

## **STATISTICS FOR THE FRCA**

(Written by Ian Wrench 2002)

Syllabus for statistics for the final part of the FRCA

### **Knowledge**

Candidates will be expected to understand the statistical fundamentals upon which most clinical research is based. They may be asked to suggest suitable approaches to test problems, or to comment on experimental results. They will not be asked to perform detailed calculations or individual statistical tests.

### **Data collection and analysis**

Simple aspects of study design defining the outcome measures and the uncertainty of measuring them.

### **Application to clinical practice**

Distinguishing statistical from clinical significance

Understanding the limits of clinical trials

The basis of systematic review and its pitfalls

### **Study design**

Defining a clinical research question

Understanding bias

Controls, placebos, randomisation, blinding, exclusion criteria

Statistical issues, especially sample size, ethical issues.

## INTRODUCTION:

Statistics are used for two main purposes:- i) to describe data - **descriptive statistics**  
 ii) to test for significant differences between data sets - **inferential statistics**

### Descriptive statistics:

#### i) NUMERICAL DATA

There are four main ways to describe numerical data:

1. **nominal (categorical)**: data that can only be named and put into categories with no scale between them, e.g. ABO blood groups or different types of fruit.
2. **ordinal**: data which can be put in an order from the least to the greatest but not at equal intervals, e.g. pain - nil/mild/moderate/severe.
3. **interval**: data described in terms of a numerical scale with equal intervals but with no absolute zero so that description in terms of ratio is meaningless, e.g. Celsius and Fahrenheit temperature scales.
4. **ratio**: data described in terms of a numerical scale with equal intervals and with an absolute zero so that it is possible to use ratio, e.g. measures of weight, length and the Kelvin temperature scale.

Nominal and ordinal are qualitative data whereas interval and ratio are quantitative data.

#### ii) MEASURES OF CENTRAL TENDENCY:

a) **mean** (average) - the sum of the observations divided by the total number of observations. Uses all of the data but is readily influenced by outliers (data points at the extremes of the spread of data).

b) **median** - the value exceeded by half the number of observations (e.g. if there are ten observations then the value exceeded by five of them would be the median). It is not readily influenced by outliers but does not make use of all the individual data values.

c) **mode** - the most frequently occurring score in the sample. It is not used much in statistical testing.

## i) MEASURES OF SPREAD

### **Parametric data:**

characteristics of parametric statistics:-

- a) mean = median = mode
- b) interval or ratio scale
- c) normal / gaussian distribution (figure 1)

measures of spread of parametric data:

#### i) **Standard deviation (SD):-**

( $\frac{\sum (x - \text{mean})^2}{n}$ ) = the average difference from the mean,  
 $n$  = the number of observations and  $\Sigma$  = sum of,  $n-1$  is only used if the number of observations is 30 or less, for higher numbers  $n$  alone is used)

1 SD = 68 % of the population, 2 SD = 96 % and 3 SD = 99.6 %

ii) **95 % confidence limits** = mean  $\pm$  SD x 1.96 (95 % confident that the true population mean lies within these limits)

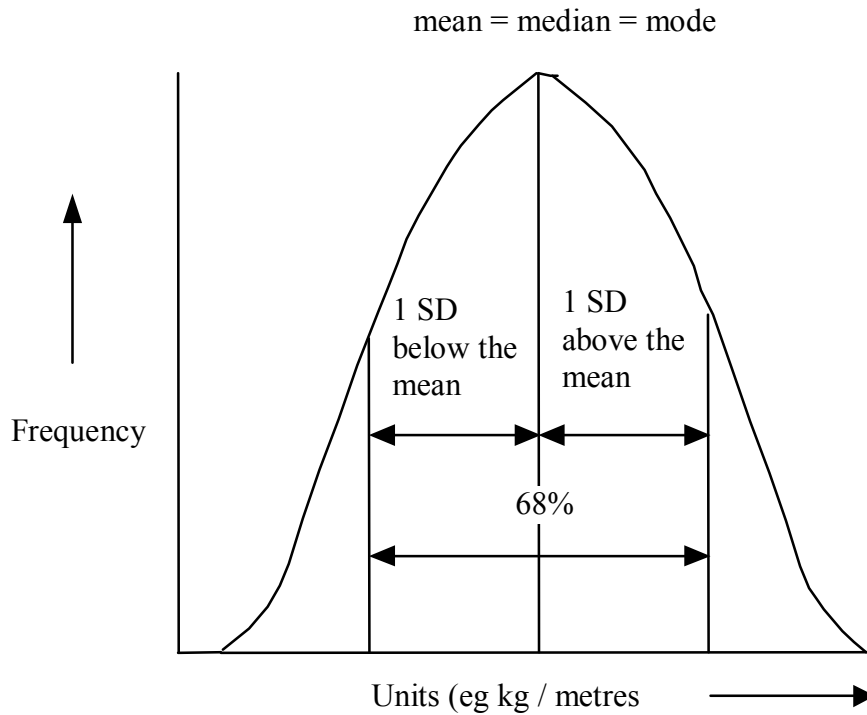
iii) **Variance** = (SD)<sup>2</sup>

**Standard error of the mean (SEM)** is a measure of how precisely the mean is estimated and not a measure of spread of the data. SEM is equal to the standard deviation divided by the square root of the number of observations, so that as the number of observations increases so the SEM becomes smaller indicating that we can be more certain that the sample mean is close to the population mean.

