**Summary of all Statistics**

Power = 1 - the probability of a type II error

The power of a study may be defined in a number of ways:

* in general terms, the probability that a statistically significant difference will be detected
* probability of (correctly) rejecting the null hypothesis when it is false
* which also means the probability of confirming the alternative hypothesis when the alternative hypothesis is true
* power = 1 - the probability of a type II error or 1 - β

#### Study design

| **Study type** | **Key features** |
| --- | --- |
| **Randomised controlled trial** | Participants randomly allocated to intervention or control group (e.g. standard treatment or placebo)Practical or ethical problems may limit use |
| **Cohort study** | Observational and prospective. Two (or more) are selected according to their exposure to a particular agent (e.g. medicine, toxin) and followed up to see how many develop a disease or other outcome.The usual outcome measure is the relative risk.Examples include Framingham Heart Study |
| **Case-control study** | Observational and retrospective. Patients with a particular condition (cases) are identified and matched with controls. Data is then collected on past exposure to a possible causal agent for the condition.The usual outcome measure is the odds ratio.Inexpensive, produce quick resultsUseful for studying rare conditionsProne to confounding |
| **Cross-sectional survey** | Provide a 'snapshot', sometimes called prevalence studiesProvide weak evidence of cause and effect |

You are asked to design a study to assess whether living near electricity pylons is a risk factor for childhood leukaemia. What is the most appropriate type of study design?

A: Case-control study

#### Study design: new drugs

When a new drug is launched there are a number of options available in terms of study design. One option is a placebo controlled trial. Whilst this may provide robust evidence it may be considered unethical if established treatments are available and it also does not provide a comparison with standard treatments.

If a drug is therefore to be compared to an existing treatment a statistician will need to decide whether the trial is intended to show superiority, equivalence or non-inferiority:

* superiority: whilst this may seem the natural aim of a trial one problem is the large sample size needed to show a significant benefit over an existing treatment
* equivalence: an equivalence margin is defined (-delta to +delta) on a specified outcome. If the confidence interval of the difference between the two drugs lies within the equivalence margin then the drugs may be assumed to have a similar effect
* non-inferiority: similar to equivalence trials, but only the lower confidence interval needs to lie within the equivalence margin (i.e. -delta). Small sample sizes are needed for these trials. Once a drug has been shown to be non-inferior large studies may be performed to show superiority

It should be remembered that drug companies may not necessarily want to show superiority over an existing product. If it can be demonstrated that their product is equivalent or even non-inferior then they may compete on price or convenience.

####

#### Study design: evidence and recommendations

Levels of evidence

* Ia - evidence from meta-analysis of randomised controlled trials
* Ib - evidence from at least one randomised controlled trial
* IIa - evidence from at least one well designed controlled trial which is not randomised
* IIb - evidence from at least one well designed experimental trial
* III - evidence from case, correlation and comparative studies
* IV - evidence from a panel of experts

Grading of recommendation

* Grade A - based on evidence from at least one randomised controlled trial (i.e. Ia or Ib)
* Grade B - based on evidence from non-randomised controlled trials (i.e. IIa, IIb or III)
* Grade C - based on evidence from a panel of experts (i.e. IV)

####

####

####

####

####

#### Odds and odds ratio

Odds - remember a ratio of the number of people who incur a particular outcome to the number of people who do not incur the outcome

NOT a ratio of the number of people who incur a particular outcome to the total number of people

A study looks at the chance of having a myocardial infarction (MI) in patients with known ischaemic heart disease. Group A are given standard treatment. After 5 years 20 of the 100 patients have had a MI. Group B have standard treatment plus a new cardiac drug. After 5 years 10 of the 60 patients have had an MI. What is the odds ratio of having a MI whilst taking the new drug compared to those who do not?Odds of MI in group B = 10/50 = 1/5

Odds of MI in group A = 20/80 = 1/4

Odds ratio of having a MI = 1/5 divided by 1/4 = 0.8

Odds are a ratio of the number of people who incur a particular outcome to the number of people who do not incur the outcome. The odds ratio may be defined as the ratio of the odds of a particular outcome with experimental treatment and that of control.

**Odds vs. probability**

In contrast, probability is the fraction of times you'd expect to see an event in many trials. When expressed as a single number probability is always between 0 and 1. So, if we take the example of rolling a dice:

* the probability of rolling a six is 1/6 or 0.166666
* the odds of rolling a six is 1/5 or 0.2

Odds ratios are the usual reported measure in case-control studies. It approximates to relative risk if the outcome of interest is rare.

For example, if we look at a trial comparing the use of paracetamol for dysmenorrhoea compared to placebo we may get the following results

|  | **Total number of patients** | **Achieved = 50% pain relief** |
| --- | --- | --- |
| **Paracetamol** | 60 | 40 |
| **Placebo** | 90 | 30 |

The odds of achieving significant pain relief with paracetamol = 40 / 20 = 2

The odds of achieving significant pain relief with placebo = 30 / 60 = 0.5

Therefore the odds ratio = 2 / 0.5 = 4The following table highlights the main features of the main types of study:

#### ANOVA

ANOVA is a statistical test to demonstrate statistically significant differences between the means of several groups. It is similar to a student's t-test apart from that ANOVA allows the comparison of more than just two means.

