Tutorials in Quantitative Methods for Psychology
2006, Vol 2 (1), p. 11-19

DOI: 10.20982/tqmp.02.1.p011

Confidence Intervals: From tests of statistical significance to confidence intervals, range hypotheses and substantial effects

Dominic Beaulieu-Prévost
Université de Montréal

For the last 50 years of research in quantitative social sciences, the empirical evaluation of scientific hypotheses has been based on the rejection or not of the null hypothesis. However, more than 300 articles demonstrated that this method was problematic. In summary, null hypothesis testing (NHT) is unfalsifiable, its results depend directly on sample size and the null hypothesis is both improbable and not plausible. Consequently, alternatives to NHT such as confidence intervals (CI) and measures of effect size are starting to be used in scientific publications. The purpose of this article is, first, to provide the conceptual tools necessary to implement an approach based on confidence intervals, and second, to briefly demonstrate why such an approach is an interesting alternative to an approach based on NHT. As demonstrated in the article, the proposed CI approach avoids most problems related to a NHT approach and can often improve the scientific and contextual relevance of the statistical interpretations by testing range hypotheses instead of a point hypothesis and by defining the minimal value of a substantial effect. The main advantage of such a CI approach is that it replaces the notion of statistical power by an easily interpretable three-value logic (probable presence of a substantial effect, probable absence of a substantial effect and probabilistic undetermination). The demonstration includes a complete example.

Tests of statistical significance, also known as null hypothesis testing (NHT) have been highly criticized during the last decades. The APA Task force on statistical inference even suggested to avoid using NHT as much as possible and to replace it with alternative procedures such as confidence intervals and measures of effect size (Wilkinson et al., 1999). However, many introductory courses and manuals in statistics for psychology still teach NHT as the only paradigm. The purpose of this article is, first, to provide the conceptual tools necessary to implement an approach based on confidence intervals, and second, to briefly demonstrate why such an approach is an interesting alternative to an approach based on NHT.

Understanding confidence intervals

Confidence intervals are mathematically equivalent to tests of significance. Indeed, for every test of significance, an equivalent confidence interval can be constructed. However, instead of providing a p value to evaluate if an effect is statistically different from zero, confidence intervals provide information about the unstandardized effect size as observed in the sample (i.e. the effect size) and the precision of the estimation of the effect size for the population (i.e. the parameter). The basic model of an effect size is:

$$CI = ES \pm V_C \times SE \tag{1}$$

where the confidence interval (CI) is constructed by adding and subtracting from the unstandardized size of an effect as observed in the sample (ES) the product of its standard error (SE) and the two-tailed critical value at the chosen alpha level of statistical significance (V_C). Every value around the unstandardized effect size and between the upper and lower boundaries of the interval is included in the confidence

Département de psychologie, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec, CANADA H3C 3J7; electronic mail may be sent to: dominic.beaulieu-prevost@umontreal.ca.

interval. When the CI excludes zero, the equivalent test of significance is statistically significant and vice versa. Although the term *effect size* is now often used to refer strictly to standardized, or metric-free, indexes of the size of an effect as observed in a sample such as the *d* statistic (Rosenthal, 1994), it is used in the present article in its older and simpler form, i.e. to refer to the unstandardized size of an effect as observed in a sample (e.g. a difference between means or a correlation). The term *parameter* will be used to refer to the unstandardized size of an effect in the population.

A confidence interval can be conceptually defined as a range of plausible values for the corresponding parameter (i.e. for the unstandardized size of the effect in the population). We could also say that conclusions that a parameter lies within a CI will err in [corresponding alpha] of the occasions. However, to interpret CIs beyond these simple definitions, one has to clarify what is meant by the notion of probability. Indeed, there are two radically different ways to interpret CIs that are related to two different interpretations of probability. The most commonly taught (but least understood) interpretation comes from the *frequentist approach*. It is indeed the approach on which traditional CIs are based. According to this approach, probability represents a long-term relative frequency. More explicitly, if CIs could be calculated for an infinity of random samples coming from the same population, the parameter of the population would be included in [1-alpha] of them. However, when a single CI is interpreted, it is inadequate to say that there is a probability of 95% that the parameter is included in the CI. From a frequentist point of vue, it makes no sense to speak about probabilities for a specific CI, it either includes the parameter or it does not. The only meaning that can be given to a specific CI is as a representation of the amount of sampling error associated with that estimate within a specified level of uncertainty. It is thus said that all the values included in a CI can be considered to be equivalent with a level of confidence of [1-alpha].

