SYNOPSIS OF MEDICAL STATISTICS

From Wikipedia, the free encyclopedia

Sample size

The **sample size** of a statistical sample is the number of observations that constitute it. It is typically denoted *n*, a positive integer (natural number). Typically, all else being equal, a larger sample size leads to increased precision in estimates of various properties of the population. This can be seen in such statistical rules as the law of large numbers and the central limit theorem. Repeated measurements and Replication of independent samples are often required in measurement and experiments to reach a desired precision.

A typical example would be when a statistician wishes to estimate the arithmetic mean of a continuous random variable (for example, the height of a person). Assuming that they have a random sample with independent observations, then if the variability of the population (as measured by the standard deviation ó) is known, then the standard error of the sample mean is given by the formula:

$$\sigma/\sqrt{n}.$$

It is easy to show that as *n* becomes large, this variability becomes very small. This yields to more sensitive hypothesis tests with greater statistical power and smaller confidence intervals.

Central Limit Theorem

The central limit theorem is a significant result which depends on sample size. It states that as the size of a sample of independent observations approaches infinity, provided data come from a distribution with finite variance, that the sampling distribution of the sample mean approaches a normal distribution.

Estimating Proportions

A typical statistical aim is to demonstrate with 95% certainty that the true value of a parameter is within a distance *B* of the estimate: *B* is an error range that decreases with increasing sample size (*n*). The value of *B* generated is referred to as the 95% confidence interval.

For example, a simple situation is estimating a proportion in a population. To do so, a statistician will estimate the bounds of a 95% confidence interval for an unknown proportion.

The rule of thumb for (a maximum or 'conservative') *B* for a proportion derives from the fact the estimator of a proportion, $\hat{p} = X/n$, (where *X* is the number of 'positive' observations) has a (scaled) binomial distribution and is also a form of sample mean (from a Bernoulli distribution [0,1] which has a maximum variance of 0.25 for parameter $p = 0.5$). So, the sample mean X/n has maximum variance $0.25/n$. For sufficiently large *n* (usually this means that we need to have observed at least 10 positive and 10 negative responses), this distribution will be closely approximated by a normal distribution with the same mean and variance.

Using this approximation, it can be shown that ~95% of this distribution's probability lies within 2 standard deviations of the mean. Because of this, an interval of the form

$$(\hat{p} - 2\sqrt{0.25/n}, \hat{p} + 2\sqrt{0.25/n}) = (\hat{p} - B, \hat{p} + B)$$

will form a 95% confidence interval for the true proportion.

If we require the sampling error å to be no larger than some bound B, we can solve the equation

$$\varepsilon \approx B = 2\sqrt{0.25/n} = 1/\sqrt{n}$$

to give us

$$1/\varepsilon^2 \approx 1/B^2 = n$$

So, $n = 100$ <=> $B = 10\%$, $n = 400$ <=> $B = 5\%$, $n = 1000$ <=> $B = \sim 3\%$, and $n = 10000$ <=> $B = 1\%$. One sees these numbers quoted often in news reports of opinion polls and other sample surveys.

Extension to other cases

In general, if a population mean is estimated using the sample mean from *n* observations from a distribution with variance ó², then if *n* is large enough (typically >30) the central limit theorem can be applied to obtain an approximate 95% confidence interval of the form

$$(\bar{x} - B, \bar{x} + B), B = 2\sigma/\sqrt{n}$$

If the sampling error å is required to be no larger than bound *B*, as above, then

$$4\sigma^2/\varepsilon^2 \approx 4\sigma^2/B^2 = n$$

Note, if the mean is to be estimated using P parameters that must first be estimated themselves from the same sample, then to preserve sufficient "degrees of freedom," the sample size should be at least $n + P$.

Required Sample Sizes for Hypothesis Tests

A common problem facing statisticians is calculating the sample size required to yield a certain power for a test, given a predetermined Type I error rate á. A typical example for this is as follows:

Let X_i, $i = 1, 2, ..., n$ be independent observations taken from a normal distribution with mean ì and variance ó² . Let us consider two hypotheses, a null hypothesis:

H_0:μ = 0

and an alternative hypothesis:

H_a:μ = μ *

for some 'smallest significant difference' ì* >0. This is the smallest value for which we care about observing a difference. Now, if we wish to (1) reject H_0 with a probability of at least 1-â when H_a is true (i.e. a power of 1-â), and (2) reject H_0 with probability á when H_0 is true, then we need the following:

