Medical Statistics for MRCP

1. Basic Stats

The **mean** is the average.

The **median** is the middle value, when the values are arranged in increasing order.

The **mode** is the value that appears the most often.

> Ex 1
> I have a bag of seven cats. The weights of the cats, in kg is: 2 3 3 4 7 10 13.
> The mean weight is (2 + 3 + 3 + 4 + 7 + 10 + 13) / 7 = 6kg.
> The median weight is the middle value, 4kg.
> The mode is the most common value, 3kg.

The **minimum** and **maximum** values are the smallest and largest.

The **range** is the distances between the smallest and largest value.

> Ex 1 (continued)
> The minimum weight is 2kg and the maximum 13kg.
> The range is 13 - 2 = 11kg.

The **variance** is a measure of the spread of the data. For a population this is the total of each value's squared distance from the mean, divided by the number of values.

The **standard deviation** is the square root of the variance.

> Ex 1 (continued)
> The population variance is $\sigma^2 = \frac{(2-6)^2 + (3-6)^2 + \ldots + (13-6)^2}{7} = \frac{104}{7} \approx 14.9$
> The standard deviation is therefore $\sigma \approx 3.9$

A **sample** is a subset of the population. Often it is not practical to get data on the whole population, so a sample is used instead. The sample can be used to estimate the mean and variance of the population.

To estimate the **population mean** the **sample mean** is used.

To estimate the **population variance** a slightly modified version of the sample variance is used (where the sum of squared differences from the mean is divided by one less than the number of values).

Ex 1 (continued)
Suppose there is a sample of just three cats with weights 2 3 10.
From this sample estimates can be made about the full population of seven cats.
Estimate the population mean by (2 + 3 + 10) / 3 = 5 kg.

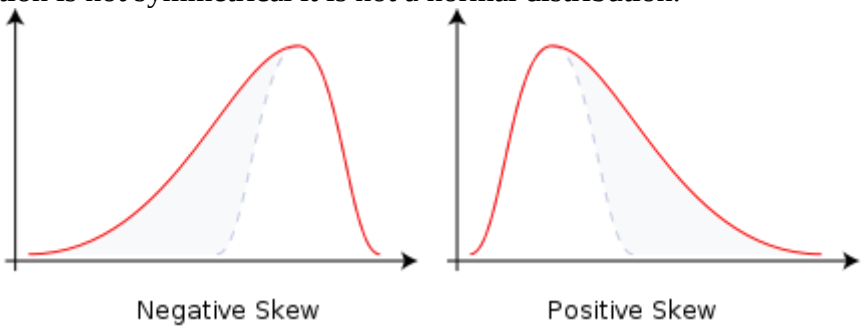
Estimate the population variance by $\sigma^2 = \frac{(2-5)^2 + (3-5)^2 + (10-5)^2}{2} = \frac{38}{2} \approx 19$

Estimate the population standard deviation by $\sigma = 4.36$

The **normal** distribution is a symmetrical bell shaped curve. On a normal distribution the value of the mean is the same as the median, both occurring at the peak of the curve.

For a normal distribution the best way to summarize the data is with the estimates of the mean and variance.

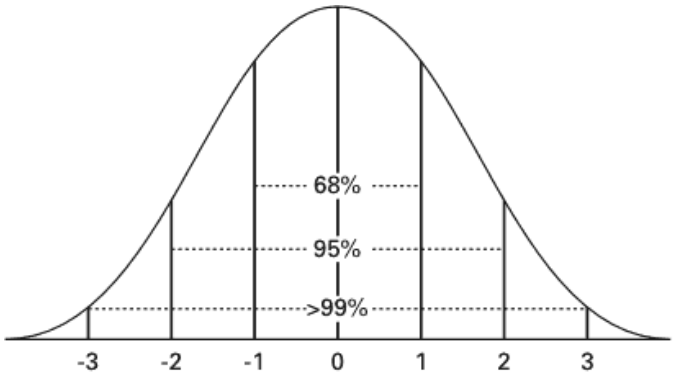
If a distribution is not symmetrical it is not a normal distribution.



A distribution with **negative skew** has some very small values, so mean < median.
A distribution with **positive skew** has some very large values, so mean > median

For these distributions the best way to summarize the data is with the median and the range (or where possible the more advanced interquartile range).

Ex1 (continued)
This data is not normally distributed. It has positive skew (mean > median) as there are some very large values (fat cats). The best way to summarize the data is by stating the median is 4kg and the range 11kg.



In a normal distribution most of the data is close to the mean.

68% (two-thirds) of all results lie within 1 S.D. (standard deviation) of the mean.
95% of all results lie within 2 S.D. of the mean.