Researchers and decision makers are often more interested to know the probability that a specific CI includes the related parameter than to measure the sampling error of their study. What they crave for is the probability from a *subjective approach* or, more simply, a reasonable estimation of the odds of being correct if they conclude that the parameter is included in a specific CI. Using that definition, probability takes place in the eye of the beholder, not in the empirical world. The subjective approach to probability is generally called the *bayesian approach* because it is mathematically based on Bayes' theorem. It is indeed possible to calculate a bayesian CI for which it can be reasonably assumed that there is [1-alpha] chances that the parameter is included. However, to adequately calculate such a CI, one has to take into account both the experiment's data and all the previous knowledge one has about that parameter. It is a process extremely similar to a meta-analysis, in which the resulting CI is calculated by combining the results of all the previous studies. There is still one case for which a bayesian CI coincides with its frequentist counterpart: It is when the bayesian CI is based on an agnostic prior, i.e. a judgment that one has no useful prior knowledge or belief about a parameter's possible value. It can thus be said that when only the experiment's data are taken into account to estimate a parameter (i.e. when an agnostic prior is postulated), a traditional CI represents an interval for which it is reasonable to assume that there is [1-alpha] chances that the parameter is included. By extension, the distribution related to the CI can be understood as the distribution of the probable values of the parameter according to an agnostic prior.

Calculating Confidence Intervals

As explained in the previous section, confidence intervals can be calculated for any traditional test of significance. Although the specific formula used to calculate a confidence interval depends on the type of data, these formulae are all based on the same general model (see equation 1).

The following section will present the specific formulae used to calculate the most commonly used confidence intervals for means, correlations, proportions and their differences. The other confidence intervals can often be constructed using the same general principle.

Confidence intervals for means

The CI equivalent of a *one sample T-test* uses the following formulae:

$$CI = \overline{X} \pm t_C \times S_{\overline{x}} \tag{2}$$

$$S_{\overline{x}} = \frac{S_x}{\sqrt{n}} = \sqrt{\frac{\sum (x - \overline{X})^2}{n(n-1)}} \tag{3}$$

where $\overline{X}$ is the observed mean; t_C is the t value corresponding to the alpha level; $S_{\overline{x}}$ is the standard error of the mean; S_x is the standard deviation; and *n* is the number of cases.

The CI equivalent of an *independent samples T-test* uses the following formulae:

$$CI = \overline{X}_2 - \overline{X}_1 \pm t_C \times S_{\overline{X}_2 - \overline{X}_1} \tag{4}$$

$$S_{\overline{X}_2 - \overline{X}_1} = \sqrt{\frac{S_{x_1}{}^2}{n_1} + \frac{S_{x_2}{}^2}{n_2}} = \sqrt{\frac{\sum(x_1 - \overline{X}_1)^2}{n_1(n_1 - 1)} + \frac{\sum(x_2 - \overline{X}_2)^2}{n_2(n_2 - 1)}} \tag{5}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the observed means of each group; t_C is the t value corresponding to the alpha level; $S_{\overline{X}_2 - \overline{X}_1}$ is the standard error of the difference between the means; ${S_{x_1}}^2$ and ${S_{x_2}}^2$ are the variances of each group; and n_1 and n_2 are the number of cases in each group. If the homogeneity of variances cannot be assumed, the formula is used as is. It is called the separate variances formula. However, if the homogeneity of variances is assumed (i.e. ${S_{x_1}}^2 = {S_{x_2}}^2$), the pooled variance (see equation 5) replaces both ${S_{x_1}}^2$ and ${S_{x_2}}^2$ in the standard error formula (i.e. equation 5).

$$S_{pooled}^{\ 2} = \frac{{S_{x_1}}^2(n_1-1)+{S_{x_2}}^2(n_2-1)}{n_1+n_2-2} \qquad (6)$$

The CI equivalent of a *paired samples T-test* uses the following formulae**:**

$$CI = \overline{D} \pm t_C \times S_{\overline{D}} \qquad (7)$$

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n}} \qquad (8)$$

where $\overline{D}$ is the observed mean of the differences between the paired observations; t_C is the t value corresponding to the alpha level; $S_{\overline{D}}$ is the standard error of the mean of the differences; S_D is the standard deviation; and n is the number of pairs. The procedure is exactly the same as for a *one sample T-test* except that the differences between the paired scores replace the scores.