### **Non-parametric data:**

i) **Range** - the difference between the highest and the lowest scores. Not very efficient in the use of data and easily influenced by outliers.

ii) **Interquartile range** - the data is **ranked** (i.e. put in order from lowest to highest) and divided into quarters. The interquartile range is the range between the lowest and the highest quarters (i.e. encompassing the middle half of the data). If a dividing point is between two numbers then an average is taken. This method is less susceptible to outliers.



**Figure 1a:** An example of gaussian/ normal distribution. One standard deviation (SD) refers to one SD above and below the mean. Two SD = 96% of the population and 3 SD = 99.6%.

## Study design and statistical testing (Inferential statistics):

### STEPS IN PERFORMING A RESEARCH STUDY:

1. **Define the problem** that needs to be addressed? For example it may have been noted that the incidence of postoperative pain or nausea following a particular procedure is unacceptably high. Audit or a clinical impression may have established this.
2. Perform a **literature search** to find out what work has been done in the area of interest.
3. Form a **working hypothesis** on how to improve the problem in question – e.g. if we use a new pain killer/ antiemetic it will be better than our current treatment. By making a definite statement of intent it makes it easier to design a study as this will need to be set up to test the working hypothesis. In other words by the end of the study we should know whether or not the working hypothesis is true.
4. Define the **primary outcome**. For example in a study looking at pain this could be pain scores measured on a visual analogue scale. However, pain is difficult to measure reproducibly as one patients “10” could be another patients “5”. To get round this you could measure the amount of pain killer (e.g. morphine) needed. This can be measured much more easily but things other than pain could affect it, e.g. patients may feel nauseated by morphine and may not ask for as much as they need to control their pain. Morphine requirement is a surrogate end point, in other words it is an indirect measure of pain. Another example of this would be using antiemetic requirement instead of measuring nausea in a PONV study.
5. **Write a protocol** for the study detailing how the study will be conducted. Among other things, the protocol should make the case for why such a study would be useful and the target population with inclusion/ exclusion criteria should be described. Ideally the study should be blinded, randomised and placebo controlled:

**Randomisation** - avoids any **Eg** in the allocation of treatments to patients. It guarantees that the probabilities obtained from statistical analysis will be valid. Randomisation prevents the influence of any **confounding factors**, i.e. when the effects of two processes are not separated. For example asthmatics are less likely to have lung cancer but this is because they are also less likely to be smokers rather than a protective effect of the disease itself. The lack of randomisation with audit data means that statistical analysis must be interpreted with caution.

**Blinding** - double-blind trials are designed so that neither the observer nor the subject are aware of the treatment and thus may not **bias** the result. A trial is single-blind if the patient is unaware of the treatment. Sometimes blinding is difficult if a drug has a characteristic side-effect or surgical techniques are being compared which makes the treatment obvious.

**Controls** - this provides a comparison to assess the effect of the test treatment. In many instances the most appropriate control is a **placebo** i.e. a tablet \ medicine

with no clinical effect. If the omission of treatment is thought to be unethical (e.g. not giving an analgesic for a painful procedure) a standard treatment may be used as control e.g. a new analgesic may be compared with morphine.

6. Estimate the number of patients required to perform the study and avoid a type 2 error:

**Type 2 ( $\beta$ ) error** - finding that there is no difference when one exists.

It would take a very large number of patients to show that there is not even the smallest difference between two treatments. Whilst it may be possible to show that there is a **statistically significant** but very small difference between two painkillers (for example), if the difference was very small then this would not be **clinically significant**. For this reason, before embarking on a study the investigators should always decide what a clinically significant difference between treatments would be. Once this has been done, the number of patients required to show whether or not there is such a difference may be calculated. This process is known as estimating the **power** of the study and is done so that it is possible to have confidence in a negative result and thus avoid a type 2 error. When calculating the power of a study it may be that the test treatment could be better or worse than the control (a **two tailed hypothesis** e.g. comparing a pain killer with a standard analgesic). Alternatively, it may be possible to assume that any difference is likely to be in one direction only (a **one tailed hypothesis** e.g. comparing a pain killer with placebo). More patients are required to prove or disprove a two tailed hypothesis than a one tailed hypothesis.