ANOVA assumes that the variable is normally distributed. The nonparametric equivalents to this method are the Kruskal-Wallis analysis of ranks, the Median test, Friedman's two-way analysis of variance, and Cochran Q test

It works by comparing the variance of the means. It distinguishes between within group variance (the variance of the sample mean) and between group variance (the variance between the separate sample means). The null hypothesis assumes that the variance of all the means are the same and that within group variance is the same as between group variance. The test is based on the ratio of these two variances (known as the F statistic).

#### Data types

| **Data type** | **Description** |
| --- | --- |
| Nominal | Observed values can be put into set categories which have no particular order or hierarchy. You can count but not order or measure nominal data (for example birthplace) |
| Ordinal | Observed values can be put into set categories which themselves can be ordered (for example NYHA classification of heart failure symptoms) |
| Discrete | Observed values are confined to a certain values, usually a finite number of whole numbers (for example the number of asthma exacerbations in a year) |
| Continuous | Data can take any value with certain range (for example weight) |
| Binomial | Data may take one of two values (for example gender) |
| Interval | A measurement where the difference between two values is meaningful, such that equal differences between values correspond to real differences between the quantities that the scale measures (for example temperature) |

#### Significance tests

A null hypothesis (H0) states that **two treatments are equally effective** (and is hence negatively phrased). **A significance test uses the sample data to assess how likely the null hypothesis is to be correct.  therefore  a null hypothesis - theres NO difference - could be true, or it could be wrong. rejecting it --> i.e saying THERE IS a difference, when infact it is true creates a false positive. this is called a type 1 error. accepting the null hypothesis ---> i.e. saying THERE IS NO difference, when infact THERE IS a difference, creates a falsely negative result. this is called a type II error.**

For example:

* 'there is no difference in the prevalence of colorectal cancer in patients taking low-dose aspirin compared to those who are not'

The alternative hypothesis (H1) is the opposite of the null hypothesis, i.e. There is a difference between the two treatments

The **p value** is the probability of obtaining a result by chance at least as extreme as the one that was actually observed, **assuming that the null hypothesis is true. It is therefore equal to the chance of making a type I error** (see below).

**Two types of errors may occur when testing the null hypothesis**

* **type I:** **the null hypothesis is rejected when it is true** - i.e. Showing a difference between two groups when it doesn't exist, **a false positive  aka - there is NO difference which is what the null hypothesis said**. This is determined against a preset significance level (termed alpha). As the significance level is determined in advance the chance of making a **type I error is not affected by sample size.** It is however increased if the number of end-points are increased. For example if a study has 20 end-points it is likely one of these will be reached, just by chance.
* **type II:** **the null hypothesis is accepted when it is false** - i.e. Failing to spot a difference when one really exists, **a false negative**. The probability of making a type II error is termed beta. It is determined by both sample size and alpha

|  | **Study accepts H0** | **Study rejects H0** |
| --- | --- | --- |
| **Reality H0** |   | Type 1 error (alpha) |
| **Reality H1** | Type 2 error (beta) | Power (1 - beta) |

#### Significance tests types:

The power of a study is the probability of (correctly) rejecting the null hypothesis when it is false, i.e. the probability of detecting a statistically significant difference

* power = 1 - the probability of a type II error
* power can be increased by increasing the sample size

**The type of significance test used depends on whether the data is parametric (something which can be measured, usually normally distributed) or non-parametric**

**Parametric tests**

* Student's t-test - paired or unpaired\*
* Pearson's product-moment coefficient - correlation

**Non-parametric tests**

* Mann-Whitney U test - unpaired data
* Wilcoxon signed-rank test - compares two sets of observations on a single sample
* chi-squared test - used to compare proportions or percentages
* Spearman, Kendall rank - correlation

\*paired data refers to data obtained from a single group of patients, e.g. Measurement before and after an intervention. Unpaired data comes from two different groups of patients, e.g. Comparing response to different interventions in two groups

eg. A study is designed to see whether the degree of chest pain is linked to the troponin I value for patients admitted following a myocardial infarction. The pain is assessed using a scale of 1-10, with 10 representing the worst pain that the patient has ever experienced. Which one of the following significance tests is it most appropriate to use to investigate this link?

A: Spearman's rank correlation coefficient

A study is designed to assess the efficacy of a new anti-hypertensive medication. Two groups of patients are randomly assigned, one to take the established drug for 3 months whilst the other takes the new drug for 3 months. Blood pressure is measured before and after the intervention. There is then a period off medication for 1 month. After this period has elapsed the medication that the groups receive is swapped around and again blood pressure is measured before and 3 months later. The difference in blood pressure after the respective medications is calculated for each patient. Which one of the following significance tests is it most appropriate to apply?

Student's paired t-test

A study is designed to look at the efficacy of a mandible advancement device in reducing snoring. The severity of snoring was assessed by the partner using a 10 point scale before and after using the device. Fifty people were involved in the study. What is the most appropriate statistical test to apply to this data?

Wilcoxon signed-rank test

The data in this study is non-parametric, paired and comes from the same population. These factors make the Wilcoxon signed-rank test the most appropriate statistical hypothesis test to use.