Confidence intervals for correlations

Confidence intervals cannot be directly calculated for correlations because the distribution of probable values varies as a function of the size of the correlation. A simple solution to this problem, described by Fisher (1925), is to:
(a) transform the correlation (*r*) into Fisher's *z′* using the following formula:

$$z' = \frac{\ln(1+r)-\ln(1-r)}{2} \qquad (9)$$

(b) calculate the CI for the z' value using the following formulae:

$$CI = z' \pm z_C \times SE_{z'} \qquad (10)$$

$$SE_{z'} = \frac{1}{\sqrt{n-3}} \qquad (11)$$

where z' is the observed z' value; z_C is the z value corresponding to the alpha level; $SE_{z'}$ is the standard error of the z' ; and *n* is the number of pairs.
(c) and retransform the lower and upper boundaries of the CI in correlations using the following formula:

$$r = \frac{e^{2z'}-1}{e^{2z'}+1} = \tanh(z') \qquad (12)$$

where $\tanh(z')$ is the hyperbolic tangent of z' .

The basic procedure to calculate the CI of a *difference between two correlations with independent samples* uses the following formulae after each correlation has been transformed into a Fisher's *z′* (see equation 9):

$$CI = z'_2 - z'_1 \pm z_C \times SE_{z'_2 - z'_1} \qquad (13)$$

$$SE_{z'_2 - z'_1} = \sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}} \qquad (14)$$

where $z'_2 - z'_1$ is the observed difference between the two *z′* values; z_C is the z value corresponding to the alpha level; $SE_{z'_2 - z'_1}$ is the standard error of the difference between the correlations; and n_1 and n_2 are the number of pairs for each sample.

One problem with the resulting CI is that it cannot be directly interpreted. Indeed, although the resulting boundaries of the CI could be transformed back into correlations, the result would be meaningless, first, because a difference between two Fisher's *z′* values does not directly translate into a difference between two correlations and, second, because a difference between two correlations does not directly translate into a specific amount of difference in explained variance. One way to transform the CI of a difference between two Fisher's *z′* values into a CI of difference in explained variance is to use an adaptation of Tryon's (2001) *inferential confidence intervals*. Inferential confidence intervals are mathematically equivalent to standard confidence intervals of differences and, consequently, to tests of statistical significance. However, they represent the statistical significance of a difference as confidence intervals around each effect size. The difference is said to be statistically significant (i.e. excluding zero) when the confidence intervals are not overlapping and non-significant when they overlap. Technically, to obtain the difference in explained variance for each boundary of a confidence interval in a difference between correlations, the two corresponding inferential confidence intervals must first be calculated using the following formulae:

$$CI_{z'_1(\text{inferentia l})} = z'_1 \pm z_{C(\text{inferentia l})} \times SE_{z'_1} \qquad (15)$$

$$CI_{z'_2(\text{inferential})} = z'_2 \pm z_{C(\text{inferential})} \times SE_{z'_2} \qquad (16)$$

$$z_{C(\text{inferential})} = z_c \times \frac{SE_{z'_2 - z'_1}}{SE_{z'_1} + SE_{z'_2}} \qquad (17)$$

where z'_1 and z'_2 are the Fisher's *z′* calculated for each correlation (using equation 9); $SE_{z'_1}$ and $SE_{z'_2}$ are the standard errors of each z' (equation 11); $z_{C(\text{inferential})}$ is the inferential equivalent of the critical *z* value; z_C is the *z* value corresponding to the alpha level; and $SE_{z'_2 - z'_1}$ is the standard error of the difference (equation 14).

When the boundaries of the two inferential confidence intervals are calculated, each boundary has to be transformed into explained variance by transforming it into a correlation (using equation 12) and squaring the result. The two boundaries of the difference in explained variance can then be calculated (1) by subtracting the upper boundary of the first inferential confidence interval from the lower boundary of the second inferential confidence interval and (2) by subtracting the lower boundary of the first inferential confidence interval from the upper boundary of the second inferential confidence interval.

Confidence intervals for proportions

Confidence intervals for proportions are the CI equivalent of chi-squares. Technically, proportions and probabilities are based on the binomial distribution. However, the calculations for their CIs are generally based on the normal distribution because its continuous scale (as opposed to the discrete scale of the binomial distribution) makes it easier to use. Traditionally, these CIs were calculated using the following formula, called the simple asymptomatic approximation or Wald method (Vollset, 1993):

$$CI = P \pm z_C \times \sqrt{\frac{P(P-1)}{n}} \tag{18}$$

where P is the observed proportion; z_C is the z value corresponding to the alpha level; and n is the number of cases.