Ex 2
In a sample of 25 bags of extra strong white flour the mass of flour was found to be normally distributed, with mean 1010g and standard deviation 10g.
Therefore 68% of all bags contain between 1000g and 1020g of flour.
Therefore 95% of all bags contain between 990g and 1030g of flour.

The **standard error of the mean** (**SEM)** is a measure of how much uncertainty there is in the estimate of the mean. It depends on the size of the sample, and the standard deviation of the sample.

$$\text{SEM} = \frac{\sigma}{\sqrt{n}}.$$

A **95% confidence interval** is a range within something is expected to be 95% of the time. The 95% confidence interval for the mean is given by

$$\mu \pm 2(SEM)$$

This is the estimate of the mean, plus or minus two times the standard error of the mean.

Ex 2 (continued)
The sample size here is $n = 25$ and the standard deviation $\sigma = 10\,\text{g}$. Hence the standard error of the mean is $\text{SEM} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$.
Thus the 95% confidence interval for the mean is $1010 \pm 2 \times 2 = 1010 \pm 4$ which is from 1006 g to 1014 g. So there is 95% confidence the average amount of flour in a bag is between 1006g and 1014g. Good news, most bags have over 1 kg.

2. Clinical Studies

A **cohort study** involves following two cohorts over time, one control group and one in which an intervention has occurred.

Ex 3
I secretly injected 25 people with mercury, and left 30 unaffected. After ten years some of them complained of death, as shown in the table below. Note that the result variable goes across the top.

	Dead Yes	Dead No
Injected Yes	10	15
Injected No	6	24

The **risk** of a treatment is the percentage of people who got the outcome with each treatment.

The **absolute risk difference** is the difference in the risks.

The **relative risk difference** is the ratio of the risks**.**
It is also sometimes called the **risk ratio.**

Ex 3 (continued)
25 people were injected and 10 died, so mercury risk is 10 / 25 = 0.40 or 40%.
30 people were not injected, and 6 died, so control risk is 6 / 30 = 0.20 or 20%.

The absolute risk difference is 0.20 or 20%.
The relative risk difference is 0.40 / 0.20 = 2. Injecting people with mercury makes them twice as likely to die within ten years.

If intervention is beneficial, then the **number needed to treat (NNT)** is the number of people who need to be treated to get one extra positive outcome.

If intervention is not beneficial, then the **number needed to harm (NNH)** is the number of people who need to be treated to get one extra positive outcome.

The NNT or NNH is calculated as the reciprocal of the absolute risk difference.

Ex 3 (continued)
The intervention here is harmful, as more people die than usual because of it. Thus the NNH, and not the NNT is appropriate here.
The NNH is $\frac{1}{0.2} = 5$. Thus on average 5 extra people need to be injected to cause one more death.

An **observational study** involves analyzing the effect of an environmental factor, comparing those who were exposed to it with those who weren't.

The **odds** of any event happening are the number of times it happens divided by the number of times it doesn't.

The **odds ratio** is the odds of getting the disease when exposed, divided by the odds of getting it when not exposed. This is approximately equal to the relative risk difference.

Ex 4
Determining if passive smoking causes lung cancer

	Lung cancer Yes	Lung cancer No
Exposed Yes	10	30
Exposed No	20	80

The odds of getting cancer as a passive smoker are 10/30 = 0.33
The odds of getting cancer as a non-passive smoker are 20/80 = 0.25
The odds ratio is 0.33/0.25 = 1.33

(Calculating the relative risk difference here gives 1.25, which is similar. In general, the relative risk difference, which uses probabilities rather than odds, makes a lot more sense to most people. Also, it's worth knowing that the odds ratio is always higher than the relative risk)

3. Statistical Significance

Suppose there are two different groups, which both have some measurable statistic *x*.

> Ex 5
> A herd of ten elephants in the wild was found to have average life expectancy of 51 years.
> A similar herd in a zoo was found to have average life expectancy of 60 years.

The **null hypothesis, H_0**, is that there is no difference between the two groups (so any observed difference is just by chance).

The **alternative hypothesis, H_1**, is that there really is an underlying difference.

> Ex 5 (continued)
> H_0 is that there is no effect on life expectancy of putting elephants in a zoo.
> H_1, is that there is an effect.
> Here the statistic of interest *x* is the difference in life expectancy years between the wild group and the group in the zoo.

In a **two-sided test** the alternate hypothesis is simply that there is a difference ($x \neq 0$).

