Correlation:

Correlation analysis is concerned with measuring the degree of association between two variables

Pearson correlation coefficient (r): measures how close the correlation is. For normally distributed numerical variables of sufficient sample size.
Value lies between -1 and +l.
Its sign indicates whether one variable increases as the other variable increases (positive r) or whether one variables decreases as the other increases (negative r).
Its magnitude indicates how close the points are to the straight line. In particular if r = +1 or -1, then there is perfect correlation with all the points lying on the line (this is most unusual, in practice); if r = 0, then there is no linear correlation (although there may be a non-linear relationship).The closer r is to the extremes, the greater the degree of linear association .
It is dimensionless, i.e. it has no units of measurement.
Its value is valid only within the range of values of x and y in the sample.
x and y can be interchanged without affecting the value of r.
Correlation does not suggest a causal relationship.
Pearson correlation coefficient is not appropriate for non-linear relationship, multiple values of the same variable in the same individual, extremes (outliers).

Spearman rank correlation coefficient is the non-parametric equivalent. It is appropriate when at least one variable is ordinal, sample size is small or when the observations are not normally distributed or when the correlation is non-linear.

Linear regression: measures the association between 2 continuous variables, where one is dependant on the other.

Tests of significance:

Start by assuming the null hypothesis holds, ie there is no difference between the 2 observations. If we can prove the null hypothesis is wrong, then the alternative hypothesis holds, ie that there is a difference.

Type 1 error (alpha)= rejecting the null hypothesis when it is true, ie detecting a difference when there is none. Value =5%

Type 2 error (beta)= accepting the null hypothesis when it is false, ie ruling out a difference when there is one. Power of the study = (1-beta). Value=70-80%.

The 95% confidence interval provides a range of values in which we are 95% certain that the true population mean lies

Numerical data: 2 unrelated groups

Normally distributed=Parametric test=Student t-test
Not normally distributed/ small number of observations= Non-parametric test=Wilcoxon rank-sum test/ Mann-Whitney U test

Numerical data: 2 related groups

Normally distributed=Parametric test=Paired t-test
Not normally distributed/ small number of observations= Non-parametric test=Wilcoxon signed rank test

Numerical data: more than 2 groups

Parametric=ANOVA
Non-parametric=Kruskal-Wallis test

Categorical data: single proportion

Sign test

Categorical data: 2 or more unrelated proportions

Chi square test
[Fischer's exact test (for number of observations in any one call of the 2x2 table <5)]

Categorical data: 2 related proportions

McNemar's test

The odds ratio is an estimate of **relative risk.** If the relative risk is equal to one (unity), then the 'risks' of having and not having the disease are the same when xl increases by one unit.

The **Kolmogorov-Smirnov** and **Shapiro-Wilk** tests, , can be used to assess Normality more objectively.

Study Power Calculation: **Lehr's formula**: 16/ (standardized difference)^2 for 2-tailed t-test/ chi-square test, significance value of 0.05 and power of 80%.

Sensitivity = proportion of individuals with the disease who are correctly identified by the test

Specificity = proportion of individuals without the disease who are correctly identified by the test
Positive predictive value = proportion of individuals with a positive test result who have the disease
Negative predictive value = proportion of individuals with a negative test result who do not have the disease

Box plot (often called a box-and-whisker plot) -This is a vertical or horizontal rectangle, with the ends of the rectangle corresponding to the upper and lower quartiles of the data values . A line drawn through the rectangle corresponds to the median value. Whiskers, starting at the ends of the rectangle, usually indicate minimum and maximum values but sometimes relate to particular percentiles, e.g. the 5th and 95th percentiles. Outliers may be marked.

Median: If the number of observations,n, is **odd,** the median is the *(n* + 1)12th observation in the ordered set. If *n* is **even** then, strictly, there is no median. However, we usually calculate it as the arithmetic mean of the two middle observations in the ordered set [i.e. the nl2th and the (n/2 + l)th].

The median is similar to the mean if the data are symmetrical , less than the mean if the data are skewed to the right and greater than the mean if the data are skewed to the left.

Variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The **standard deviation** (s) is the square root of the variance.
The **Standard Error of the Mean** (SEM) is

$$\text{SEM} = s/\sqrt{n}$$

Larger SEM= less precise, due to small sample size or larger variation

Confidence intervals:

A wide confidence interval indicates that the estimate is imprecise; a narrow one indicates a precise estimate. The width of the confidence interval depends on the size of the standard error, which in turn depends on the sample size and, when considering a numerical variable, the variability of the data. Therefore, small studies on variable data give wider confidence intervals than larger studies on less variable data. The upper and lower limits provide a means of assessing whether the results are clinically important

Bias:
Observer bias-one observer consistently under- orover-reports a particular variable;

Confounding bias-where a spurious association arises due to a failure to adjust fully for factors related to both the risk factor and outcome;

Selection bias-patients selected for inclusion into a study are not representative of the population to which the results will be applied;

Information bias -measurements are incorrectly recorded in a systematic manner; and

Publication bias-a tendency to publish only those papers that report positive or topical results.

Other biases may, for example, be due to recall , healthy entrant effect, assessment and allocation