The simple asymptomatic approximation for a difference of proportions between two independent samples uses the following formulae:

$$CI = P_2 - P_1 \pm z_C \times \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} \tag{19}$$

where P_1 and P_2 are the observed proportions; z_C is the z value corresponding to the alpha level; and n_1 and n_2 are the number of cases for each proportion.

However, studies have shown that the simple asymptomatic method do not provide a very good approximation, especially when the number of cases is small aod/or the observed proportion is near 0 or 1. The Wilson method is an interesting method that avoids these problems (for a review, see Brown, Cai & DasGupta, 2001; Newcombe, 1998a; 1998b). *Wilson score CIs* for proportions use the following formulae:

$$CI = \frac{2nP + z_C^2}{2(n + z_C^2)} \pm z_C \times SE_P \tag{20}$$

$$SE_P = \sqrt{\frac{z_C^2 + 4nP(1-P)}{2(n + z_C^2)}} \tag{21}$$

where P is the observed proportion; n is the number of cases; z_C is the z value corresponding to the alpha level; and SE_P is the standard error of the proportion.

To calculate a *Wilson score CI* for a difference of proportions between two independent samples, one has first to calculate the *Wilson score CI* of each proportion (using equations 20 and 21) and then use the following formulae:

$$LB_{P_2-P_1} = P_2 - P_1 - \sqrt{(P_2 - LB_2)^2 + (UB_1 - P_1)^2} \tag{22}$$

$$UB_{P_2-P_1} = P_2 - P_1 - \sqrt{(P_1 - LB_1)^2 + (UB_2 - P_2)^2} \tag{23}$$

where $LB_{P2\text{-}P1}$ and $UB_{P2\text{-}P1}$ are, respectively, the lower and upper boundaries of the confidence interval of the difference; P_1 and P_2 are the observed proportions for each sample; LB_1 and LB_2 are the lower boundaries for each sample; and UB_1 and UB_2 are the upper boundaries for each sample.

Using Confidence Intervals to Validate Your Hypotheses

Why would you need an alternative to NHT?

The logic of null hypothesis testing (NHT) has been used for more than 50 years. However a growing number of criticisms (i.e. more than 300 articles) pinpoint major problems that question its usefulness as a general model of theory appraisal and its capacity to answer many of our research questions. Although an exhaustive coverage of problems related to NHT is impossible to include in this article because of space constraints, four classical problems will be briefly described below. Interested readers are referred to reviews of the topic (e.g. Kline, 2004; Beaulieu-Prévost, in press).

The relation to sample size. In a test of significance, the *p* value is the probability of having a result at least as "extreme" as the one observed IF WE SUPPOSE that the data are the result of a totally random process. It is thus simply an index of "surprise" and it is related both to the effect size and to the sample size. More specifically, the *p* value becomes smaller as the effect size increases and as the sample size increases. A problematic consequence is that a statistically significant result will ALWAYS be obtained if the sample is big enough, unless the effect size is EXACTLY zero. An irrelevant effect can thus be highly significant just because of sample size.

The lack of plausibility of the null hypothesis. A second major problem with significance testing is the lack of plausibility of the null hypothesis (H_0), especially in the "soft" sciences. This notion called the crud factor (Meehl, 1990), can be summarized by the following statement: *In the sciences of the living (i.e. from biology to sociology), almost all of the variables that we measure are correlated to some extent*. H_0 is thus rarely plausible. It is important to specify that the crud factor does not refer to random sampling error nor to measurement error. A resulting consequence of the situation is that the emergence of a statistically significant effect cannot be claimed as support for a specific theory because we should at least expect a small effect (e.g. correlation or difference) for most studies in the "soft" sciences.

The logical improbability of the null hypothesis. While there is only one specific parametric value associated with H_0 (i.e. zero), there is a range of possible parametric values associated with H_1 (i.e. anything except zero). H_0 is thus said to be a *point hypothesis* while H_1 is a *range hypothesis* (Serlin, 1993; Kline, 2004). The main problem with point hypotheses is that they are logically improbable on a continuous scale. Since a continuous scale is composed of an infinity of specific values (or points), there is only one chance out of the infinite that a specific point hypothesis is true. When defining point hypotheses as a special case of range hypotheses (i.e. hypotheses with the smallest possible range), the problem can be summarized by the following statement: *The precision of a hypothesis limits its logical probability of being true*. Indeed, restricting the range of possible values for a hypothesis reduces its probability of being true. If the logic is applied to significance testing, a major problem of the approach becomes obvious. As a point hypothesis on a continuous scale, H_0 is ALWAYS false, since $1/\infty$ can clearly be considered a negligible probability. Indeed, the probability that an intervention has an effect size of EXACTLY zero is infinitesimal. Therefore, H_1 is ALWAYS true and the concepts of type I and type II errors are nearly meaningless!