A clinical trial is conducted to study the benefits of a new oral medication to improve the symptoms of patients with chronic obstructive pulmonary disease (COPD). In the trial 300 patients with COPD are given the new medication and a further 300 COPD patients are given a placebo. Three months later they are asked to rate their symptoms using the following scale: much improved, slight improvement, no change, slight worsening, significantly worse. What is the most appropriate statistical test to see whether the new medication is beneficial?

Mann-Whitney U test

The first point to note is that the outcome measure is not normally distributed, i.e. it is non-parametric. This excludes the Student's t-tests. We are not comparing percentages/proportions so the chi-squared test is excluded.
The Wilcoxon signed-rank test is used to compares two sets of observations on a single sample or matched samples.

Bias

Bias describes the situation in a trial where one outcome is systematically favoured. It should be noted that there is considerable variation in the definitions and classification of bias. The table below lists some of the more common types of bias.

| **Type** | **Description** |
| --- | --- |
| **Selection bias** | Error in assigning individuals to groups leading to differences which may influence outcome. Subtypes include **sampling bias** where the subjects are not representative of the population. This may be due to **volunteer bias**. An example of volunteer bias would be a study looking at the prevalence of Chlamydia in the student population. Students who are at risk of Chlamydia may be more, or less, likely to participate in the study. A similar concept is **non-responder bias**. If a survey on dietary habits was sent out in the post to random households it is likely that the people who didn't respond would have poorer diets than those who did.Other examples include * loss to follow up bias
* prevalence/incidence bias (Neyman bias): when a study is investigating a condition that is characterised by early fatalities or silent cases. It results from missed cases being omitted from calculations
* admission bias (Berkson's bias): cases and controls in a hospital case control study are systematically different from one another because the combination of exposure to risk and occurrence of disease increases the likelihood of being admitted to the hospital
* healthy worker effect
 |
| **Recall bias** | Difference in the accuracy of the recollections retrieved by study participants, possibly due to whether they have disorder or not. E.g. a patient with lung cancer may search their memories more thoroughly for a history of asbestos exposure than someone in the control group. A particular problem in case-control studies. |
| **Publication bias** | Failure to publish results from valid studies, often as they showed a negative or uninteresting result. Important in meta-analyses where studies showing negative results may be excluded. |
| **Work-up bias (verification bias)** | In studies which compare new diagnostic tests with gold standard tests, work-up bias can be an issue. Sometimes clinicians may be reluctant to order the gold standard test unless the new test is positive, as the gold standard test may be invasive (e.g. tissue biopsy). This approach can seriously distort the results of a study, and alter values such as specificity and sensitivity. Sometimes work-up bias cannot be avoided, in these cases it must be adjusted for by the researchers. |
| **Expectation bias (Pygmalion effect)** | Only a problem in non-blinded trials. Observers may subconsciously measure or report data in a way that favours the expected study outcome. |
| **Hawthorne effect** | Describes a group changing it's behaviour due to the knowledge that it is being studied |
| **Late-look bias** | Gathering information at an inappropriate time e.g. studying a fatal disease many years later when some of the patients may have died already |
| **Procedure bias** | Occurs when subjects in different groups receive different treatment |
| **Lead-time bias** | Occurs when two tests for a disease are compared, the new test diagnoses the disease earlier, but there is no effect on the outcome of the disease |

#### Confidence interval and standard error of the mean

Standard error of the mean = standard deviation / square root (number of patients)

A follow-up study is performed looking at the height of 100 adults who were given steroids during childhood. The average height of the adults is 169cm, with a standard deviation of 16cm. What is the standard error of the mean?

The standard error of the mean is calculated by the standard deviation / square root (number of patients)

= 16 / square root (100) = 16 / 10 = 1.6

The confidence interval is a common and sometimes misunderstood principle in medical statistics.

* a formal definition may be: a range of values for a variable of interest constructed so that this range has a specified probability of including the true value of the variable. The specified probability is called the confidence level, and the end points of the confidence interval are called the confidence limits\*
* in simpler terms: a range of values within which the true effect of intervention is likely to lie

The likelihood of the true effect lying within the confidence interval is determined by the confidence level. For example a confidence interval at the 95% confidence level means that the confidence interval should contain the true effect of intervention 95% of the time.

**How is the confidence interval calculated?**

The standard error of the mean (SEM) is a measure of the spread expected for the mean of the observations - i.e. how 'accurate' the calculated sample mean is from the true population mean

Key point

* SEM = SD / square root (n)
* where SD = standard deviation and n = sample size
* therefore the SEM gets smaller as the sample size (n) increases

A 95% confidence interval:

* lower limit = mean - (1.96 \* SEM)
* upper limit = mean + (1.96 \* SEM)

The above formula is a slight simplification:

* if a small sample size is used (e.g. n < 100) then it is important to use a 'Student's T critical value' look-up table to replace 1.96 with a different value
* if a different confidence level is required, e.g. 90% then 1.96 is replaced by a different value. For 90% this would 1.645

Results such as mean value are often presented along with a confidence interval. For example, in a study the mean height in a sample taken from a population is 183cm. You know that the standard error (SE) (the standard deviation of the mean) is 2cm. This gives a 95% confidence interval of 179-187cm (+/- 2 SE).

\*Last JM. A dictionary of epidemiology. Oxford: International Journal of Epidemiology, 1988

#### Incidence and prevalence

These two terms are used to describe the frequency of a condition in a population.