If z_γ is the upper γ percentage point of the standard normal distribution, then

$$\Pr(\bar{x} > z_\alpha \sigma/\sqrt{n} | H_0 \text{ true}) = \alpha$$

and so

'Reject H_o if our sample average ($\bar{x}$) is more than

$$z_\alpha \sigma/\sqrt{n}$$

is a decision rule which satisfies (2). (Note, this is a 1-tailed test)

Now we wish for this to happen with a probability at least 1-â when H_a is true. In this case, our sample average will come from a Normal distribution with mean ì*. Therefore we require

$$\Pr(\bar{x} > z_\alpha \sigma/\sqrt{n} | H_a \text{ true}) \geq 1 - \beta$$

Through careful manipulation, this can be shown to happen when

$$n \geq \left(\frac{\Phi^{-1}(1-\beta) + z_\alpha}{\mu/\sigma} \right)^2$$

where Ö is the normal cumulative distribution function.

Stratified Sample Size

With more complicated sampling techniques, such as Stratified sampling, the sample can often be split up into sub-samples. Typically, if there are *k* such sub-samples (from *k* different strata) then each of them will have a sample size n_i, $i = 1, 2, ..., k$. These n_i must conform to the rule that $n_1 + n_2 + ... + n_k = n$ (i.e. that the total sample size is given by the sum of the sub-sample sizes). Selecting these n_i optimally can be done in various ways, using (for example) Neyman's optimal allocation.

According to Leslie Kish,[1] there are many reasons to do this; that is to take sub-samples from distinct sub-populations or "strata" of the original population: to decrease variances of sample estimates, to use partly non-random methods, or to study strata individually. A useful, partly non-random method would be to sample individuals where easily accessible, but, where not, sample clusters to save travel costs.

In general, for H strata, a weighted sample mean is

$$\bar{x}_w = \sum_{h=1}^{H} W_h \bar{x}_h,$$

with

$$\mathrm{Var}(\bar{x}_w) = \sum_{h=1}^{H} W_h^2 \, \mathrm{Var}(\bar{x}_h).$$

The weights, W(h), frequently, but not always, represent the proportions of the population elements in the strata, and W(h)=N(h)/N. For a fixed sample size, that is n=Sum{n(h)},

$$\mathrm{Var}(\bar{x}_w) = \sum_{h=1}^{H} W_h^2 \, \mathrm{Var}(h) \left(\frac{1}{n_h} - \frac{1}{N_h} \right),$$

which can be made a minimum if the sampling rate within each stratum is made proportional to the standard deviation within each stratum: $n_h / N_h = kS_h$.

An "optimum allocation" is reached when the sampling rates within the strata are made directly proportional to the standard deviations within the strata and inversely proportional to the square roots of the costs per element within the strata:

$$\frac{n(h)}{N(h)} = \frac{KS(h)}{\sqrt{C(h)}},$$

or, more generally, when

$$n(h) = \frac{K'W(h)S(h)}{\sqrt{C(h)}}.$$

Statistical Power

The power of a statistical test is the probability that the test will reject a false null hypothesis (that it will not make a Type II error). As power increases, the chances of a Type II error decrease. The probability of a Type II error is referred to as the false negative rate (â). Therefore power is equal to 1 " â.

Power analysis can either be done before (*a priori*) or after (*post hoc*) data is collected. *A priori* power analysis is conducted prior to the research study, and is typically used to determine an appropriate sample size to achieve adequate power. *Post-hoc* power analysis is conducted after a study has been completed, and uses the obtained sample size and effect size to determine what the power was in the study, assuming the effect size in the sample is equal to the effect size in the population.

Statistical tests attempt to use data from samples to determine if differences or similarities exist in a population. For example, to test the null hypothesis that the mean scores of men and women on a test do not differ, samples of men and women are drawn, the test is administered to them, and the mean score of one group is compared to that of the other group using a statistical test. The power of the test is the probability that the test will find a statistically significant difference between men and women, as a function of the size of the true difference between those two populations. Despite the use of random samples, which will tend to mirror the population due to mathematical properties such as the central limit theorem, there is always a chance that the samples will appear to support or refute a tested hypothesis when the reality is the opposite. This risk is quantified as the power of the test and as the statistical significance level used for the test.

Statistical power depends on:

- the statistical significance criterion used in the test.
- the size of the difference or the strength of the similarity (that is, the effect size) in the population.
- the sensitivity of the data.

