Glossary of Statistical Terms

Barbara Butland and Martin Bland

Binomial Distribution: the probability distribution for the number of 'successes' in n independent trials where the probability of success is *p* and therefore the probability of 'failure' is 1–*p*. *n* and *p* are called the parameters of the distribution. By varying the values of *n* and *p* we obtain many different examples of the Binomial distribution.

Case-control study: In a case-control study we take a group of people with the disease, the cases, and a second group without the disease, the controls. We then obtain information for each subject concerning exposure to possible causative factors and look for differences in exposure between the two groups. The way in which we select controls is obviously very important. We want a group of subjects that do not have the disease under study but who are otherwise comparable to our cases. The design is usually retrospective (relating current disease status to past exposures).

Case-Control and Cohort Studies: Advantages and Disadvantages.

i) A cohort study allows you to calculate the rate of disease among exposed and unexposed individuals. Rates cannot be calculated from a case-control study.

ii) A cohort study requires a large number of subjects followed up over a long period if sufficient numbers are to develop the diseases of interest. Since the case-control study starts with cases it requires comparatively few subjects and is comparatively quick to conduct.

iii) With a case-control study there is the problem of selecting a suitable control group.

iv) The need in most case-control studies to collect information on past exposures by asking subjects to recall past events may lead to poor quality information and to bias. Differences in the ability of cases and controls to recall past events may lead to recall bias. Knowledge of case-control status by the interviewer may lead to assessment bias. Recall and assessment bias are not a problem in prospective cohort studies as information on the possible causal factor is collected at the time of exposure and prior to the onset of disease.

Central Limit Theorem: States that if we have any series of independent identically distributed random variables then their sum tends to a Normal distribution as the number of variables increases.

Chi-squared test: Used to test the null hypothesis of no relationship or no association between two qualitative variables. The two qualitative variables are cross tabulated to form a contingency table. The test statistic =

$$\sum \frac{(O-E)^2}{E}$$

where *E*= the expected frequencies and *O*=the observed frequencies in the contingency table. If 80% of expected frequencies are > 5 and all expected frequencies are > 1 the test statistic can be assumed to follow a Chi-squared distribution with degrees of freedom given by (number of rows – 1) × (number of columns – 1).

Clinical trial: An EXPERIMENTAL study. Allocate patients to groups e.g. treatment or placebo, and compare outcomes.

Cohort study: Take a population of individuals selected usually by a common link e.g. living in the same geographical area or working in the same factory. (Note: not chosen on medical grounds). Include in the study either the entire population or a representative sample. Collect information on the study subjects concerning exposure to the possible causative factor of interest. Follow the subjects forward in time to see whether they develop disease. Information on exposure can then be related to subsequent disease experience. The design is often prospective (cohort defined in the present and followed into the future).

Confidence Intervals: We can use a random sample to estimate characteristics of the population from which the sample was drawn. For example we might use the proportion of females in a sample to estimate the proportion of females in the population. However estimates vary from sample to sample. A 95% confidence interval gives us some idea of the extent of this variability allowing us to assess the accuracy of the estimate provided by a single sample. The wider the interval the poorer the estimate. A 95% confidence interval is constructed such that if we had an infinite number of samples of a given size selected at random from a population and calculated a 95% confidence interval for each sample, 95% of the intervals obtained would contain the true population value. Confidence intervals can be calculated using probabilities other than 95% (e.g. 90%, 99%). The larger the probability associated with a confidence interval, the wider the interval.

Correlation coefficient *r*: a measure of how closely two continuous variables are linearly related. It must lie between +1 and −1. A value of +1 or −1 shows that the two variables are perfectly linearly related and if the two variables were plotted against each other all points would lie exactly on a straight line. If r=+1 the slope is positive and if one variable has a high value the other variable has a high value. If $r = -1$ the slope is negative and if one variable has a high value the other variable has a low value. If $r = 0$ the variables are not linearly related. We can test the null hypothesis of no linear association between two continuous variables i.e. $r = 0$, using a significance test provided at least one of the continuous variables follows a Normal distribution in the population.

Independent Events: Two events are said to be independent if knowing one has happened tells us nothing about whether the other happens.

Intention to treat: When patients are randomly allocated to treatment groups, the groups are comparable. If some patients do not get the allocated treatment we must still include them in the analysis in the group to which they were allocated, or groups will not be comparable.