The **incidence** is the number of new cases per population in a given time period.

For example, if condition X has caused 40 new cases over the past 12 months per 1,000 of the population the annual incidence is 0.04 or 4%.

The **prevalence** is the total number of cases per population at a particular point in time.

For example, imagine a questionnaire is sent to 2,500 adults asking them how much they weigh. If from this sample population of 500 of the adults were obese then the prevalence of obesity would be 0.2 or 20%.

Relationship

* prevalence = incidence \* duration of condition
* in chronic diseases the prevalence is much greater than the incidence
* in acute diseases the prevalence and incidence are similar. For conditions such as the common cold the incidence may be greater than the prevalence

#### Normal distribution

A study is performed to find the normal reference range for IgE levels in adults. Assuming IgE levels follow a normal distribution, what percentage of adults will have an IgE level higher than 2 standard deviations from the mean?

2.3%

For normally distributed data 95.4% of values lie within 2 standard deviations of the mean, leaving 4.6% outside this range. Therefore 2.3% of values will be higher and 2.3% will be lower than 2 standard deviations from the mean. This figure is sometimes approximated to 2.5%

The normal distribution is also known as the Gaussian distribution or 'bell-shaped' distribution. It describes the spread of many biological and clinical measurements

Properties of the Normal distribution

* symmetrical i.e. Mean = mode = median
* 68.3% of values lie within 1 SD of the mean
* 95.4% of values lie within 2 SD of the mean
* 99.7% of values lie within 3 SD of the mean
* this is often reversed, so that within 1.96 SD of the mean lie 95% of the sample values
* the range of the mean - (1.96 \*SD) to the mean + (1.96 \* SD) is called the 95% confidence interval, i.e. If a repeat sample of 100 observations are taken from the same group 95 of them would be expected to lie in that range

Standard deviation

* the standard deviation (SD) is a measure of how much dispersion exists from the mean
* SD = square root (variance)

#### Association and causation

Two variables are said to be associated when one is found more commonly in the presence of the other.

There are three types of association.

* Spurious - an association that has arisen by chance and is not real
* Indirect - the association is due to the presence of another factor (a confounding variable)
* Direct - a true association not linked by a third (confounding) variable

Once the association has been established, the next question is whether the association is causal.

In order to establish causation, the Bradford Hill Causal Criteria (1) are used, these include:-

* Strength - The stronger the association the more likely it is to be truly causal.
* Temporality - Does the exposure precede the outcome?
* Specificity - Is the suspected cause associated with a specific outcome/ disease?
* Coherence - Does the association fit with other biological knowledge?
* Consistency - Is the same association found in many studies?

#### Screening: Wilson and Junger criteria

1. The condition should be an important public health problem

2. There should be an acceptable treatment for patients with recognised disease

3. Facilities for diagnosis and treatment should be available

4. There should be a recognised latent or early symptomatic stage

5. The natural history of the condition, including its development from latent to declared disease should be adequately understood

6. There should be a suitable test or examination

7. The test or examination should be acceptable to the population

8. There should be agreed policy on whom to treat

9. The cost of case-finding (including diagnosis and subsequent treatment of patients) should be economically balanced in relation to the possible expenditure as a whole

10. Case-finding should be a continuous process and not a 'once and for all' project

#### forrest plot

 The forest plot, below, is from a meta-analysis comparing the effectiveness of clozapine and conventional antipsychotic drugs in treatment resistant schizophrenia (negative numbers favour clozapine).

|  |
| --- |
|   |



|  |
| --- |
|   |
|   |   |



|  |
| --- |
|   |

Use the forest plot to answer the following 3 questions

|  |  |
| --- | --- |
| **39.** | How many studies show a statistically significant benefit of clozapine over conventional antipsychotics?  |



|  |
| --- |
|  4 |
| **40.** | How many studies show a statistically significant benefit of conventional antipsychotics over clozapine?  |
|  0 |  |
| **41.** | How many studies did not achieve a statistically significant result? |

|  |
| --- |
|  5 |

#### Forest plots

A forest plot (aka a blobbogram) is a graphical display of a number of results from different studies. It is the main method for illustrating the results of a meta-analysis.

|  |
| --- |
|   |



|  |
| --- |
|   |
|   |   |

|  |
| --- |
|   |

The name of the trials is listed down the left hand side, usually in chronological order. On the right hand side the results of the studies are shown as squares centred on the point estimate of the result of each trial. The size of the square is proportional to the weight of the study in the meta-analysis. The line running through the square shows the confidence interval, usually at 95%.

The large vertical line is the line of no effect. Results with confidence intervals which cross this line could potentially have mean values which are beyond this line and therefore insignificant. Beneath the individual trials is the summary result (i.e. The result of the meta-analysis) represented by a diamond.

#### Confounding

In statistics confounding refers to a variable which correlates with other variables within a study leading to spurious results.

For example

* a case-control study looks at whether low-dose aspirin can prevent colorectal cancer
* the proportion of people diagnosed with colorectal who took aspirin is compared to the proportion of people without colorectal cancer who took aspirin
* if the case and control groups are not matched for age then age could be said to be a confounding factor as older people are more likely to take aspirin and also more likely to develop cancer

In another example a study which finds that people who drink coffee are more likely to develop heart disease. The confounding factor in this study is smoking. Smoking is associated with both drinking coffee and heart disease. People who drink coffee are also more likely to smoke. In this case smoking confounds the apparent relationship between coffee and heart disease.