In a **one-sided test** the alternate hypothesis is in one direction only, for example $x > 0$ or perhaps $x > 10$

> Ex 5 (continued)
> As any difference in life expectancy is of interest (without knowing in what direction the difference is), this is a two-tailed test. If some other scientists wanted to show that elephants live *longer* in captivity, that would be a one-tailed test.

Even when the null hypothesis is true, it is likely by chance that *x* will be at least a bit away from zero.

The ***p*-value** is the probability of getting the given value of *x*, given that the null hypothesis is true.

If the *p*-value is small, typically $p < 0.05$, then the results found are unlikely to have occurred if the null hypothesis is true, so the null hypothesis is rejected.

If the *p*-value is large, typically $p > 0.05$, then the results found could plausibly have occurred if the null hypothesis is true, so the null hypothesis cannot be rejected. Notice the careful language here; the null hypothesis is never declared false, merely not rejected (not rejected yet, as it might still be rejected if there is more data or a bigger study).

Ex 5 (continued)
In the elephant comparison it was found that $x = 9$ meaning captive elephants live on average 9 years longer.

Suppose after a relevant significance test this result was found to have $p = 0.10$. This means that there is an 10% chance of getting x as high as 9 by chance, given that really x is 0. This value of $p > 0.05$, so there is not (yet) enough evidence to reject the null hypothesis, so we conclude that it cannot be rejected

A **Type 1 Error (being overconfident)** occurs when a small p-value is produced and the null hypothesis is rejected, but actually this was just lucky and really there is no underlying difference.

For example, using peanut butter to cure acne might appear effective as it happened to work in a trial of ten people, although really this was just a fluke. Researchers concluding it is effective would have made a Type 1 Error (which would only come to light when a larger, more definitive, trial is done).

The probability of a Type 1 Error is denoted by **alpha** (α).
Normally the null hypothesis is rejected when $p < 0.05$, which will happen by chance 5% of the time anyway. Hence the chance of a Type 1 Error is $\alpha = 0.05$.

A **Type 2 Error (not being confident enough)** occurs when it is not possible to reject the null hypothesis, even though it is false. This could happen for example if a particular study doesn't show a significant effect, even though one does exist. This is typically because the sample is too small (as may be the case with the elephants in the example above).

The probability of a Type 2 Error is denoted by **beta** (β).
The **power** of a study is $1-\beta$, and can be thought of as the ability of the study to detect a positive result, if one exists.

Ex 5 (continued)
The original trial with 20 elephants found a difference in life expectancy of $x = 9$. This produced a p-value of 0.10, which was not significant.

Suppose a new experiment with 200 elephants still found $x = 9$.
This might produce a p-value of 0.02, which is now significant. The larger sample has greater power so results in a significant finding.

The first experiment had a small sample size, hence a low power, and failed to detect significance. It made a Type 2 Error.

Adjusting the **significance criteri**a (the p-value required to reject the null hypothesis) to make a Type 1 Error less likely makes a Type 2 Error more likely, and vice versa. You can't have it all.

Normally α is fixed at 0.05 and β is set to 0.20 or 0.10 (with corresponding power of 80% or 90%) by picking an appropriately large sample size.

An intervention is **clinically significant** if it will make a difference in real life.
This is often a different thing from statistical significance.

For example, an intervention that reduces the risk of death from a falling piano from two in a trillion to one in a trillion may be statistically significant, but not clinically significant.



4. Statistical Tests

Qualitative data is categorical or nominal data, i.e. the names of things.
The only way it can be presented is in a table (no graphs can be drawn).
Note that things like gender are qualitative, even though numbers could be assigned to them, for example 0 = male and 1 = female.

Ex 6
In the 2000s 48 men were nominated for the Academy Award for Best Director, and nine won. Two women were nominated, with one winner. Winning or losing is qualitative data.

	Won	Lost
Male	9	39
Female	1	1

[Note: Director Ang Lee is a man]

For **qualitative data** like this not much analysis can be done. But it is possible to answer the natural question of whether the two different variables are related (i.e. do nominated men win more often than nominated women).

This is typically done with a **Chi-squared test**.

For a 2x2 data table (2 rows and 2 columns) like the one above, it is usual to perform **Chi-squared with Yates Correction**. This of course is now done by computer.

A more exact (but computationally difficult) alternative is the **Fisher Exact Test**.

Ex 6 (continued)
The null hypothesis H_0 is that male and female directors that are nominated are equally likely to win (i.e. there's no gender bias).

Although women seem more likely to win (½ the nominated women win), the sample size is small and this result could well have happened by chance.

Using the Fisher Exact Test (e.g. on http://www4.stat.ncsu.edu/~boos/exact/) the *p*-value is 0.3667. This is > 0.05 so the null hypothesis cannot be rejected, and there is not enough evidence to claim a gender bias.