7. Present the protocol to the **ethics committee**. Ethics and research is an enormous topic but in brief for a study to be ethical – i) it should have received ethics committee approval, ii) informed written consent should have been obtained beforehand, iii) the individuals rights should be preserved, iv) respect for participants confidentiality should be maintained.
8. Once ethical approval is obtained perform the study and **collect data**.
9. **Analyse the data** – in particular find whether groups differ significantly in terms of the **primary outcome** (e.g. pain scores or nausea). Statistical analysis is undertaken to prove or disprove the **null hypothesis** (the hypothesis that there is no difference between groups). A type 1 error could occur at this stage:

**Type 1 ( $\alpha$ ) error** - finding a difference when one does not exist.

When comparing two groups of data, statistical testing is undertaken to establish the probability that they are taken from two different populations, this is expressed as the **p value**. The p value may vary from 1 (the groups are the same) to 0 (100% certainty that the groups are different). Usually a value for p is obtained between these two extremes and **p < 0.05** is taken to be "statistically significant". This is an arbitrary figure which means that there is a 1 in 20 (5%) chance that there really is no difference between groups. As the p value becomes lower the possibility of there

being no difference when one has been found becomes more and more remote (e.g.  $p = 0.01$ , 1 in 100 and  $p = 0.001$ , 1 in 1000). Thus the lower the  $p$  value the less likely that a type 1 error has been made.

#### WHICH STATISTICAL TEST:

(the list contains commonly used tests and is not exhaustive)

	<b>Parametric data</b>	<b>Non-parametric data</b>
<b>Correlation</b> (association between two variables)	<i>Pearson's r</i>	<i>Spearman's rho</i>
<b>Statistical testing</b>	(comparing two sets of data to decide whether or not they come from the same population):-	
<u>Paired data*</u> (e.g. crossover trial)	<i>Paired t-test</i>	<i>Sign test or Wilcoxon's test</i>
<u>Unpaired / independent data</u>	<i>Independent t-test</i>	<i>Mann-Whitney U test or Kruskal-Wallis test</i>

For parametric tests the criteria are: a) Interval or ratio data b) data normally distributed c) equal variance of the two data sets

\*paired data is data where two observations have been made on the same group such as a crossover trial where subjects receive both treatments which are being compared. Obviously there will be the same number of observations in each set of data.

10. Draw **conclusions, publish** and plan further research if necessary.

## Correlation and contingency tables:

### CORRELATION:

Correlation is a technique to assess whether or not there is a linear association between two continuous variables (variables which may take on any value in a given range). Neither of the variables should have been determined in advance. The data is collected in pairs and plotted on a graph. A **correlation coefficient** is calculated which may range from -1 (a negative correlation) to 0 (no correlation) to +1 (a positive correlation) (figure 2). A probability is then derived for this by referring to standard tables. Such a calculation is only possible if the data conform approximately to a linear pattern. **Correlation is not equivalent to causation.**

Linear regression is similar to correlation except that one of the variables is determined in advance as with a dose response.

### CONTINGENCY TABLES

Sometimes it is necessary to compare proportions. The simplest case would be a comparison between two groups where the variable is a yes or no answer. For example, patients with lung cancer may be allocated to receive one of two treatments (A or B), and the endpoint might be whether they were alive or dead after five years. Such data may be presented in terms of a 2 x 2 table:

