Short Glossary in Medical Statistics for 2nd year medical students

Alpha α	The alpha value is the P value and should be interpreted in the same way.
Alternative hypothesis – H_1 or H_A (experimental hypothesis)	It's the hypothesis for which the researcher is trying to gain support through statistical analysis, by rejecting the null hypothesis. **H_1** states that there is a difference between the groups or a relationship between dependent and independent variables.
ANOVA (ANanysis Of Variance)	This is a grup of techniques used to compare the means of two or more samples to see wherter they come from the same population.
Association	A word used to describe a relationship between two variables.
Beta, β	The beta value is the probability of accepting a hypothesis that is actually false. $1- \beta$ is known as a power of the study.
Binary variable	Binary variable is a categorical variable with only two categories.
Bi-modal distribution	Where there are 2 modes in a set of data, it is said to be bi-modal.
Binominal distribution	When data can only take one of two values (for instance, male or female), it is said to follow a binominal distribution.
Box and whisker plot	A graph showing the median, range and interquartile range of a set of values.
Case-control study	A retrospective study which investigates the relationship between the eisk factor and one or more outcomes. This is done by selecting patients who already have the disease or outcome (cases), matching them to patients who do not (controls) and then comparing the effect of the risk factor on the two groups.
Cases	This usually refer to patients but could refer to hospitals, wards, countries, regions, blood samples etc.
Categorical variable	A variable whose values represent different categories of the same feature. Examples include different blood groups, different eye colours, different ethnic groups etc.

	When the variable has only two categories, it is termed **binary or dichotomous** (e.g. gender). When the variable has more than two categories, it is said to be **polychotomous** (e.g. profession). When there is some inherent ordering (e.g. mild, moderate, severe), it is called an **ordinal variable.**
Causation	The direct relationship of the **cause (independent variable)** to the **effect (dependent variable)** that it produces, usually established in experimental studies.
Central tendency	The central "scores" in a set of figures. **Mean, median and mode are measures of central tendency.**
Chi-squared test, χ^2	Chi-squared test is a test of association between two categorical variables.
Coefficient of variation	It expresses the sample standard deviation as a proportion or percentage of the mean value and can be calculated very easily by the following formula: $C_V = \frac{s}{X} \times 100$
Confidence interval, CI	A range of values within which we are fairly confident the true population value lies. For example, a 95% CI means that we can be 95% confident that the population value lies within those limits.
Cohort study	A prospective observational study that follows a group (cohort) over a period of time and investigates the effect of a treatment or risk factor. From epidemiological point of view, the design of a cohort study may be also retrospective and ambispective.
Confounding	A confounding factor is the effect of a covariate or factor that cannot be separated out. For example, if women with a certain condition received a new treatment and men received placebo, it would not be possible to separate the treatment effect due to gender. Therefore gender would be a confounding factor.
Continuous variable	A variable which can take any value within a given range, for instance BP.
Correlation	When there is a linear relationship between two variables there is said to be a correlation between them. Examples are height and weight in children, or socio-economic class and mortality.

	Measured on a scale from -1 (perfect negative correlation), through 0 (no correlation at all) to +1 (perfect positive correlation.
Correlation coefficient	A measure of the strength of the linear relationship between two variables.
Covariate	A covariate is a continuous variable that is not of primary interest but is measured because it may affect the outcome and may therefore need to be included in the analysis.
Database	A collection of records that is organized for ease and speed of retrieval.
Degrees of freedom, df	The number of degrees of freedom, often abbreviated to df, is the number of independent pieces of information available for the statistician to make the make the calculations.
Directional hypothesis (one-tailed)	It asserts that differences between groups in the data will occur in a particular direction, e.g. smokes die younger than non-smokers.
Descriptive statistics	Descriptive statistics are those which describe the data in the sample. They include means, medians, modes, standard deviations, quartiles and histograms. They are designed to give the reader an understanding of the data.
Discrete variable	A variable which data can only be certain values, usually whole numbers, for example the number of children in families.
Distribution	A distinct pattern of data may be considered to follow a distribution. Many patterns of data have been described, the most useful of which is the **normal distribution.**
Fisher's exact test	Fisher's exact test is an accurate test for association between categorical variables.
Histogram	A graph of continuous data with the data categorized into a number of classes.
Hypothesis	A statement which can be tested that predicts the relationship between variables.
Hypothesis testing	It is the process of deciding statistically whether the findings of an investigation reflect chance or 'real' effects at a given level of probability.
Incidence	The rate or proportion of a group developing a condition within a given period.