Quantitative data is numerical data, which can be **discrete** (taking only certain values, like the size of shoes in a shop) or **continuous** (taking any value, like the size of home made shoes).

In **parallel trials** one group is given treatment, the other group is a control group.

This gives **independent data**.

The natural question is if one group has a higher average value than the other group.

If the data is normally distributed a **t-test** is used.
If it is not normally distributed a **Mann-Whitney-U test** is used.

> Ex 7
> 20 Doctors are in an unheated hospital, and 20 a normal roasting hot hospital. The number of cups of tea each group takes is normally distributed. What test should be used to determine if the average number of teas drunk is more in the cold hospital?
> (Answer: use a t-test as data is normally distributed).

> Ex 7 (continued)
> Suppose a few Doctors drink an excessive amount of tea, so the data is no longer normally distributed (it has positive skew, hence a big tail out the right hand side). Now what test should be used?
> (Answer: use a Mann-Whitney U-test as data is not normally distributed).

In **crossover trials** the same group is given both treatments, one after the other in time.
This gives **dependent data**, also called **paired data**.

The natural question is if the group has improved between the two measurements.

If the data is normally distributed a **paired t-test**.is used.
If it is not normally distributed the **Wilcoxon signed rank test**.

> Ex 8
> Fifty subjects were given a fizzy drink and the number of burps measured. After a suitable wash-out period the same thing was done with another fizzy drink. Some subjects started with Coke then had Pepsi, some started with Pepsi then had Coke.
>
> If the burps in each case were normally distributed, what test should be used to determine if the average number of burps is different with Coke and Pepsi?
> (Answer: The paired t-test of course).

> Ex 8 (continued)
> If the number of burps was not normally distributed, what test should be used?
> (Answer: The Wilcoxon signed rank test)

5. Correlation

The **correlation coefficient** is the strength of the linear relationship between two numerical variables.

It can be expressed as a number *R*, where $-1 \leq R \leq 1$.

If ***R* = 1** the two variables have perfect positive correlation.
For example, height in cm and height in inches.

If **R > 0** there is some correlation.
For example, a child's height in cm and weight in kg.

If ***R* < 0** the variables have negative correlation. When one goes up the other goes down.
For example, the distance from the North Pole and the distance from South Pole.

If ***R* = 0** the variables are completely uncorrelated.
For example, the time in weeks since bathing and the number of skittles in your pocket.

Correlation does not imply causation. Two correlated variables might both be being influenced by a third variable, called a **confounding variable.**

For example, there is a correlation between long life expectancy and reading the FT, but this doesn't mean that reading the FT will make you live longer (in fact both effects are caused by the confounding variable of being rich).

Use the **Pearson correlation coefficient** to determine if correlation between two variables is statistically significant, when either variable is normally distributed.

Use the **Spearman rank correlation coefficient** if the data is not normally distributed, or is not numerical at all.

Ex 9
Ten people are ranked by number of shoes owned and number of hats owned.
The first person in the list has the most shoes and third most hats etc.

Rank in shoes	1	2	3	4	5	6	7	8	9	10
Rank in hats	3	1	7	5	2	4	9	8	10	6

Here $R = 0.66$ (http://www.wessa.net/rankcorr.wasp) showing there is a fairly strong positive correspondence. The *p*-value is 0.04, showing this correlation is significant (as the p-value is less than 0.05).
There is a correlation between having lots of shoes and having lots of hats.

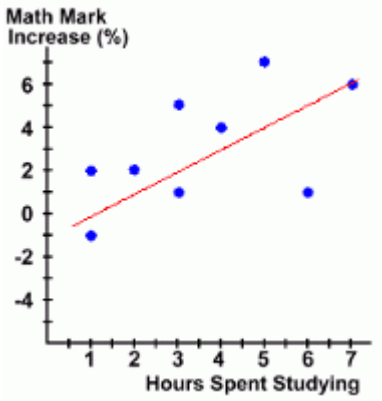
Regression analysis is an attempt to determine if two numerical variables are related.

The **independent variable** is the one varied by the researcher, shown on the *x*-axis.
The **dependent variable** is the result of the test, and is shown on the *y*-axis.

The **line of best fit** is the straight line that goes through the middle of all the points (or as close as possible). The better the line fits the points, the higher the value of *R* (hence the stronger the correlation).

The **slope** of the line is the gradient, which will be positive if $R > 0$ and negative if $R < 0$.
The **intercept** of the graph is where it crosses the y-axis, which gives the change in the dependent variable if there is no change in the independent variable.