Confounding occurs when there is a non random distribution of risk factors in the populations. Age, sex and social class are common causes of confounding.

In the design stage of an experiment, confounding can be controlled by randomisation which aims to produce an even amount of potential risk factors in two populations.

In the analysis stage of an experiment, confounding can be controlled for by stratification.

#### Skewed distributions

Normal (Gaussian) distributions: mean = median = mode

Positively skewed distribution: mean > median > mode

Negatively skewed distribution mean < median < mode

To remember the above note how they are in alphabetical order, think positive going forward with '>', whilst negative going backwards '<'

The prevalence is the proportion of a population that have the condition at a point in time whilst the incidence is the rate at which new cases occur in a population during a specified time period.

#### Pre- and post- test odds and probability

**Which one of the following is equivalent to the pre-test probability?**

The prevalence of a condition

The prevalence is the proportion of a population that have the condition at a point in time whilst the incidence is the rate at which new cases occur in a population during a specified time period.

**Pre-test probability**

The proportion of people with the target disorder in the population at risk at a specific time (point prevalence) or time interval (period prevalence)

For example, the prevalence of rheumatoid arthritis in the UK is 1%

#### Pre-test probability

This is the probability of having the disease before a diagnostic test is done. For example a 56-year-old man who smokes comes to the surgery looking pale and clammy and complaining of a severe chest pain. The pre-test probability of him having an MI is quite high. ECGs and troponins will make this diagnosis more or less likely (change the pre-test probability).

The pre-test probability can be calculated from a two by two table (as above) like this:

Pre-test probability = (TP + FN) / (TP + FP + FN + TN)

Or

All those with the disease divided by all patients with the symptoms (both those with and without the disease).

**Post-test probability**

The proportion of patients with that particular test result who have the target disorder

Post-test probability = post test odds / (1 + post-test odds)

**Pre-test odds**

The odds that the patient has the target disorder before the test is carried out

Pre-test odds = pre-test probability / (1 - pre-test probability)

**Post-test odds**

The odds that the patient has the target disorder after the test is carried out

Post-test odds = pre-test odds x likelihood ratio

where the likelihood ratio for a positive test result = sensitivity / (1 - specificity)

#### Screening test statistics

Patients and doctors need to know if a disease or condition is present or absent. Tests can be used to help us decide. Tests generally guide us by indicating how likely it is that the patient has the condition.

In order to interpret test results we need to have a working knowledge of the statistics used to describe them.

Contingency tables (also known as 2 \* 2 tables, see below) are used to illustrate and calculate test statistics such as sensitivity. It would be unusual for a medical exam not to feature a question based around screening test statistics. Commit the following table to memory and spend time practicing using it as you will be expected to make calculations using it in your exam.

TP = true positive; FP = false positive; TN = true negative; FN = false negative

|  | **Disease present** | **Disease absent** |
| --- | --- | --- |
| **Test positive** | TP | FP |
| **Test negative** | FN | TN |

The table below lists the main statistical terms used in relation to screening tests:

| **Measure** | **Formula** | **Plain english** |
| --- | --- | --- |
| **Sensitivity** | TP / (TP + FN ) | Proportion of patients with the condition who have a positive test result |
| **Specificity** | TN / (TN + FP) | Proportion of patients without the condition who have a negative test result |
| **Positive predictive value** | TP / (TP + FP) | The chance that the patient has the condition if the diagnostic test is positive |
| **Negative predictive value** | TN / (TN + FN) | The chance that the patient does not have the condition if the diagnostic test is negative |
| **Likelihood ratio for a positive test result** | sensitivity / (1 - specificity) | How much the odds of the disease increase when a test is positive |
| **Likelihood ratio for a negative test result** | (1 - sensitivity) / specificity | How much the odds of the disease decrease when a test is negative |

Positive and negative predictive values are prevalence dependent. Likelihood ratios are not prevalence dependent.

**Precision**

The precision quantifies a tests ability to produce the same measurements with repeated tests.

Which of the following is most affected by the prevalence of a condition?

|  |
| --- |
| Positive predictive value |

Precision, sensitivity, accuracy, and specificity are not affected by the prevalence of the condition (they are stable characteristics).

The positive predictive value is low in conditions whereby the prevalence is low.. This is due to the fact that as the prevalence falls, the number of true positives falls also.

Given that

PPV = TP/(TP+FP)

One can see how low numbers of true positives decreases the size of the numerator thus lowering the PPV.

#### Confounding

In statistics confounding refers to a variable which correlates with other variables within a study leading to spurious results.

For example

* a case-control study looks at whether low-dose aspirin can prevent colorectal cancer
* the proportion of people diagnosed with colorectal who took aspirin is compared to the proportion of people without colorectal cancer who took aspirin
* if the case and control groups are not matched for age then age could be said to be a confounding factor as older people are more likely to take aspirin and also more likely to develop cancer

In another example a study which finds that people who drink coffee are more likely to develop heart disease. The confounding factor in this study is smoking. Smoking is associated with both drinking coffee and heart disease. People who drink coffee are also more likely to smoke. In this case smoking confounds the apparent relationship between coffee and heart disease.