Large Sample Comparisons using the Normal Distribution: When samples are large we can assume that the sample means and proportions are observations from Normal distributions - this follows from the Central Limit Theorem. We can also assume that the sample standard errors are good estimates of the standard deviations of these Normal distributions. Thus the ratios of sample means and proportions to their sample standard errors can be assumed to follow Normal distributions. These methods are also known as z tests.

Large Sample Comparison of Two Means: Based on two large independent samples selected at random from two populations. The confidence interval is given by

$$\bar{x}_1 - \bar{x}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where x_1 and x_2 are the sample means, s_1 and s_2 the sample standard deviations and n_1 and n_2 the sample sizes. To test the null hypothesis that the means of the two populations are equal i.e. that the difference between the two means equals zero, the test statistic = (difference in means) / standard error (difference):

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Provided the samples are large ($n_1 > 50$ and $n_2 > 50$), we can assume that the test statistic follows the Standard Normal distribution when the null hypothesis is true.

Large Sample Comparison of Two Proportions: Based on two large independent samples (sizes n_1 and n_2) selected at random from two populations, with sample proportions p_1 and p_2. The confidence interval for the difference is given by:

$$p_1 - p_2 \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

To test the null hypothesis that the proportion of subjects exhibiting a particular characteristic is the same in both populations, we first find the common proportion p where

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

The test statistic = (difference in proportions) / standard error (difference):

$$\frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

If the samples are large, the test statistic can be assumed to follow the Standard Normal distribution when the null hypothesis is true.

Linear regression: a method of estimating the linear relationship between continuous variables. $Y = a + bX$ is the regression equation in **simple** linear regression. The X is the predictor (independent) variable, b is the regression coefficient (slope of the line) and a is the intercept (the point where the line cuts the y-axis). Y is the outcome (dependent) variable and for any given value of X it should be Normally distributed with the same variance. The regression line is fitted to the data using the method of least squares.

Matching: It is clearly important that cases and controls should be comparable with respect to potential confounding factors like age and sex. Comparisons between cases and controls must take into account any differences e.g. by adjustment in the analysis. However if each case is paired or matched with a control who is deliberately chosen to be of the same age and sex, adjustment, at least for age and sex, may be avoided.

Mutually Exclusive Events: Two events are said to be mutually exclusive if when one happens the other cannot happen.

Normal Distribution: uniquely defined by its mean and standard deviation. It is symmetrical about the mean and may be represented graphically as a bell shaped curve, known as the Normal curve. The area under the curve = 1. Most of the area under the curve is within ± one SD of the mean, the large majority (95%) is within ± 1.96 SD (often written as 2 SD for short) of the mean, almost all is within ± 3 SD of the mean.

NS: Stands for Not Significant. It indicates that the null hypothesis has not been rejected. In other words the value of p is above the significance level. Not significant does not mean there is no effect, only that we have failed to find one.

Null Hypothesis: The hypothesis of 'no difference' or 'no effect' which is rejected or not rejected according to the results of a significance test.

Observational study: Literally you OBSERVE subjects in their current state i.e. no drugs are prescribed or interventions applied. Examples are Cohort studies and Case-control studies.

Odds of Disease: The odds of an individual developing a disease are the probability that any given individual will develop the disease divided by the probability that any given individual will not develop the disease.

Odds Ratio: is the ratio of the odds of disease among those with the factor of interest to the odds of disease among those without the factor. The odds ratio can be calculated from a case-control study as the ratio of cross-products, *ad/bc*, where *a*, *b*, *c*, and *d* are the four frequencies in the 2 by 2 table (see MB notes or book). If the disease under study is rare the odds ratio provides a good approximation to the relative risk.

One sample (paired) t-test: Given one sample selected at random from the population of interest, the one sample (paired) t test can be used to test the null hypothesis that the mean difference in the population between measurements obtained for the same subjects under two conditions is zero. The test statistic = $\bar{d} / (s^2 / n)$, where $\bar{d}$ is the mean difference in the sample, *s* is the standard deviation of the differences in the sample and *n* is the sample size. Provided differences are Normally distributed in the population, the test statistic follows a t distribution on $n-1$ degrees of freedom when the null hypothesis is true. The test can also be used for two matched samples.