	Treatment A	Treatment B
Alive	150	10
Dead	50	190

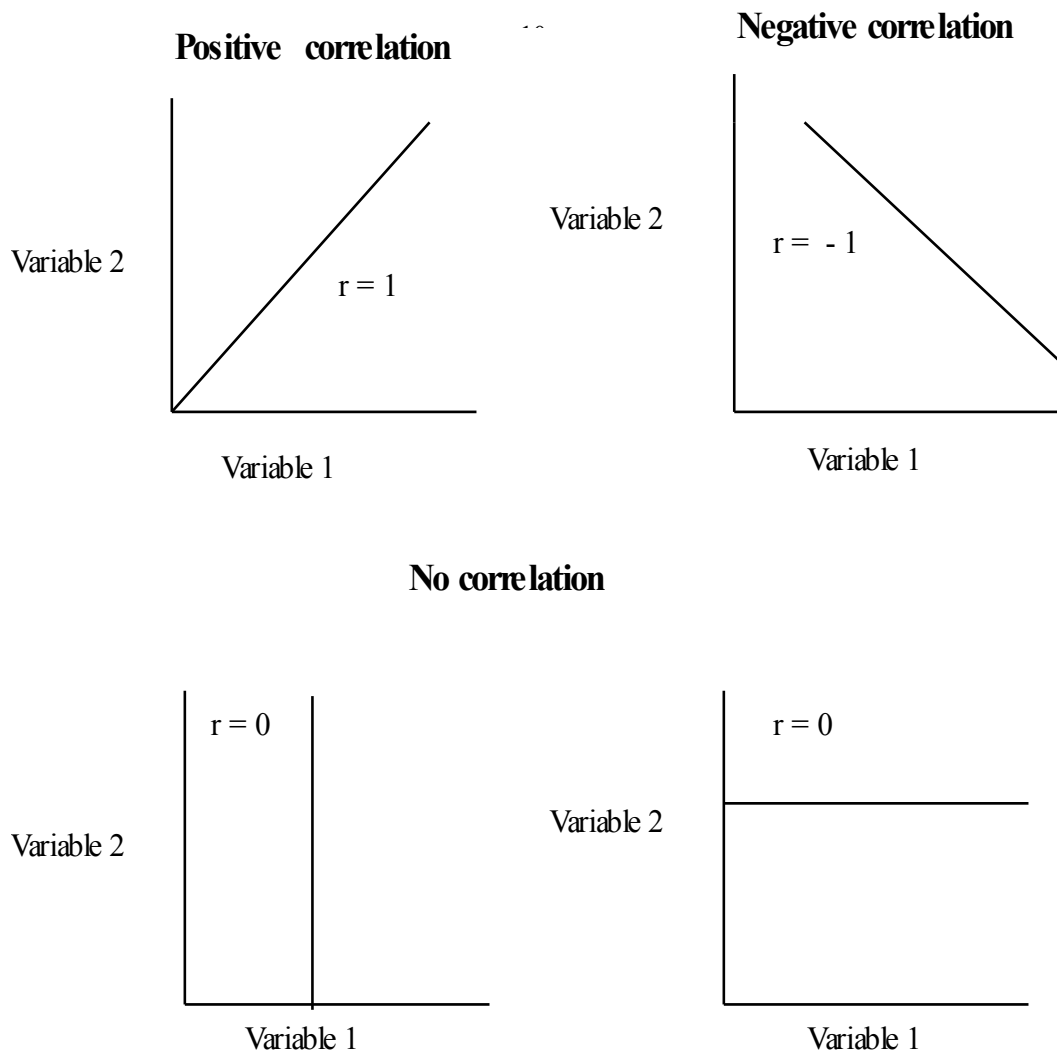
As it may be seen, there is a clear difference between the two treatments, these are called the *observed numbers*. In order to test the significance of this difference a **Chi square test** is applied. This involves calculating what the numbers would be if there were no difference between the groups, and comparing this to the actual numbers obtained. The total number who survived are added and divided by 2 as are the total number who died to give a 2x2 table of *expected numbers*:

	Treatment A	Treatment B
Alive	80	80
Dead	120	120

The observed and expected numbers are then compared using the Chi square test and a Chi square number is derived. The p value is obtained by reference to tables. This test may only be applied to the raw data so that derived data such as percentages must not be used. If the expected numbers for more than one cell of the table comes to less



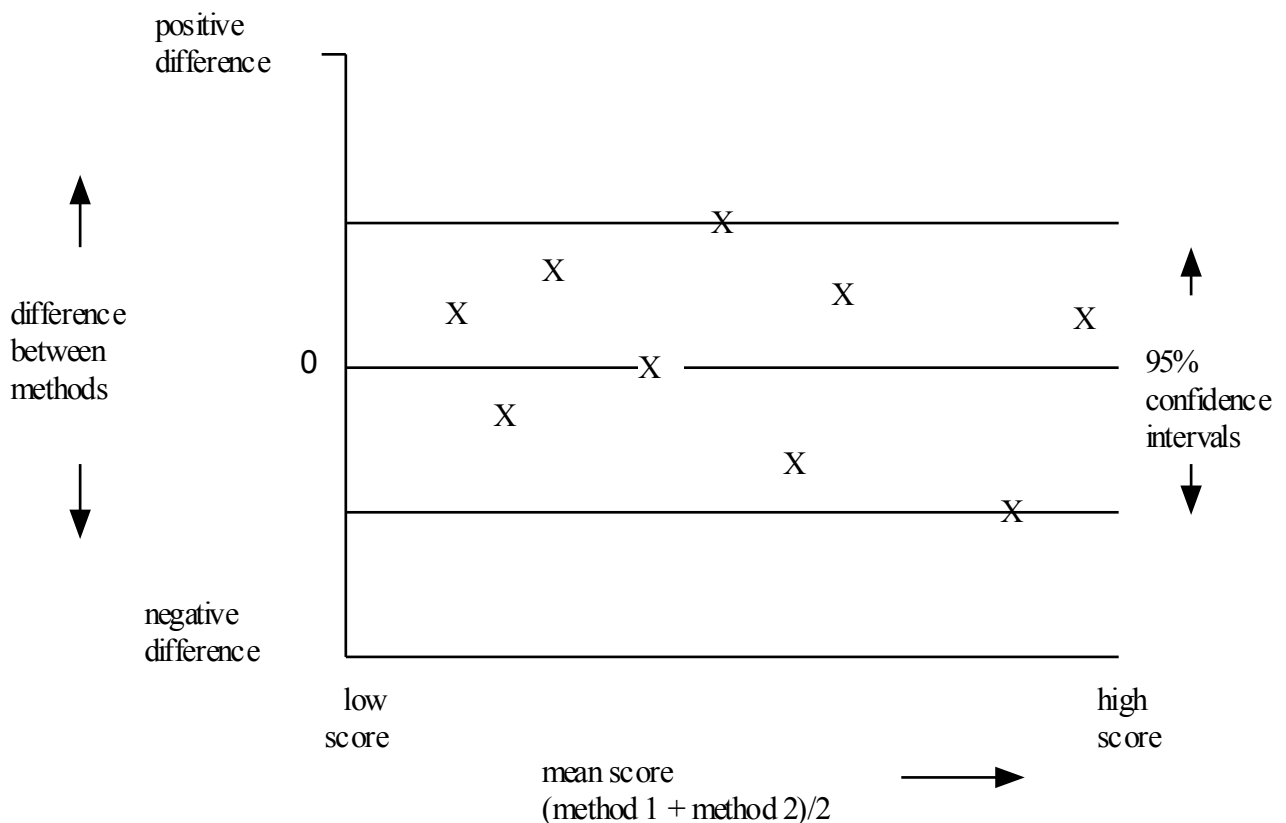
than 5 then it is necessary either to use Chi square with **Yates correction** or to use **Fishers exact test**.