Confounding occurs when there is a non random distribution of risk factors in the populations. Age, sex and social class are common causes of confounding.

In the design stage of an experiment, confounding can be controlled by randomisation which aims to produce an even amount of potential risk factors in two populations.

In the analysis stage of an experiment, confounding can be controlled for by stratification.

#### Graphical representations of statistical data

The table below gives a brief summary of the main types of graphs used to represent statistical data.

| **Box-and-whisker plot** | **Graphical representation of the sample minimum, lower quartile, median, upper quartile and sample maximum** |
| --- | --- |
| **Funnel plot** | Used to demonstrate the existence of publication bias in meta-analyses |
| **Histogram** | A graphical display of continuous data where the values have been categorised into a number of categories |
| **Forest plot** | Forest plots are usually found in meta-analyses and provide a graphical representation of the strength of evidence of the constituent trials |
| **Scatter plot** | Graphical representation using Cartesian coordinates to display values for two variables for a set of data |
| **Kaplan-Meier survival plot** | A plot of the Kaplan-Meier estimate of the survival function showing decreasing survival with time |



box and whisker plot



funnel plot



 histogram



forrest plot

#### Forest plots

A forest plot (aka a blobbogram) is a graphical display of a number of results from different studies. It is the main method for illustrating the results of a meta-analysis.

|  |
| --- |
|   |



|  |
| --- |
|   |
|   |   |



|  |
| --- |
|   |

The name of the trials is listed down the left hand side, usually in chronological order. On the right hand side the results of the studies are shown as squares centred on the point estimate of the result of each trial. The size of the square is proportional to the weight of the study in the meta-analysis. The line running through the square shows the confidence interval, usually at 95%.

The large vertical line is the line of no effect. Results with confidence intervals which cross this line could potentially have mean values which are beyond this line and therefore insignificant. Beneath the individual trials is the summary result (i.e. The result of the meta-analysis) represented by a diamond.

scatter plot



kaplan myer survival plot

#### Association and causation

Two variables are said to be associated when one is found more commonly in the presence of the other.

There are three types of association.

* Spurious - an association that has arisen by chance and is not real
* Indirect - the association is due to the presence of another factor (a confounding variable)
* Direct - a true association not linked by a third (confounding) variable

Once the association has been established, the next question is whether the association is causal.

In order to establish causation, the Bradford Hill Causal Criteria (1) are used, these include:-

* Strength - The stronger the association the more likely it is to be truly causal.
* Temporality - Does the exposure precede the outcome?
* Specificity - Is the suspected cause associated with a specific outcome/ disease?
* Coherence - Does the association fit with other biological knowledge?
* Consistency - Is the same association found in many studies?

#### Skewed distributions

Normal (Gaussian) distributions: mean = median = mode

Positively skewed distribution: mean > median > mode

Negatively skewed distribution mean < median < mode

To remember the above note how they are in alphabetical order, think positive going forward with '>', whilst negative going backwards '<'

Skewed distributions

* alphabetical order: mean - median - mode
* '>' for positive, '<' for negative

|  |
| --- |
| The chance that the patient does not have the condition if the diagnostic test is negative |



|  |
| --- |
|  Negative predictive value |
| **28.** | Proportion of patients with the condition who have a positive test resultSensitivity |
| **29.** | How much the odds of the disease decrease when a test is negativeLikelihood ratio for a negative test result |

# Summary of Statistics

**If the very word 'statistics' fills you with horror, read on... these pages are written with you in mind!** This section is included to help you appraise the statistics in the rest of the handbook. I do not offer it in any way as a comprehensive guide to statistics - just a basic guide from a non-statistician to help you along the way!

**The following terms are covered and explained in alphabetical order below:**

|  |  |
| --- | --- |
| * Absolute risk (AR)
* Composite endpoints
* Confidence intervals (CI)
* Forest plots
* Hazard ratios (HR)
* Interquartile range (see 'mean')
* Likelihood ratios
* Mean and median
* Non-inferiority trials
* Number needed to treat (or harm) (NNT/NNH)
 | * Odds ratios (OR)
* Positive & negative predictive value (PPV & NPV)
* Pre-test probability
* Rate ratio
* Relative risk (RR)
* Relative risk reduction (RRR)
* Sensitivity and specificity
* Systematic reviews & meta-analysis
 |

#### Absolute risk reduction (ARR)

Absolute risk reduction is difference in rate of events between the two groups

 ARR = risk of event in control group - risk of event in Rx group

 If ARR = 0 there is no difference between the two groups (no treatment effect)

Example:  Death rate in control group: 15% or 0.15

  Death rate in treatment group: 10% or 0.10

  ARR = Risk in control group - risk in treatment group = 0.15 - 0.10 = 0.05 or 5%

#### Comparing absolute and relative risk reductions

Note the differences between relative risk and absolute risk reduction in this study:

*'You could take this extra tablet, dipyridamole, twice daily for the next year and you could reduce your risk of having a further event by up 20% compared to not taking it'. (Relative risk reduction 20%).*

*'You could take this extra tablet, dipyridamole, twice daily for the next year and at the end of the year you are 1% less likely to have had an event'. (Absolute risk reduction 1%)*

Both come from exactly the same data but you can see why pharmaceutical companies prefer to use relative risk reductions!

