STATISTICS NOTES

Assuming *normal distribution* -
99% of subjects are within 3 standard deviations of the mean
95% of patients within 2 standard deviations
68% are within 1 standard deviation

Example: The mean Hb value for a group of 150 patients is 15 g/dl. The standard deviation is 2 g/dl. 95% of patients would have a Hb between 11 and 19.

Z scores can be used instead of standard deviation to estimate distribution.
If an observation lies 1 standard deviation away from the mean, its z score is 1. Hence, a score of 2 is 2 standard deviations more than the population, and a score of -2 is to standard deviations less than the mean.

Standard Deviation is the square of *variance*.

Standard error of the mean is the standard deviation of the sampling distribution of the *mean* - which gives an estimate of how close the sample mean is to the true population mean. It increases with sample size and increases with stand deviation.

The S.E.M. is the standard deviation divided by the *square root* of the sample size SEM = $\sigma / \sqrt{N}$ where σ is the standard deviation of the original distribution and N is the sample size).

Coefficient of variation is expressed in %. The definition coefficient of variation V = SD/mean. In this example 15/150 = 10%. It is a statistical measure of the deviation of a variable from its mean.

Sensitivity is the probability that a test will be positive when a patient has the condition. Alternatively, it is the number of true positives detected by the test divided by the number of all true positives in the population tested.

Specificity is the probability that a test will be negative when a patient does not have the condition. Alternatively it is the number of true negatives detected by the test divided by the number of all true negatives in the population.

Positive predictive value suggests the probability that if a test is positive, it is true

The null hypothesis states that there is no difference between treatments.
A **type 1 error** occurs when 'the null hypothesis is falsely rejected'. This means that the study claims to find a difference that does not really exist.

A **type 2 error** occurs when 'the null hypothesis is falsely accepted'. This means that although it is suggested that there is no difference between two groups, the study is actually too small to detect a difference.

Comparing two groups, all who have a risk factor, the relative risk is the ratio of those who had the disease, compared to those who did not have a disease. For example, the **relative risk** in a group of patients is the ratio of those who had CVA in a hypertension group compared to those who had CVA without hypertension (e.g. 10 patients / 2 patients = 5).

The **attributable risk** is the difference in incidence of a disease, among patients who have or do not have a risk factor. It is expressed in %. For example, 20% with hypertension had CVA and 5% without hypertension had CVA, hence 20% - 5% - 15%.

Another example, if a drug reduces the incidence of stroke from 10% to 5%. The **relative risk reduction** (RRR) is 50%. The **absolute risk reduction** (ARR) is 5%. The **number needed to treat** (NNT) is the total divided by the **absolute risk reduction**, which is 100% / 5% = 20. NNT is defined as number needed to treat to prevent 1 death.

Secondary prevention does treat patients with pre-existing disease, such as heart attack. It also involves identification of patients who are high risk of further development of cardiovascular disease

Chi-squared tests are used to compare percentages or proportions of *categorical data*. Data such as the above can be organised into a 2 x 2 contingency table. From the chi-squared value a p value is read off a statistical table (depends on degree of freedom) to give the degree of significance.

The **paired students t test** can be used to compare two groups of patients with parametric data (Null hypothesis being that any difference is due to chance).

Parametric means that the data will be of normal distribution and parallels the normal or bell curve). In addition, it means that numbers can be added, subtracted, multiplied, and divided. Normally distributed data can be compared with a Student's t-test.

Skewed continuous data can be compared with a **Wilcoxon rank-sum test** or a **Mann-Whitney U-test.** They should be described with a median and range, but NOT a mean. The tests are also known as **non parametric tests**.

Logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. ANOVA is an example of a logistic regression analysis.

Case control studies are not good at identifying rare causes. In order to identify whether a rare exposure causes a disease then the appropriate design is a large *cohort* study, where one group with the particular exposure of interest is compared with a control group without that exposure. The advantage of case control studies is that they can be used with rare diseases and can examine multiple risk-factors. Poor control selection often makes a case control study uninterpretable.

Crossover trials are designed where a patient can have one drug, have a washout period and then a second drug. It is applicable to chronic conditions, such as multiple sclerosis and not acute conditions such as MI requiring thrombolysis.

Intention to treat reduces bias by including the data from original allocation of treatment, e.g. analyzing the patient in a treatment group even if they did not complete the trial (dropped out because of side effects).