Inferential statistics	All statistical methods that test something are inferential. They estimate whether the results suggest that there is a real difference in the populations.
Inter-quartile range, IQR	A measure of spread given by the difference between the first quartile (the value below which 25% of cases lie) and the third quartile (the value below 75% of cases lie.
Kolmogorov Smirnov test	Kolmogorov Smirnov test is used to test the hypothesis that the collected data are from a normal distribution. It is therefore used to assess whether parametric statistics can be used.
Kruskal Wallis test	This is a non-parametric test which compared two or more independent groups.
Mann-Whitney U test	A non-parametric test to see whether there is a difference between two sets of data that have come from two different sets of subjects.
Mean	The sum of the observed values divided by the number of observations.
Median	The middle observation when the observed values are ranked from smallest to largest. **When the number of cases is odd**, the median is the value in the middle. **When the number of cases is even**, the median is just a halfway of values of the two middle observations.
Meta-analysis	Meta-analysis is a method of combining results from a number of independent studies to give one overall estimate of effect.
Mode	The most commonly occurring observed value.
Negative predictive value, NPV	If a diagnostic test is negative, the NPV is the chance that a patient does not have the condition.
Nominal data	Data that can be placed in named categories that have no particular order, for example eye colour.
Non-directional hypothesis (two-tailed)	It asserts that there are differences between groups in the data but with no direction specified, e.g. smokers and non-smokers have different life expectancies.
Non-parametric test	A test that is not dependent on the distribution (shape) of the data.

	Non-parametric tests – suitable for the analysis of nominal or ordinal data.
Normal distribution	This refers to a distribution of data that is symmetrical. In a graph it forms a characteristic bell shape.
Null hypothesis	A hypothesis that there is no difference between the groups being tested. The result of the test either supports or rejects that hypothesis. Paradoxically, the null hypothesis is usually the opposite of what we are actually interested in finding out. If we want to know whether there is a difference between two treatments, then the null hypothesis would be that there is no difference. The statistical test would be used to try to disprove this.
Odds	The ratio of the number of times an event happens to the number of times it does not happen in a group of patients. Odds and risk have similar values when considering rare events (e.g. winning the lottery), but may be substantially different in common events (e.g. not winning the lottery).
One-tailed test	The test whether the null hypothesis can only be rejected in one direction, for example if new treatment is worse than current treatment but not if it is better. **One-tailed test -** a statistical test where a difference between two groups is tested in a particular direction of the difference, e.g. to test a **directional hypothesis** – when not only the significance of differences is tested but also the direction of these differences is determined. It should only rarely be used to test non-directional hypothesis.
Ordinal data	Data that can be allocated to categories that can be "ordered", e.g. from least to strongest. Examples: the staging of malignancy or many other diseases.
***P* value**	Usually used to test the null hypothesis, the P value gives the probability of any observed differences having happened by chance.
Parametric test	Any test that has an assumption that the data needs to follow a certain distribution can be considered to be a parametric test. The most common distribution that the data need to follow is the normal distribution. Examples are t-test and ANOVA. **Parametric tests** – suitable for the analysis of interval or ratio data.