The unfalsifiability of the alternate hypothesis. *We can never prove a theory although we can refute it*. This statement that summarizes the limits of inductive inference is used since Fisher to justify the logic of significance testing. Since we cannot prove H_1, we will do our best to refute H_0. And it could have been an interesting idea if H_0 was not already known to be false! It is basically correct to argue that a statement cannot be inductively proven but that it can be refuted, but it is paradoxical to empirically test a statement's truth value when it is already known. As we have seen above, H_1 is always true because it includes the whole continuum of possible results (except one point). Furthermore, if we fail to reject H_0, we can always claim that the sample was not big enough. H_1 is thus unfalsifiable, which makes it a scientifically problematic hypothesis if we follow Popper's (1959) philosophy of science.

In conclusion. Using significance testing to appraise the validity of a scientific hypothesis implies using a decision criterion (i.e. the *p* value) that confounds effect size and sample size to test a hypothesis already known to be false and unrealistic. And when we successfully reject this false hypothesis, we can be tempted to infer that this test improves the plausibility/credibility of our "scientific" hypothesis, although such inference is generally unwarranted. Thus, significance testing as a general model of theory appraisal is, at least, a problematic procedure. However, the problem of significance testing is not so much in the statistical principles used to evaluate the probability of an event, but in the specific hypotheses that are systematically tested (i.e. H_0 and H_1). As will be shown in the following paragraphs, a CI approach to statistics is an interesting alternative to NHT that can be used to avoid most of these major problems.

Testing scientifically useful hypotheses

If we summarize, a scientifically useful hypothesis has to be probable, plausible and falsifiable. All point hypotheses (e.g. H_0) are thus scientifically problematic since they are improbable to the point of being false. Hypotheses that include every possible result except one (e.g. H_1) are also scientifically problematic since they are unfalsifiable. In fact, the only way to construct a probable and falsifiable hypothesis is to construct a range hypothesis that both includes and excludes a significant amount of possible results. This type of hypothesis has the best of both worlds: it is falsifiable because it excludes a significant amount of possible results and it is logically probable because it also includes a significant amount of possible results. As will be demonstrated in the next section, range hypotheses can be easily tested using a CI approach.

Even though an infinity of possible range hypotheses could be constructed, most scientifically meaningful hypotheses can be summarized by one of the following types: (1) There is (or not) a substantial effect, (2) There is (or not) a harmful effect and (3) There is (or not) a trivial effect.

To understand the meaning of these types of hypotheses, the notion of substantial effect has first to be clarified. Basically, the concept of "substantial effect" is the equivalent of "clinically significant effect" although it is not limited to clinical settings. A substantial effect is simply an effect whose size is large enough to be of interest. However, it is important to mention that the minimal value of a substantial effect is always context-dependent. To adequately quantify the minimal value of a substantial effect (or the maximal value of a trivial effect), one has to assess the important aspects of the study such as the theoretical importance of the effect, the practical purpose of the phenomenon, the potential cost of an intervention and, minimally, the sensitivity of the scale. For example, if the effect of an intervention on depression is measured with a depression scale from 1 to 10, it might be decided that an effect size of one would be the smallest interesting value since it is the smallest possible difference that can be detected by the scale. However, if the intervention is extremely costly, it might be decided that the effect size would need to be of at least 2.5 for the intervention to be interesting. Two different minimal values can often be quantified for the same study: The minimal value to consider that an effect is theoretically interesting and the minimal value to consider that an effect has a contextual usefulness. For example, if you are interested to investigate a potential link between self-esteem and school performance, you might be satisfied with correlations of 0.09 (i.e. 1% of explained variance) or more, but if you plan to increase school performance through a large-scale self-esteem intervention, you might evaluate that only correlations of at least 0.30 (i.e. 9% of explained variance) are deemed to be interesting. A major advantage of having to define the minimal value of a substantial effect is that it forces researchers to take into account the purpose of their study because such a value cannot be defined for meaningless studies.