**FIGURE 2:** Examples of different correlation coefficients ( $r$ ). Usually the data is scattered, but it must be approximately linear in order for a test of correlation to be valid

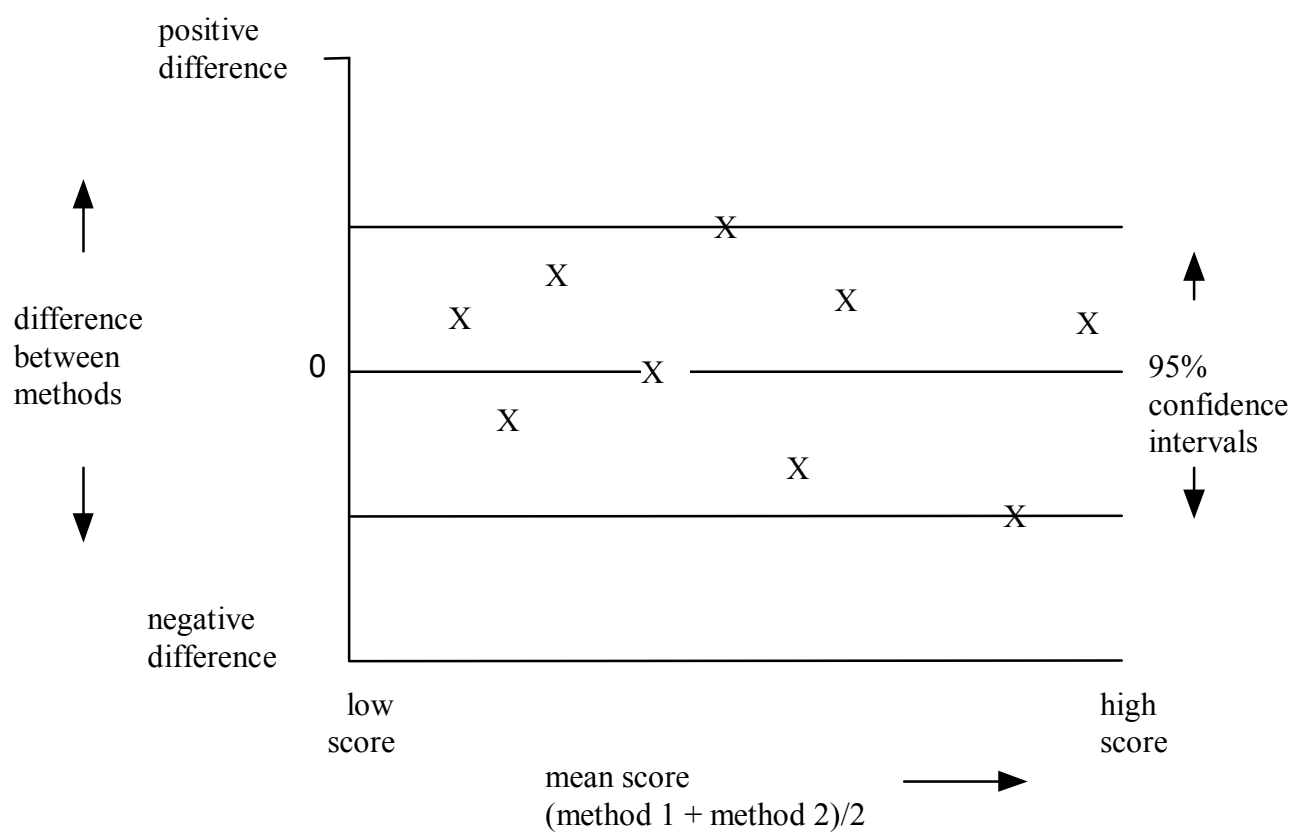
### Method comparison studies (Bland and Altman).