Ex 10
There is a positive correlation ($R = 0.6$) between hours spent studying and increase in Maths mark since the last test.



The *x*-axis is the independent variable (the one controlled), of hours studying.
The *y*-axis is the dependent variable (the result), the increase in mark.
The line of best fit goes through all the points, as best as possible.

The slope of this line is 1 (meaning each extra hour gives 1% increase in mark).
The intercept is -1 (meaning that studying 0 hours gives a -1% improvement on mark)

6. Diagnostic Tests

A diagnostic test determines whether or not someone has a disease.
However, tests are not 100% accurate.

A **false positive** is when a test returns a positive result when the person is disease free
This is similar to a type-1 error (being too confident).

A **false negative** is when a test returns a negative result when the person has a disease.
This is similar to a type-2 error (not being confident enough).

Ex 11
100 people are tested for killer sneezes. 10 actually have it, 1 of whom is not detected by the test. Of the 90 who do not have the disease 18 still show up as positive (it's not a great test).
All this information is shown in the table. Notice as usual the result variable goes along the top (as in Ex 3).

	Disease Yes	Disease No
Test Yes	9	18
Test No	1	72

In a perfect test there would only be numbers in the top left and bottom right cells. The other entries are when the test doesn't work well.

Here there are 18 false positives, and 1 false negative.

The **sensitivity** of a test is how sensitive it is to people with the disease.
To find the sensitivity calculate the proportion of accurate results in the "Disease Yes" column of the table.

The **specificity** of a test is how specific it is at only detecting the disease (and not giving a positive result when there shouldn't be one).
To find the specificity calculate the proportion of accurate results in the "Disease No" column of the table.

The **positive predictive rate (PPV)** is the proportion of positive tests that actually have the disease.
To find the PPV calculate the proportion of accurate results in the "Test Yes" row.

The **negative predictive rate (NPV)** is the proportion of negative tests that are really disease free.
To find the NPV calculate the proportion of accurate results in the "Test No" row.

Ex 11 (continued)
The sensitivity, specificity, PPV and NPV are calculated for this example:

	Disease Yes	Disease No	
Test Yes	9	18	PPV 33%
Test No	1	72	NPV 99%
	Sensitivity 90%	Specificity 80%	

The test has 90% sensitivity. This means if you have killer sneezes there's a 90% chance the test will come back positive. You can think of sensitivity as how sensitive the test is at picking up the disease.

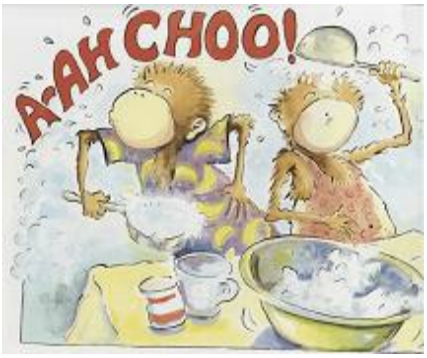
The test has 80% specificity. This means if you don't have killer sneezes there's an 80% chance of it coming back negative. You can think of specificity as how specific the test is at picking up killer sneezes. For example, a test with a low specificity would come back as positive if you had any of killer sneezes, killer coughs etc.

The test has a PPV of 33%. That means that if you get a positive test result back, there's only a 33% chance you have killer sneezes! This is very low. The reason it is low is because killer sneezes are so rare, that a positive test result is more likely to come from a faulty test on one of the many people without the disease.

The NPV is 99%. That means that if you get a negative test result back, there's a 99% chance they don't have the disease.

The example shows that the **NPV** and **PPV** depend on how common the disease is in the population. If the disease is very rare, it's hard to separate out those who really have it from the false positives, hence the PPV is low.

However, the **sensitivity** and **specificity** are not affected by how common the disease is in the population. They are properties of the test itself.



Ex 11 (continued)
Design a test for killer sneezes with 100% sensitivity.

(Answer: Simply say everyone has killer sneezes. Not a very useful test, but 100% sensitivity as everyone who does have the disease will indeed come back positive).

Ex 11 (continued)
Design a test for killer sneezes with 100% specificity.

(Answer: Simply say no one has killer sneezes. Not a very useful test, but 100% specificity as everyone who does not have the disease will indeed come back negative).

Sensitivity and specificity have **negative correlation**. The more sensitive a test is (the more likely it is to return positive if someone has the disease) the less specific it is (it's going to return positive tests far too often). Hence false positives occur.

The more specific a test is (the less likely it is to return positive if someone doesn't have the disease) the less sensitive it is (it will fail to return positive even when they do have the disease). Hence false negatives occur.