A significance criterion is a statement of how unlikely a result must be, if the null hypothesis is true, to be considered significant. The most commonly used criteria are probabilities of 0.05 (5%, 1 in 20), 0.01 (1%, 1 in 100), and 0.001 (0.1%, 1 in 1000). If the criterion is 0.05, the probability of the difference must be less than 0.05, and so on. One way to increase the power of a test is to increase (that is, weaken) the significance level. This increases the chance of obtaining a statistically significant result (rejecting the null hypothesis) when the null hypothesis is false, that is, reduces the risk of a Type II error. But it also increases the risk of obtaining a statistically significant result when the null hypothesis is in fact true; that is, it increases the risk of a Type I error.

Calculating the power requires first specifying the effect size you want to detect. The greater the effect size, the greater the power.

Sensitivity can be increased by using statistical controls, by increasing the reliability of measures (as in psychometric reliability), and by increasing the size of the sample. Increasing sample size is the most commonly used method for increasing statistical power. Although there are no formal standards for power, most researchers who assess the power of their tests use 0.80 as a standard for adequacy.

A common misconception by those new to statistical power is that power is a property of a study or experiment. In reality any statistical result that has a p-value has an associated power. For example, in the context of a single multiple regression, there will be a different level of statistical power associated with the overall r-square and for each of the regression coefficients. When determining an appropriate sample size for a planned study, it is important to consider that power will vary across the different hypotheses.

There are times when the recommendations of power analysis regarding sample size will be inadequate. Power analysis is appropriate when the concern is with the correct acceptance or rejection of a null hypothesis. In many contexts, the issue is less about determining if there is or is not a difference but rather with getting a more refined estimate of the population effect size. For example, if we were expecting a population correlation between intelligence and job performance of around .50, a sample size of 20 will give us approximately 80% power (alpha = .05, two-tail). However, in doing this study we are probably more interested in knowing whether the correlation is .30 or .60 or .50. In this context we would need a much larger sample size in order to reduce the confidence interval of our estimate to a range that is acceptable for our purposes. These and other considerations often result in the recommendation that when it comes to sample size, "More is better!"

Funding agencies, ethics boards and research review panels frequently request that a researcher perform a power analysis. The argument is that if a study is inadequately powered, there is no point in completing the research.

Student's t-test

A ***t*-test** is any statistical hypothesis test in which the test statistic has a Student's *t* distribution if the null hypothesis is true. It is applied when sample sizes are small enough that using an assumption of normality and the associated z-test leads to incorrect inference.

Use

A *t*-test is any statistical hypothesis test in which the test statistic has a Student's t-distribution if the null hypothesis is true. It is applied when sample sizes are small enough that using an assumption of normality and the associated z-test leads to incorrect inference. Among the most frequently used *t* tests are:

- A test of the null hypothesis that the means of two normally distributed populations are equal. Given two data sets, each characterized by its mean, standard deviation and number of data points, we can use some kind of *t* test to determine whether the means are distinct, provided that the underlying distributions can be assumed to be normal. All such tests are usually called **Student's *t* tests**, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's *t* test. There are different versions of the *t* test depending on whether the two samples are
  - unpaired, independent of each other (e.g., individuals randomly assigned into two groups), or
  - paired, so that each member of one sample has a unique relationship with a particular member of the other sample (e.g., the same people measured before and after an intervention, or IQ test scores of a husband and wife).

If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05, or 0.01 level), then the null hypothesis which usually states that the two groups do not differ is rejected in favor of an alternative hypothesis, which typically states that the groups do differ.

- A test of whether the mean of a normally distributed population has a value specified in a null hypothesis.
- A test of whether the slope of a regression line differs significantly from 0.

Once a *t* value is determined, a p-value can be found using a table of values from Student's t-distribution.

Assumptions

- Normal distribution of data, tested by using a normality test, such as Shapiro-Wilk and Kolmogorov-Smirnov test.
- Equality of variances, tested by using either the F test, the more robust Levene's test, Bartlett's test, or the Brown-Forsythe test.

- Samples may be independent or dependent, depending on the hypothesis and the type of samples:
  - Independent samples are usually two randomly selected groups
  - Dependent samples are either two groups matched on some variable (for example, age) or are the same people being tested twice (called repeated measures)

Since all calculations are done subject to the null hypothesis, it may be very difficult to come up with a reasonable null hypothesis that accounts for equal means in the presence of unequal variances. In the usual case, the null hypothesis is that the different treatments have no effect — this makes unequal variances untenable. In this case, one should forgo the ease of using this variant afforded by the statistical packages. See also Behrens-Fisher problem.

One scenario in which it *would* be plausible to have equal means but unequal variances is when the 'samples' represent repeated measurements of a single quantity, taken using two different methods. If systematic error is negligible (e.g. due to appropriate calibration) the effective population means for the two measurement methods are equal, but they may still have different levels of precision and hence different variances.

Determining Type

For novices, the most difficult issue is often whether the samples are independent or dependent. Independent samples typically consist of two groups with no relationship. Dependent samples typically consist of a matched sample (or a "paired" sample) or one group that has been tested twice (repeated measures).

Dependent *t*-tests are also used for **matched-paired samples**, where two groups are matched on a particular variable. For example, if we examined the heights of men and women in a relationship, the two groups are matched on relationship status. This would call for a dependent *t*-test because it is a paired sample (one man paired with one woman). Alternatively, we might recruit 100 men and 100 women, with no relationship between any particular man and any particular woman; in this case we would use an independent samples test.

Another example of a matched sample would be to take two groups of students, match each student in one group with a student in the other group based on an achievement test result, then examine how much each student reads. An example pair might be two students that score 90 and 91 or two students that scored 45 and 40 on the same test. The hypothesis would be that students that did well on the test may or may not read more. Alternatively, we might recruit students with low scores and students with high scores in two groups and assess their reading amounts independently.

An example of a **repeated measures** *t*-test would be if one group were pre- and post-tested. (This example occurs in education quite frequently.) If a teacher wanted to examine the effect of a new set of textbooks on student achievement, (s)he could test the class at the beginning of the year (pretest) and at the end of the year (posttest). A dependent *t*-test would be used, treating the pretest and posttest as matched variables (matched by student).

Calculations

Independent one-sample *t*-test

This equation is used to compare one sample mean to a specific value μ_0.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Where *s* is the grand standard deviation of the sample. *n* is the sample size. The degrees of freedom used in this test is *n* " 1.

Independent two-sample *t*-test

Equal sample sizes, equal variance

This equation is only used when both:

- the two sample sizes (that is, the *n* or number of participants of each group) are equal;
- it can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below.

The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$S_{X_1X_2} = \sqrt{\frac{S_{X1}^2 + S_{X2}^2}{2}}.$$ [1]

Here $S_{X_1X_2}$ is the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two. The denominator of *t* is the standard error of the difference between two means. For significance testing, the degrees of freedom for this test is 2*n* " 2 where *n* = # of participants in each group.

Unequal Sample Sizes, Equal Variance

This equation is only used when it can be assumed that the two distributions have the same variance. (When this assumption is violated, see below.) The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$S_{X_1X_2} = \sqrt{\frac{n_1 S_{X1}^2 + n_2 S_{X2}^2}{n_1 + n_2}} \quad [2]$$

Note that the formulae above are generalizations for the case where both samples have equal sizes (substitute n_1 and n_2 for n and you'll see).

$S_{X_1X_2}$ is the unbiased estimator of the variance of the two samples, n = number of participants, 1 = group one, 2 = group two. n " 1 is the number of degrees of freedom for either group, and the total sample size minus 2 ($n_1 + n_2$ " 2) is the total number of degrees of freedom, which is used in significance testing.

The statistical significance level associated with the t value calculated in this way is the probability that, under the null hypothesis of equal means, the absolute value of t could be that large or larger just by chance—in other words, it's a two-tailed test, testing whether the means are different when, if they are, either one may be the larger (see Press et al, 1999, p. 616).

Unequal Sample Sizes, Unequal Variance

This equation is only used when the two sample sizes are unequal and the variance is assumed to be different. See also Welch's t test. The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where s^2 is the unbiased estimator of the variance of the two samples, n = number of participants, 1 = group one, 2 = group two. Note that in this case, is not a pooled variance. For use in significance testing, the distribution of the test statistic is approximated as being an ordinary Student's t distribution with the degrees of freedom calculated using:

$$\text{D.F.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

This equation is called the Welch-Satterthwaite equation. Note that the true distribution of the test statistic actually depends (slightly) on the two unknown variances: see Behrens-Fisher problem.

This test can be used as either a one-tailed or two-tailed test.

Dependent t-test

This equation is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired".

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{N}}.$$

For this equation, the differences between all pairs must be calculated. The pairs are either one person's pre-test and post-test scores or between pairs of persons matched into meaningful groups (for instance drawn from the same family or age group: see table). The average (X_D) and standard deviation (s_D) of those differences are used in the equation. The constant ì$_0$ is non-zero if you want to test whether the average of the difference is significantly different than ì$_0$. The degree of freedom used is N-1.