Pearson correlation coefficient	A method of calculation a correlation coefficient if the values are sampled from a normal distribution.
Percentage	The number of items in a category, divided by the total number in the group, then multiplied by 100.
Percentiles	Points that divide an array into 100 equal parts.
Poisson distribution	This distribution represents the number of events happening in a fixed time interval, for instance the number of deaths in a year.
Population	The complete set of subjects from which a sample is drawn.
Positive predictive value, PPV	If a diagnostic test is positive, the PPV is the chance that a patient has the condition.
Power	The power of a study is the probability that it will detect a statistically significant difference.
Prevalence	The proportion of a group with a condition at a single point in time.
Proportions	It is the frequency of one category over that of the total numbers in the sample or the population – A/A+B Percentages - the same as the proportions but multiplied by 100.
Quantiles (Q)	Quantiles are special measures of location - points that divide the ordered series of data (from the lowest to the highest value) into subgroups of equal size.
Quartiles	Quartiles may be given with the median. The first quartile has ¼ of the data below it, the 3rd quartile has ¾ of the data below it.
r	Where there is a linear relationship between two variables there is to be a correlation between them. The correlation coefficient gives the strength of that relationship.
R^2	Coefficient of determination – an estimate of the amount of the variation that is being explained by a regression model.
Range	The difference between the maximum and a minimum score in a set of figures.
Rank	A numerical value given to an observation showing its relative order in a set of data.

Rate	The number of times that an event happens in a fixed period of time.
Ratios	Statistics which express the relative frequency of one set of frequencies, A, in relation to another, B. Ratios are useful in the health sciences when we are interested in the distribution of illnesses or symptoms or the categories of subjects requiring or benefiting from some treatment.
Regression	Regression analysis is a technique for finding the relationship between two variables, one of which is dependent on the other,
Relative risk	**Risk ratio** is often referred to as **relative risk.** However, **odds ratio**s are also a measure of relative risk.
Risk	The probability of occurrence of an event. Calculated by dividing the number of events by the number of people at risk.
Risk ratio, RR	The risk of an event happening in one group (e.g. incidence in exposed), divided by the risk of it happening in another group (e.g. incidence in unexposed).
Sample	A small group drawn from a larger population.
Sensitivity	This is the rate of pick-up od a condition in a test. In other words, the proportion of patients with a condition having a positive test result.
Significance	The probability of getting the results if the null hypothesis is true. **Statistical significance (P)** – it's the probability over which H_0 is accepted to be true and below which H_0 is rejected. P for H_0 + P for H_1 = 1 = 100%
Spearman correlation coefficient	An estimate of correlation used for non-parametric variables (which are not normally distributed).
Specificity	The rate of estimation of the possibility of disease by a test. In other words, the proportion of patients without the condition that has a negative test result.
Skewed data	A lack of symmetry in the distribution of data.
Standard deviation, SD	A measure of the spread of scored away from the mean.

Standard error of the mean	A measure of how close the sample mean is likely to be to the population mean.
Stratified sample	A stratified sample is one that has been split into a number of subgroups. These subgroups should have the same proportions as the groups in the population from which the sample gas been drawn.
Subjects	The sample in a study.
***t* test (also known as Student's *t* test)**	***t*** test is a parametric test used to compare the means of two groups.
Transformation	A transformation is where a mathematical formula is used to change the data. This will often be done to try to make the data follow a normal distribution so that a parametric test can be used.
Type I and II errors	Any statistical test can fail in two ways. A hypothesis that is correct can be rejected (type I error), or a hypothesis that is incorrect can be accepted (type II error). The chance of making a type I error is the same as the P value.
Two-tailed test	A statistical test where a difference between two groups is tested without reference to the expected direction of the difference, e.g. **for non-directional hypothesis.** The critical area for two-sided test is a series of values that are less that the first critical value of the test and a series of values that are higher than the second critical value of test.
U criterion	The ratio of the difference between the outlier and the mean and the standard deviation s. The computed criterion U is ompared with the table of critical values of u_t and if $u \geq u_t$, the extreme value x_i is discarded as unusual.
Variable	A variable is a property, an attribute or any characteristic that differs from subject to subject or from time to time. In data analysis, variables may be called fields and refer to all the things recorded on the cases.
Variance	A measure of the spread of scores from the mean. It is the square of the standard deviation.