#### Composite endpoints: beware!

Many research trials look at composite endpoints - for example in cardiovascular research a composite endpoint might include MI, need for revascularisation and cardiovascular death. However, a paper in the BMJ points out that for patients these endpoints are not equal - how can you compare dying with any other non-fatal endpoint? Dying is an altogether different magnitude event to needing a revascularisation (BMJ 2007;334:786-8). However, the paper pointed out that most cardiovascular research in recent years has been into composite endpoints. And by combining endpoints you can present skewed data to patients.

For example, one study they looked at reduced the relative risk of death by 8% and the relative risk of other more minor events by 33%. Most patients would be most interested in the reduction in their risk of death - yet the benefits to patients of this are much smaller than that of the other events. In addition, by looking at composite endpoints you are likely to over-exaggerate the benefits the patients perceive they are getting from any given treatment.

#### Confidence intervals

Confidence intervals (CI) allow you to assess statistical significance. All confidence intervals in this book are 95% confidence intervals - that is you can expect 95% of the population to fall into the range given.

If, for example, smokers got cancer 5x more often than non-smokers (N.B. I have made this figure up!), the relative risk for smoking would be 5. If the confidence intervals for this were 3.5-8 then, **because they are greater than 1*,*** you could assume this result is unlikely to have arisen by chance. However, if the 95% confidence intervals were between 0.5 and 7 it may be that, **because they cross 1**, smoking may not actually increase the risk of cancer.

#### Forest plots

There aren't any Forest plots in this Handbook, but since they come up in the MRCGP AKT exam and are commonly presented in papers you may read, we thought we'd give you a brief guide to interpreting them.

Forest plots are used to explain the results of a meta-analysis. Results from each individual trial is presented, and then combined at the bottom.

Each square represents the outcome of a trial, with the horizontal line over which it sits representing the confidence intervals. The vertical line indicates where a treatment changes from being beneficial (to the left of the vertical line in this example) to not beneficial (to the right of the vertical line in this example). At the bottom all the trials are combined and this is shown by the diamond.

In this example the diamond is to the right of the vertical line, showing no benefit from treatment. You know this is statistically significant because the confidence intervals (the horizontal line over which the diamond sits), do not cross the vertical line.

**Important!** Depending on what is being tested, the 'treatment beneficial' side is not always on the left of the vertical line, so read the labels on each forest plot carefully!

#### Hazard ratios

**Hazard ratios** are a form of relative risk (see that section). A hazard ratio of greater than 1 means an event is more likely to happen in the treatment group than in the placebo group.

#### Likelihood ratios

Likelihood ratios are useful because they incorporate sensitivity and specificity. If you want to know how they are calculated, see below. You don't need to understand likelihood ratios - just remember that they are useful because they incorporate both sensitivity and specificity.

For a positive test, the likelihood ratio of a positive test is: sensitivity/(1 - specificity)

The likelihood of a negative test is: (1 - sensitivity)/specificity

#### Mean, median and interquartile range

This may sound basic but it is important, so I have included it here. As an example let's take resting pulse rates of 7 people (65, 68, 72, 75, 78, 83, 108bpm).

**Mean:** add all the results up and divide by the number of results you had (= 549/7 = 78.4).

**Median:** line up all the numbers in order, and the median is the middle number (in this case the 4th number = 75).

**Interquartile range:** the difference between the 25th quartile and 75th quartile of data (i.e. the middle 50% of data). In this case the 25th quartile is 68 and the 75th quartile is 83, so the interquartile range is 68-83). The interquartile range is important because although it can be similar to the median, it ignores outliers that may skew data (such as the person with the pulse of 108bpm who may well have AF).

**Medians can also be quoted for interquartile ranges.** Once again, this is useful to avoid skewing of data, although in small data sets such as the example I have given you here, the result is the same as the median for the full data set.

#### Non-inferiority trials

Most trials are superiority trials: is this new drug better than this other drug or this placebo? However, sometimes non-inferiority trials are run. This is often the case when it would be unethical to offer placebo, for example if someone has H. pylori, ethically you can't really enrol them in a trial of new drug versus placebo. However, you could offer them a non-inferiority trial, testing out this new drug versus standard H. pylori eradication therapy.

**Non-inferiority trials will tell you whether your new drug is no worse than the control treatment BUT it can't tell you if it is any better** (although you can run a non-inferiority trial that tests for non-inferiority but is also sufficiently powered to detect superiority!).

There are several inherent weaknesses in non-inferiority trials, in particular that the margin for proving effect can be set nice and wide, making almost anything look effective. Also non-inferiority trials assume the standard control therapy is effective (it may not be!). In addition, intention-to-treat analysis (deemed good in superiority trials) may blur the effect for new and old treatments still further.

#### Numbers needed to treat (NNT) or harm (NNH)

NNTs tell us how many people have to be treated for 1 person to benefit. An ideal NNT is 1; everyone treated gets better, no one given the placebo group gets better. NNHs are numbers needed to harm. NNT/H should, but don't always, quote a time frame.

An NNT (or H) of 40 over 2 years means that 40 people have to be treated for one to get a benefit (or harm) over a 2-year period.

NNTs are easy to calculate: NNT = 1/ARR (absolute risk reduction)

If risk of event in treatment group: 4%, and risk of event in placebo group: 1%

 ARR is 4 - 1=3%

 NNT=1/ARR = 1/3 (x100) = 33

#### Odds ratios

An odds ratio is a way of expressing probability or relative risk - an odds ratio of greater than 1 means an event is more likely to happen in the treatment group than in the placebo group.

#### Positive & negative predictive values

The positive predictive value of a symptom or test is the proportion of the people who test positive who actually have the disease.

The negative predictive value of a symptom or test is the proportion of people told they don't have the disease that really don't have it.

Using a 2 2 table:

|  |  |  |
| --- | --- | --- |
|   | Disease present | Disease absent |
| Test positive | True positives (TP) | False positives (FP) |
| Test negative | False negatives (FN) | True negatives (TN) |

Positive predictive values (PPV) = TP/(TP + FP)

Negative predictive value (NPV) = TN/(TN +FN)

**The higher the PPV of a symptom or test, the more likely the patient sitting in front of you really does have that disease.**

**The higher the NPV of a test, the more likely it is that the patient who has tested negative, really doesn't have the disease.**

A PPV of 10% means 10% of people with that symptom will, after investigations, actually have cancer. That means 90% of people with that symptom will not.

In the Cancer chapter we discuss how low some PPVs are for classic 'red flags' for cancer and what this means in terms of our ability to detect cancers.

#### Pre-test probability

This is the probability of having the disease before a diagnostic test is done. For example a 56-year-old man who smokes comes to the surgery looking pale and clammy and complaining of a severe chest pain. The pre-test probability of him having an MI is quite high. ECGs and troponins will make this diagnosis more or less likely (change the pre-test probability).

The pre-test probability can be calculated from a two by two table (as above) like this:

Pre-test probability = (TP + FN) / (TP + FP + FN + TN)

Or

All those with the disease divided by all patients with the symptoms (both those with and without the disease).

#### Rate ratio

Rate ratio is simply the ratio of the rate of something in one population divided by the rate in another population. It is often used for comparing the incidence of a disease in a group of people exposed to a something compared to an unexposed population. For example, the rate of cancer in a population exposed to a carcinogen may be 10/hundred person years. The rate of cancer in an unexposed population might be 3/100 person years. The rate ratio would be 3.333, suggesting that those exposed to the carcinogen were 3x more likely to get cancer than the unexposed population.

#### Relative risk (RR)

How many times more likely is it that an event will occur in the treatment group compared to control group?

 RR is risk in treatment group/risk in control group

 RR of 1 = no difference

 RR <1 means treatment reduces risk of outcome

 RR >1 means treatment increases risk of outcome

Example:  Death rate in control group: 15% or 0.15

  Death rate in treatment group: 10% or 0.10

  RR is Risk in treatment group/risk in control group = 0.10/0.15 = 0.67

#### Relative risk reduction (RRR)

Tells us reduction in the rate of the outcome in the treatment group relative to control group.

 RRR = ARR/risk of outcome in control group  OR  RRR = 1 - RR

Example:  Death rate in control group: 15% or 0.15

  Death rate in treatment group: 10% or 0.10

  RRR = ARR/risk of outcome in control group = 0.05/0.15 = 0.33 or 33%

 Or  RRR = 1 - RR = 1 - 0.67 = 33% or 0.33

#### Comparing absolute and relative risk reductions

Note the differences between relative risk and absolute risk reduction in this study:

*'You could take this extra tablet, dipyridamole, twice daily for the next year and you could reduce your risk of having a further event by up 20% compared to not taking it' (Relative risk reduction 20%).*

*'You could take this extra tablet, dipyridamole, twice daily for the next year and at the end of the year you are 1% less likely to have a had an event'. (Absolute risk reduction 1%)*

Both come from exactly the same data but you can see why pharmaceutical companies prefer to use relative risk reductions!

#### Sensitivity and specificity

Sensitivity is the proportion of people with a disease who are detected by the test.

Specificity is the people who don't have the disease and don't test positive (i.e. they test negative). Using the 2 x 2 chart from before:

Sensitivity = TP/(TP + FN) E.g. you work out the proportion of cancers detected as a proportion of all the cancers. High sensitivity - good test for cancer.

Specificity = TN/(TN + FP) E.g. you work out the proportion of the people who haven't got cancer and test negative for cancer as a proportion of all those without cancer. High specificity = few false positives.

#### What is the difference between a systematic review and a meta-analysis?

Both systematically look for all the relevant literature on a subject.

A systematic review will draw together all the literature and come to conclusions in the absence of numerical data to prove an effect.

A meta-analysis is a systematic review that uses quantitative methods to summarise the results - basically the results from a number of different studies are pooled to produce a large enough sample size to reduce the risk of any finding being down to chance alone.

Of course for both systematic reviews and meta-analysis the studies need to be similar and good quality... rubbish in equals rubbish out!

|  |
| --- |
|   |
|   | Statistics* Don't be scared of statistics: most commonly used statistics can be understood relatively easily.
 |

|  |
| --- |
|   |

|  |  |
| --- | --- |
|   | For more on statistics these two websites are really good:Bandolier: [www.medicine.ox.ac.uk/bandolier](http://www.medicine.ox.ac.uk/bandolier)Centre for Evidence-Based Medicine: [www.cebm.net](http://www.cebm.net/)  |