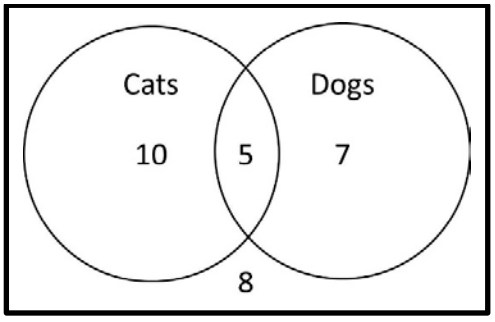
(That's a vertical bar between the B and the A, not a '1', an 'I' or a slash.)

There is a simple way to work this out:

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \quad or \quad P(B \mid A) = \frac{N(B \cap A)}{N(A)}$$

If it is given that A has happened, our 'universe' is just the Venn circle or two-way table row/column for A. Within this universe, we look for the chance of B also happening.

Back to dogs and cats:



	Cats	No cats	Total
Dogs	5	7	12
No dogs	10	8	18
Total	15	15	30

$$P(has\ a\ cat \mid has\ a\ dog) = \frac{N(has\ a\ cat\ and\ a\ dog)}{N(has\ a\ dog)} = \frac{5}{12}$$

This works in the same way for probabilities; it's all rather straightforward, really.

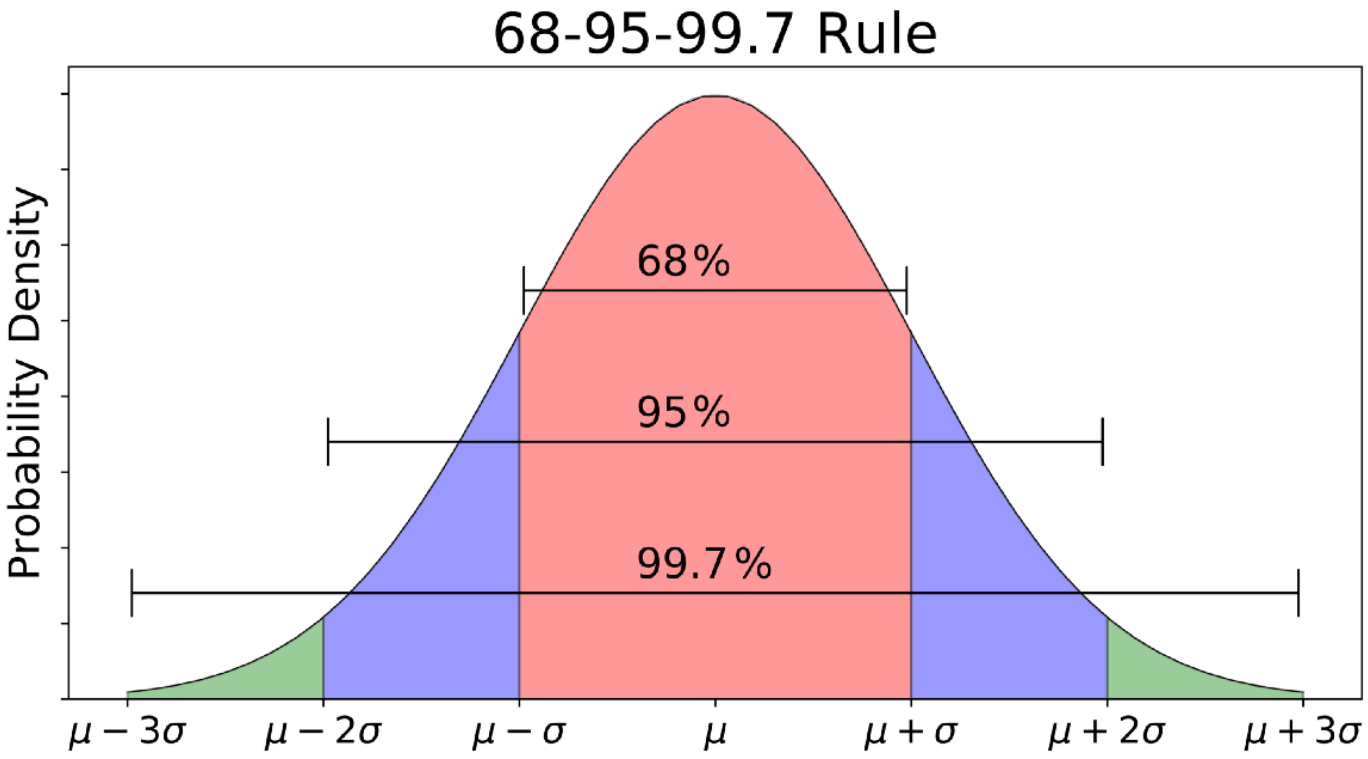
Note that if your events are **independent**, A doesn't affect B at all.

So $P(B \mid A) = P(B \mid A') = P(B)$ because B doesn't care what A gets up to…

RULES OF THUMB FOR NORMAL DISTRIBUTIONS

You need to **learn** (off by heart) the following **facts about the normal distribution** with mean μ and standard deviation σ:

- It is a **symmetrical bell-shaped** curve.
- The **total area** under the curve is **1**.
- Mean = Median = Mode = μ.
- It has **points of inflection** at **μ ± σ**.
- **50%** [half] of the values lie **above** (or below) the **mean** (because it's also the median).
- **68%** [two-thirds-ish] of the values lie **within 1 standard deviation** from the mean. [32% outside this]
- **95%** [most] of the values lie **within 2 standard deviations** from the mean. [5% outside this]
- **99.7%** [nearly all] of the values lie **within 3 standard deviations** from the mean. [0.3% outside this]



Technically, the curve extends forever to the left and right, but in practice the **probabilities** become **vanishingly small beyond 5 standard deviations** from the mean (less than 1 in a million, actually).

So if we model people's heights with a normal distribution, we don't need to worry about someone having a negative height (in Europe, the standard deviation for women's heights is 6.1 cm so most heights should be within a foot either way of the mean).

Contents Book 1:

Contents Book 2: