Statistical literacy guide
Confidence intervals and statistical significance

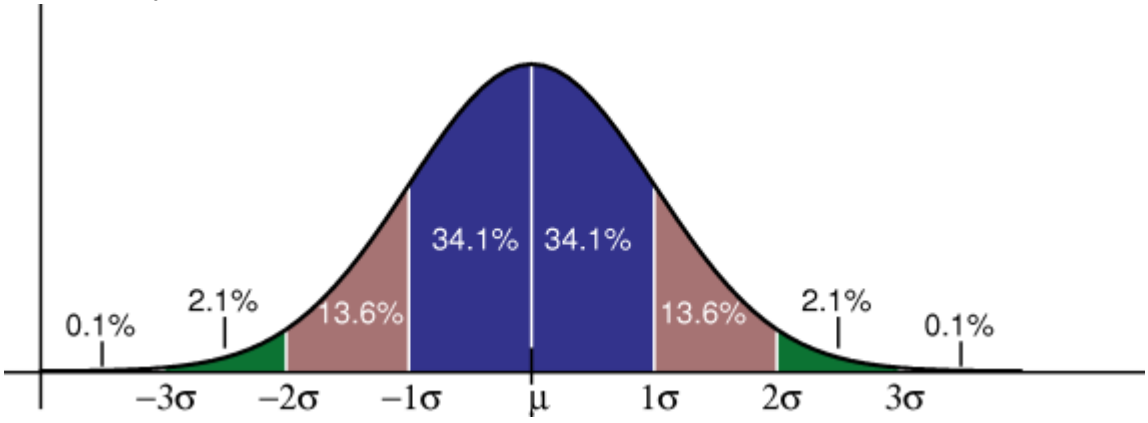
Last updated: June 2009
Author: Ross Young & Paul Bolton

This guide outlines the related concepts of confidence intervals and statistical significance and gives examples of the standard way that both are calculated or tested.

In statistics it is important to measure how confident we can be in the results of a survey or experiment. One way of measuring the degree of confidence in statistical results is to review the **confidence interval** reported by researchers. Confidence intervals describe the range within which a result for the whole population would occur for a specified proportion of times a survey or test was repeated among a sample of the population. Confidence intervals are a standard way of expressing the statistical accuracy of a survey-based estimate. If an estimate has a high error level, the corresponding confidence interval will be wide, and the less confidence we can have that the survey results describe the situation among the whole population.

It is common when quoting confidence intervals to refer to the **95% confidence interval** around a survey estimate or test result, although other confidence intervals can be reported (e.g. 99%, 90%, even 80%). Where a 95% confidence interval is reported then we can be reasonably confident that the range includes the 'true' value for the population as a whole. Formally we would expect it to contain the 'true' value 95% of the time.

The calculation of a confidence interval is based on the characteristics of a **normal distribution.** The chart below shows what a normal distribution of results is expected to look like. The mean value of all cases sampled in the survey is shown with the symbol μ and each standard deviation is shown with the symbol σ. Just over 95% of the distribution lies within 2 standard deviations of the average (mean). Thus if it can be shown, or is assumed, that the statistic is distributed normally then the confidence interval is the mean ±1.96 multiplied by the standard deviation.



So how do we work out what the **standard deviation** is? Standard deviation measures the *spread* of the results and, technically, is calculated as the square root of the variance.

Variance (σ^2) is calculated as the average squared deviation of each result from the average (mean) value of all results from our survey. The example below shows how we can work out the standard deviation.

A survey of 10 people asked respondents their current salary. These were the results:

£16,500
£19,300
£25,400
£23,200
£35,100
£46,000
£29,000
£38,200
£65,700
£49,200

The mean (average) salary is £34,760, calculated as adding together all the salaries and dividing the total by the number of cases. We calculate the variance (σ^2) by squaring the sum of the mean salary subtracted from each individual's reported salary, totaling these together, and dividing the total by the number of cases in our survey.

$$\sigma^2 = \frac{\sum (16{,}500-34{,}760)^2+(19{,}300-34{,}760)^2 \ldots}{10 \text{ (i.e. number of cases)}}$$

Otherwise, £2,130,944,000 divided by 10. The variance value is £213,094,400

The standard deviation is calculated as the square root of the variance, hence
$\sqrt{}$ £213,094,000, or £14,598

Where survey or test results are based on a larger sample size or results are less variable then the confidence interval will be smaller, other things equal.

In everyday language we often use the term "significant" to mean important and this normally involves some judgement relevant to the field of study (here **substantive significance**). However, in statistical terminology "significant" means probably true and probably not due to a chance occurrence (**statistical significance**). A finding may be statistically significant without being important or substantively significant. For instance in very large samples it is common to discover many statistically significant findings where the sizes of the effects are so small that they are meaningless or trivial.

Significance testing is normally used to establish whether a set of statistical results are likely to have occurred by chance. This may be to test whether the difference between two averages (mortality for pill 1 v mortality for pill 2) is 'real' or whether there a relationship between two or more variables. In the latter case these relationships are often expressed in terms of the **dependent variable** and one or more **independent variables**. The dependent variable is the variable against which certain effects are measured. The independent variables are those that are being tested to see what extent, if any, they have on the dependent variable. For example, we may wish to test how media coverage of political parties and electors' past voting records determine an individual's likelihood to vote for any one political party. In this case, media coverage and past voting record are the independent

variables, and their effect would be measured in terms of the dependent variable, in this case reported voting intention at the next general election.

The key to most significance testing is to establish the extent to which the **null hypothesis** is believed to be true. The null hypothesis refers to any hypothesis to be nullified and normally presumes chance results only –no difference in averages or no correlation between variables. For example, if we undertook a study of the effects of consuming alcohol on the ability to drive a car by asking a sample of people to perform basic driving skills while under the influence of large quantities of alcohol, the null hypothesis would be that consuming alcohol has *no* effect on an individual's ability to drive.

In statistics, a result is said to be statistically significant if it is unlikely to have occurred by chance. In such cases, the null hypothesis cannot be shown to be true. The most common significance level to show that a finding is good enough to be believed is 0.05 or 5%. This mans that there is a 5% chance of the observed data (or more extreme data) occurring given that the null hypothesis is true. Where findings meet this criteria it is normally inferred that the null hypothesis is false. While the 5% level is standard level used across most social sciences, the 1% level ($p<0.01$) is also fairly common.

When a null hypothesis is rejected, but it is actually true, a so-called **type I error** has occurred. In most cases this means that there is no correlation between the variables, but the test indicates that there is. Much null hypothesis testing is aimed at reducing the possibility of a type I error by reducing the p value and testing against a lower significance level. The aim of this is to reduce the possibility of falsely claiming some connection; a 'false positive' finding.

A **type II error** occurs when a false null hypothesis is accepted/not rejected. In most cases this will mean that results are not down to chance alone, there is a correlation between the variables, but the test did not detect this and gives a 'false negative' finding. The **power** of a test is one minus the type II error rate and is the probability of correctly rejecting a false null hypothesis (a 'true positive' finding). A higher power raises the chances that a test will be conclusive. It is not common for the type II error rate or power to be calculated in significance testing. Convention in many areas of social science especially is that type II errors are preferable to type I errors. There is a trade off between type I and II errors as the former can be reduced by setting a very low significance level ($p<0.01$ or $p<0.001$) but this increases the likelihood that a false null hypothesis will not be rejected.

To revisit our example**,** a survey of 10 people asked respondents their current salary but also their age, in order to investigate whether age (independent variable) has an effect on salary (dependent variable). These were the results:

Age	Salary
18	£16,500
25	£19,300
31	£25,400
33	£23,200
39	£35,100
41	£46,000
49	£29,000
52	£38,200
58	£65,700
63	£49,200

Here our null hypothesis is that age has *no* effect on salary (and the significance level is 5%. Using a simple linear regression[1] (of the type Income =β multiplied by age plus α) we get a β value of £880 –income increased on average by £880 for each additional year of age. The p value for this statistics was 0.0025. Thus the result is statistically significant at the 5% level and we reject the null hypothesis of no connection between age and salary. While the actual value of β is not always reported it helps the author start to establish importance or substantive significance if they report it. Just as important is the confidence interval of the estimate. The 95% confidence interval of β in this example is £410-£1,360, or we expect that the true value would fall in this range 95% of the time. This tells the reader both about the size of the effect and illustrates the level of uncertainty of the estimate.

There is a precise link between significance levels and confidence intervals. If the 95% confidence interval includes the value assumed for the null hypothesis (here zero) then $p \geq 0.05$ and the null hypothesis is not rejected at the 5% level. Similarly if the 99% confidence interval included zero then the hypothesis would not be rejected at the 1% level.

The type of null hypothesis testing outlined in this note is that which most readers are likely to find in the social sciences and medicine. The ritualised nature of some of this significance testing, misinterpretation of results, non-reporting of the size of coefficients, focus on random error at the expense of other sources of error, absence of alternative hypotheses and ignorance of alternative types of significance testing has been criticism by some authors.[2] The most important criticism is the equivalence that some authors are said to see between statistical significance and substantive significance. The sole focus on the presence of an effect, not what size/how important it is. Statistical significance is not sufficient for substantive significance in the field in question. It may also not be necessary in certain circumstances.

[1] See the Basic outline of regression analysis guide for more background

[2] See for instance Gerd Gigerenzer, *Mindless Statistics*, The Journal of Socio-Economics 33 (2004) 587–606. http://www.mpib-berlin.mpg.de/en/institut/dok/full/gg/mindless/mindless.pdf

Other statistical literacy guides in this series:

- What is a billion? and other units
- How to understand and calculate percentages
- Index numbers
- Rounding and significant places
- Measures of average and spread
- How to read charts
- How to spot spin and inappropriate use of statistics
- A basic outline of samples and sampling
- Confidence intervals and statistical significance
- A basic outline of regression analysis
- Uncertainty and risk
- How to adjust for inflation