Physiological variables (e.g. blood pressure or serum electrolyte levels) are commonly measured in clinical practice. Any result obtained is an estimate of the true value and for various reasons there may be a degree of inaccuracy. It is often the case that there may be more than one method of measuring a variable and it would obviously be useful to compare methods to establish the extent to which they agree. The method of Bland and Altman was developed specifically for this purpose. Although correlation may be used as a way of comparing techniques (i.e. drawing a line through a scattergram of paired measurements and calculating a correlation coefficient) it doesn't give any information as to whether agreement between methods varies with the size of the value. Using the technique of Bland and Altman a graph is constructed (figure 3) of the difference between measurements (method 1 minus method 2) against the average value for the 2 methods. From this graph it is possible to observe whether the discrepancy between measurements varies with the quantity of the variable. It is also possible to calculate a mean and standard deviation for the points on the graph and see whether the population differs significantly from zero (i.e. no difference between measurement techniques). It is also possible to calculate 95% confidence intervals for this distribution to give the amount that measurement by one technique may vary from the other 95% of the time.



**FIGUR**  
techniq

of the difference between measurements using 2 different  
magnitude of the variable.

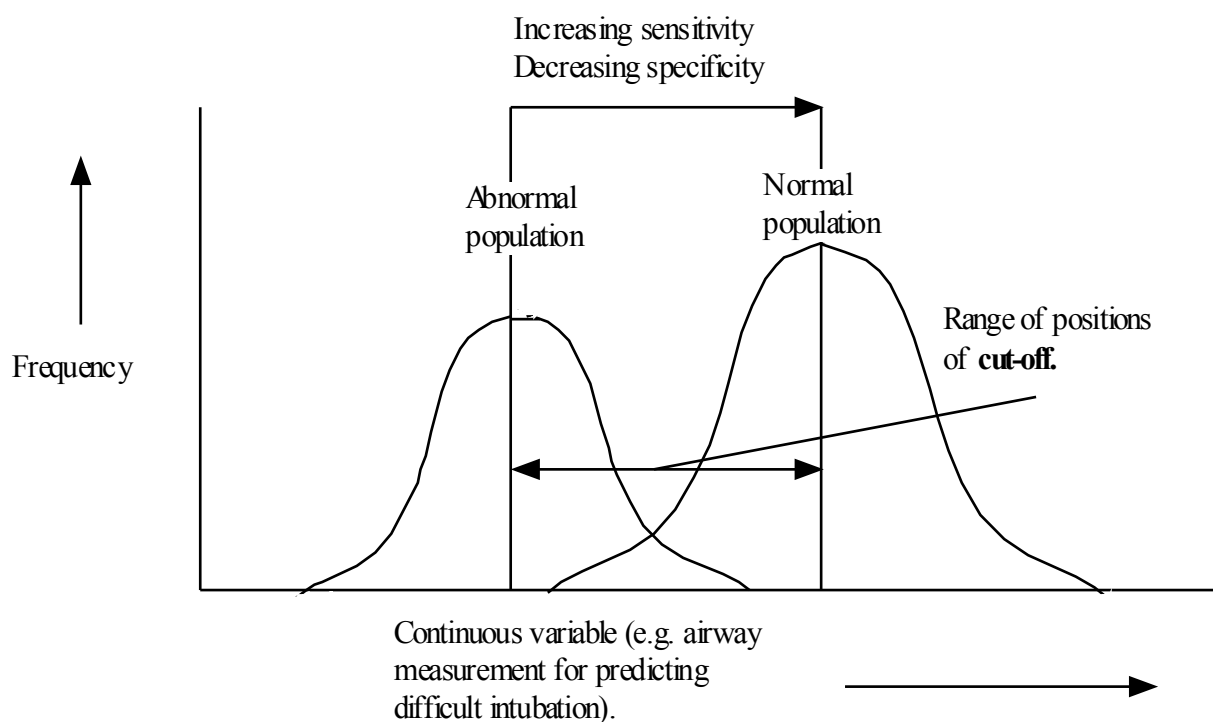


### SENSITIVITY AND SPECIFICITY:

(Specificity = the proportion of negatives which are correctly identified by the test.)