Example of matched pairs

Pair	Name	Age	Test
1	Jon	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

Example of repeated measures

Number	Name	Test 1	Test 2
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%

Example

A random sample of screws have weights

30.02, 29.99, 30.11, 29.97, 30.01, 29.99

Calculate a 95% confidence interval for the population's mean weight.

Assume the population is distributed as $N($, ó^2).

The samples' mean weight is 30.015 with standard deviation of **0.0497**. With the mean and the first five weights it is possible to calculate the sixth weight. Consequently there are five degrees of freedom.

We can lookup in the table that for a confidence range of 95% and five degrees of freedom, the value is 2.571.

i.e.

$$30.015 - 2.571\frac{0.0497}{\sqrt{6}} < \mu < 30.015 + 2.571\frac{0.0497}{\sqrt{6}}$$

$$29.96 < \mu < 30.07.$$

If we sampled many times, our interval would capture the true mean weight 95% of the time; thus, we are 95% confident that the true mean weight of all screws will fall between 29.96 and 30.07

Alternatives to the *t* test

Recall that the *t* test can be used to test the equality of the means of two normal populations with unknown, but equal, variance.

· To relax the normality assumption, a non-parametric alternative to the *t* test can be used, and the usual choices are:
  - o for independent samples, the Mann-Whitney U test
  - o for related samples, either the binomial test or the Wilcoxon signed-rank test

· To test the equality of the means of more than two normal populations, an Analysis of Variance can be performed

· To test the equality of the means of two normal populations with known variance, a Z-test can be performed

History

The *t* statistic was introduced by William Sealy Gosset for cheaply monitoring the quality of beer brews ("Student" was his pen name)[3]. Gosset was a statistician for the Guinness brewery in Dublin, Ireland, and was hired due to Claude Guinness's innovative policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness' industrial processes[3]. Gosset published the *t* test in Biometrika in 1908, but was forced to use a pen name by his employer who regarded the fact that they were using statistics as a trade secret. In fact, Gosset's identity was known not only to fellow statisticians but to his employer — the company insisted on the pseudonym[4] so that it could turn a blind eye to the breach of its rules.

Today, it is more generally applied to the confidence that can be placed in judgments made from small samples.

Most spreadsheet programs and statistics packages, such as DAP, gretl, R and PSPP include implementations of Student's t-test.

Meta-analysis

In statistics, a meta-analysis combines the results of several studies that address a set of related research hypotheses. The first meta-analysis was performed by Karl Pearson in 1904, in an attempt to overcome the problem of reduced statistical power in studies with small sample sizes; analyzing the results from a group of studies can allow more accurate data analysis.

Although meta-analysis is widely used in epidemiology and evidence-based medicine today, a meta-analysis of a medical treatment was not published until 1955. In the 1970s, more sophisticated analytical techniques were introduced in educational research, starting with the work of Gene V. Glass, Frank L. Schmidt and John E. Hunter.

The online Oxford English Dictionary lists the first usage of the term in the statistical sense as 1976 by Glass. The statistical theory surrounding meta-analysis was greatly advanced by the work of Nambury S. Raju, Larry V. Hedges, Harris Cooper, Ingram Olkin, John E. Hunter, Jacob Cohen, and Frank L. Schmidt.

Uses in Modern Science

Because the results from different studies investigating different independent variables are measured on different scales, the dependent variable in a meta-analysis is some standardized measure of effect size. To describe the results of comparative experiments the usual effect size indicator is the standardized mean difference (*d*) which is the standard score equivalent to the difference between means, or an odds ratio if the outcome of the experiments is a dichotomous variable (success versus failure). A meta-analysis can be performed on studies that describe their findings in correlation coefficients, as for example, studies of the correlation between familial relationships and intelligence. In these cases, the correlation itself is the indicator of the effect size.

The method is not restricted to situations in which one or more variables is defined as "dependent." For example, a meta-analysis could be performed on a collection of studies each of which attempts to estimate the incidence of left-handedness in various groups of people.

Researchers should be aware that variations in sampling schemes can introduce heterogeneity to the result, which is the presence of more than one intercept in the solution. For instance, if some studies used 30mg of a drug, and others used 50mg, then we would plausibly expect two clusters to be present in the data, each varying around the mean of one dosage or the other. This can be modelled using a "random effects model."