P value: The probability of getting a value of the test statistic at least as or more extreme as that observed if the null hypothesis is true.

Population Mean of a variable (μ): Average value of all the observations in the population.

Population Standard Deviation of a variable (σ): Square root of the population variance and therefore in the same units as the basic observations.

Population Variance of a variable (σ^2): Measure of the spread of the population variable values about the population mean. It is calculated by obtaining the difference of each observation from the population mean, squaring the differences and taking the average.

Probability Distribution: Suppose we have a set of all possible mutually exclusive events, then their probabilities sum to 1.0 and make up a probability distribution.

Probability of an Event: The proportion of trials in which the event will occur in the long run e.g. the probability of tossing an unbiased coin once and getting a 'head' = 1/2.

Random allocation or randomisation: Method used to achieve comparable groups. The group to which a subject is allocated does not depend on the characteristics of the subject. The only differences between the groups will be those due to chance. The effects of chance can be measured using statistical methods. Subjects can be randomly allocated into groups using random number tables.

Random sample: The sample is selected so as to give each member of the population the same probability of being chosen. Whether or not a subject is selected does not depend on the characteristics of that subject. The fact that the sample is random means that we can apply methods of probability to the data obtained. Random sampling is used to obtain a representative sample of the population. Given a list of the whole population random numbers from tables or generated by computer can be used to select a random sample.

Random Variable: A variable which can take more than one value with given probabilities e.g. the number of 'heads' in 10 tosses of an unbiased coin.

Randomly selected: the patients are a random sample of some larger population.

Relative Risk: is the ratio of the risk of disease among those with the factor of interest to the risk of disease among those without the factor. The relative risk cannot be calculated directly from a case-control study, because the risks cannot be calcualted, but can be estimated from the odds ratio.

Risk of Disease: The risk of an individual developing a disease is simply the probability that any given individual will develop the disease.

Sample Mean of a variable ($\bar{x}$): Average value of all the observations in the sample. Used as an estimator of the population mean.

Sample standard error of a variable: The value of any estimator e.g. the sample mean, will vary from sample to sample. The standard error measures this variability. It is the standard deviation of the estimator. The sample standard error of the mean $= s / \sqrt{n}$.

Sample Variance of a variable (s^2): Used as an estimator of the population variance. It is calculated by obtaining the difference of each observation in the sample from the sample mean, squaring the differences and adding them together. The total is then divided by $n - 1$ where n = sample size. Note: Dividing by $n - 1$ rather than n produces a better estimator of the population variance.

Sampling distribution of the mean: The distribution of the means of all possible samples of the same size.

Sign test: a significance test used for paired data based on signs of differences. Useful when the assumption of Normality for the one-sample paired t-test does not hold, although can be used for any ordered paired data.

Significance Level: A value of P below which there is strong evidence to reject the null hypothesis. The level usually chosen is that of 5% i.e. if P is less than 0.05 then the null hypothesis is rejected.

Significance Test: Measures the strength of evidence against the null hypothesis.

Tests for Small Samples using the t Distribution: Provided the data are Normally distributed i.e. provided the variable of interest follows a Normal distribution in the population, we can assume that means calculated from small samples follow Normal distributions. However the standard errors calculated from small samples are not good estimates of the standard deviations of these Normal distributions. If a small sample of size n is selected at random from a population in which the variable of interest follows a Normal distribution, the ratio (mean over standard error) calculated from the sample follows a t distribution on $n - 1$ degrees of freedom.

Test Statistic: A calculated 'value' whose distribution is known if the null hypothesis is true.

Transformation: if the distribution of a variable is not Normal it may be possible to transform it mathematically to improve the approximation to a Normal distribution. One way of doing this is to take the log of the variable.

Two sample (unpaired) t-test: Based on two independent samples selected at random from two populations where the variable of interest is Normally distributed. It tests the null hypothesis that the means of the two populations are equal. The test statistic = (difference in means) / standard error (difference) =

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

where $s^2=((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/ (n_1 + n_2 - 2)$ with s_1 and s_2 the sample standard deviations and n_1 and n_2 the sample sizes. Provided both populations have the same variance the test statistic follows a t distribution on $n_1+n_2 - 2$ degrees of freedom when the null hypothesis is true.