(Sensitivity = the proportion of positives which are correctly identified by the test).

Some diseases are diagnosed by means of a gold standard test (e.g. venography for venous thrombosis). Such a test may be expensive and / or invasive and therefore inappropriate for screening purposes. Consequently, a simpler / cheaper test may be developed. This screening test may be expressed on a continuous scale, i.e. one that can take on any value within a given range. A convenient cut-off must be chosen one side of which is taken to indicate that the condition is present and the other side that it is absent. It is inevitable that wherever the cut-off is placed, there will be some normal subjects with an abnormal score and visa versa (figure 4). If the cut-off is placed at a value close to the abnormal side then it is likely that most patients without the disease will test negative (true negative) i.e. it will be very specific. However, this will also mean that more of the subjects with the condition will not be picked up by the test (false negative) i.e. the test will be less sensitive. As the cut-off is moved away from the abnormal side the test will become more sensitive (pick up more of the subjects with the condition) but less specific (have an increased false negative rate):

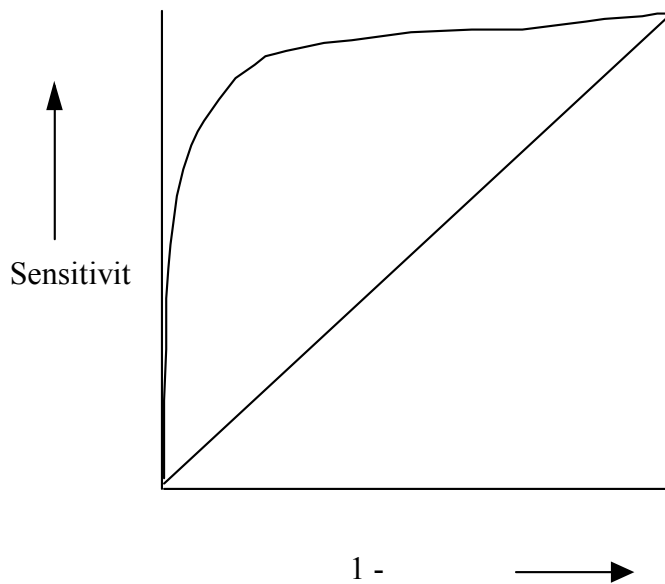


**FIGURE 4:** Distribution of test results for which normal and abnormal are defined by means of a gold standard test. The cut-off is taken to be the value below which the result would be considered to be positive for the abnormal condition. It can be seen that as the cut-off is moved to the right, so more and more of the abnormal population will be included i.e. the test becomes more sensitive. However, at the same time there

will be a rise in the proportion of false negatives (abnormal population being identified as normal) i.e. the test becomes less specific.

## CHARACTERISTICS OF A GOOD SCREENING TEST

A way of assessing whether or not a particular test is useful or not is to plot what happens to the sensitivity and specificity as the cut-off is altered on a receiver operating characteristic curve. For easier interpretation, the sensitivity (y-axis) is plotted against 1 minus the specificity (x-axis) (figure 5). For any useful test such a line should be above the diagonal i.e. have a high degree of specificity (right hand side