Results from studies are combined using different approaches. One approach frequently used in meta-analysis in health care research is termed 'inverse variance method'. The average effect size across all studies is computed as a *weighted mean*, whereby the weights are equal to the inverse variance of each studies' effect estimator. Larger studies and studies with less random variation are given greater weight than smaller studies. Other common approaches include the Mantel Haenszel method and the Peto method. A

free Excel-based calculator to perform Mantel Haenszel analysis is available at: http://www.pitt.edu/~super1/lecture/lec1171/014.htm. They also have a free Excel-based Peto method calculator at : http://www.pitt.edu/~super1/lecture/lec1171/015.htm

Cochraine and other sources provide a useful discussion of the differences between these two approaches.

Q : Why not just add up all the results across studies ?

Answer : There is concern about Simpson's paradox. Note, however that Mantel Haenszel analysis and Peto analysis introduce their own biases and distortions of the data results.

A recent approach to studying the influence that weighting schemes can have on results has been proposed through the construct of *gravity*, which is a special case of combinatorial meta analysis.

Modern meta-analysis does more than just combine the effect sizes of a set of studies. It can test if the studies' outcomes show more variation than the variation that is expected because of sampling different research participants. If that is the case, study characteristics such as measurement instrument used, population sampled, or aspects of the studies' design are coded. These characteristics are then used as predictor variables to analyze the excess variation in the effect sizes. Some methodological weaknesses in studies can be corrected statistically. For example, it is possible to correct effect sizes or correlations for the downward bias due to measurement error or restriction on score ranges.

Meta analysis leads to a shift of emphasis from single studies to multiple studies. It emphasises the practical importance of the effect size instead of the statistical significance of individual studies. This shift in thinking has been termed Metaanalytic thinking.

The results of a meta-analysis are often shown in a forest plot.

Weaknesses

A weakness of the method is that sources of bias are not controlled by the method. A good meta-analysis of badly designed studies will still result in bad statistics. Robert Slavin has argued that only methodologically sound studies should be included in a meta-analysis, a practice he calls 'best evidence meta-analysis'. Other meta-analysts would include weaker studies, and add a study-level predictor variable that reflects the methodological quality of the studies to examine the effect of study quality on the effect size. Another weakness of the method is the heavy reliance on published studies, which may increase the effect as it is very hard to publish studies that show no significant results. This publication bias or "file-drawer effect" (where non-significant studies end up in the desk drawer instead of in the public domain) should be seriously considered when interpreting the outcomes of a meta-analysis. Because of the risk of publication bias, many meta-analyses now include a "failsafe N" statistic that calculates the number of studies with null results that would need to be added to the meta-analysis in order for an effect to no longer be reliable.

Forest Plot

From Wikipedia, the Free Encyclopedia

A **forest plot** is a graphical display that shows the strength of the evidence in quantitative scientific studies. It was developed for use in medical research as a means of graphically representing a meta-analysis of the results of randomized controlled trials. In the last twenty years, similar meta-analytical techniques have been applied in observational studies (e.g. environmental epidemiology) and forest plots are often used in presenting the results of such studies also.

Although forest plots can take several forms, they are commonly presented with two columns. The left-hand column lists the names of the studies (frequently randomized controlled trials or epidemiological studies), commonly in chronological order from the top downwards. The right-hand column is a plot of the measure of effect (*e.g.* an odds ratio) for each of these studies (often represented by a square) incorporating confidence intervals represented by horizontal lines. The graph may be plotted on a natural logarithmic scale when using odds ratios, so that the confidence intervals are symmetrical about the means from each study. The size of each square is proportional to the study's weight in the meta-analysis. The overall meta-analysed measure of effect is represented on the plot as a vertical line. This meta-analysed measure of effect is commonly plotted as a diamond, the lateral points of which indicate confidence intervals for this estimate.

A vertical line representing no effect is also plotted. If the confidence intervals for individual studies overlap with this line, it demonstrates that at the given level of confidence their effect sizes do not differ from no effect. The same applies for the meta-analysed measure of effect: if the points of the diamond overlap the line of no effect the overall meta-analysed result cannot be said to differ from no effect at the given level of confidence.

Forest plots date back to at least the 1970s, although the first use in print may be 1996.[1] The name refers to the forest of lines produced. In September 1990, Richard Peto joked that the plot was named after a breast cancer researcher called Pat Forrest and the name has sometimes been spelt "**forrest plot**".[1]