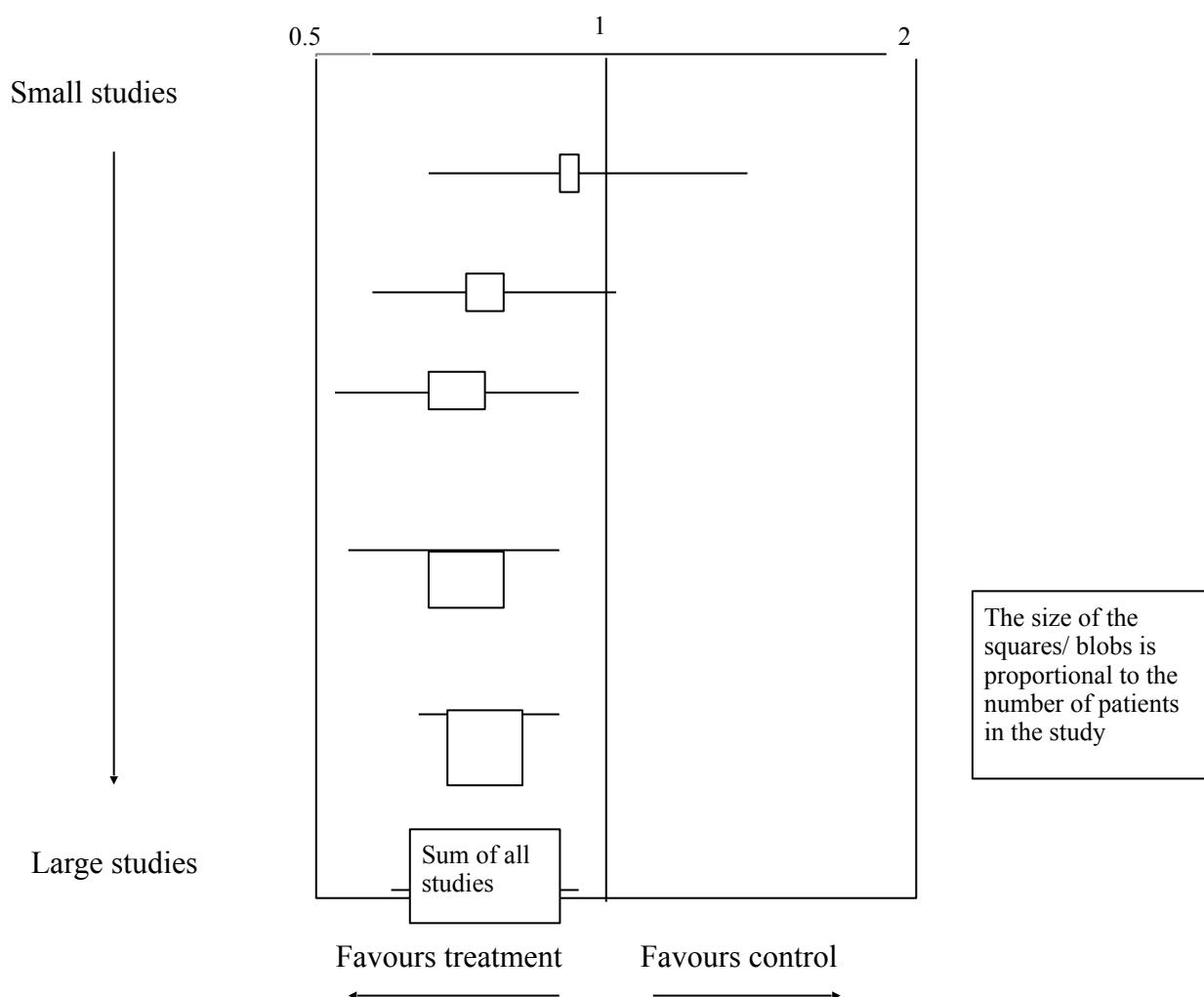
**FIGURE 5: Receiver operating characteristic curve** for a theoretical “good” screening test i.e. the line is drawn above the diagonal.

### Meta-analysis:

Meta-analysis is a statistical technique for combining the findings from a number of independent studies. Prior to performing a meta-analysis a **systematic review** must be undertaken. A systematic review is a complete, unbiased collection of original, high quality studies that examine the same therapeutic question. Systematic review is preferable to a narrative review where an individual summarises the literature according to his or her own impression of it, so that there is plenty of opportunity for bias.

Small studies often have low power and so may sometimes miss important differences in treatment. Bringing studies together allows a **pattern** to emerge and enables a more **precise estimate** of the difference between treatments. If studies are collected in a systematic way then there is **less danger of bias**. Good meta-analysis should be **transparent** allowing readers to judge for themselves the decisions that were taken to reach the final estimate of effect.

The results of meta-analysis are displayed as “**blobbograms**” where a blob/ square from an individual study represents the difference between treatments and a horizontal line represents the 95% confidence intervals. The results are brought together by means of an **odds ratio** which is like a relative risk. For example an odds ratio of 2 means that the outcome happens twice as often in the intervention group as in the control group.





**Prob Figure 6:** Mantel-Haenszel method of displaying Meta-analysis. The results

**P** of each study in the meta-analysis are displayed as odds ratios with 95%

1. confidence intervals.

studies where a difference is shown.

2. Studies included may differ in terms of methodology patient groups etc.
3. Loss of information on important outcomes – the meta-analysis seeks to simplify outcomes so that data from different studies may be combined. For example detailed information such as pain scores are simplified to reduction in pain scores by over 50% i.e. a “yes/ no” outcome.
4. Sub group analysis – defining subgroups in retrospect may be misleading.
5. Conflict with the results of large studies – Large well-conducted studies are still the best way of answering a research question. Sometimes they may contradict meta-analysis and flaws in the way that the meta-analysis was performed may be revealed in retrospect. For example meta-analysis suggested that magnesium was beneficial for patients who suffered heart attacks. However, this was contradicted by the large ISIS 4 trial.

#### **Further reading for statistics for the final FRCA:**

1. Introduction to statistics. Francis Clegg. Three papers published in the British Journal of Hospital Medicine:

i) April 1987, p 356

ii) April 1988, p334

iii) November 1988 (vol 40) p 396.

2. Medical statistics: A commonsense approach (Second edition).  
Campbell and Martin (Published by Wiley).