As soon as the minimal value of a substantial effect is defined, the three possible types of hypotheses can automatically be defined:

1) The hypothesis of a substantial effect,

which evaluates whether or not the effect is at least equal to the minimal value of the substantial effect.

2) The hypothesis of a harmful effect,

which is defined as the opposite of the hypothesis of a substantial effect. It can be used to evaluate the possibilities of a harmful or counter-intuitive effect of substantial value.

3) The hypothesis of a trivial effect,

which evaluates whether or not the effect is between the minimal substantial effect and the minimal harmful effect. When this hypothesis is tested for a comparison between two means, it is also called a test of equivalence (see Rogers, Howard & Vessey, 1993) since it evaluates whether or not the means are substantially different.

Testing hypotheses with confidence intervals

As soon as adequate range hypotheses are defined and the CI is calculated, hypothesis testing can be done at a glance! You just have to see if the CI is either (1) totally included within the range of the hypothesis, (2) totally excluded from the range of the hypothesis or (3) partly included within the range of the hypothesis. If the CI is totally included, the hypothesis is *corroborated* (i.e. $p > 0.95$ if alpha = .05), if it is totally excluded, the hypothesis is *falsified* (i.e. $p < 0.05$ if alpha = .05) and if it is partly included, the hypothesis is *undetermined* (i.e. $0.05 < p < 0.95$ if alpha = .05). Since point hypotheses have a range of zero (or $1/\infty$ to be more precise), they are always too small to totally include a confidence interval. Consequently, point hypotheses can be falsified or undetermined but they can never be corroborated.

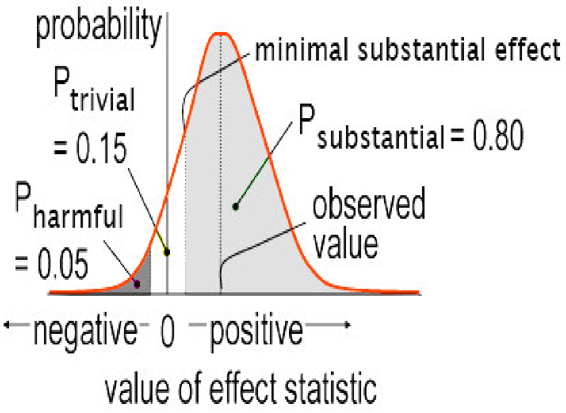
It is important to realize that this graphic method supposes that your hypotheses are two-tailed (i.e. they have both lower and upper boundaries for their range of expected values). If your hypotheses are one-tailed (e.g. $r > 0.30$), you can still use that method although it becomes slightly overconservative as it can falsely categorize a hypothesis as *undetermined*. For example, each boundary of a 95% two-tailed CI is the boundary of a 97.5% one-tailed CI because the 5% rejection area of a two-tailed CI is composed of a 2.5% rejection area on each side of the CI. To insure that the adequate probabilities are used to test your one-tailed hypotheses, one can simply calculate the one-tail values of your CI (by changing the critical value used in your calculations) and use the same logic as before.

If more precision is required, it is also possible to estimate the subjective probability of a hypothesis instead of simply comparing this probability to the 5%/95% criterion (i.e. the alpha). This type of precision is particularly appropriate with meta-analyses when the goal is to make a practical decision. Graphically speaking, the procedure requires, first, to project the values corresponding to the hypothesis on the distribution of probable scores for the parameter (i.e. the distribution used to construct the confidence interval) and, second, to quantify the area under the curve between the boundaries of the hypothesis (see Figure 1).

Mathematically speaking, the procedure requires that the values of the boundaries of the hypothesis (e.g. the minimal value of the substantial effect) be transformed into *z*-scores centered around the observed effect size. *T*-scores might have to be used instead of *z*-scores depending on the distribution on which the confidence interval was based. The second step is then to calculate the area under the curve. When the confidence interval is based on the *z* distribution (e.g. for correlations and proportions), one has simply to use a *z* distribution table to find the corresponding area under the curve. This table can be found in most introduction manual of statistics for the social sciences. When the confidence interval is based on the *t* distribution (e.g. for means), the procedure is a bit more complex because the shape of the *t* distribution depends on the degrees of freedom. The easiest way to solve the problem is to use an electronic calculator as found on the internet (e.g. West, R.W., 2006). A second option, when the number of cases exceeds 30, is simply to use the *z* distribution as an approximation of the *t* distribution. With this procedure, the precise value of the subjective probability can be calculated for each hypothesis (e.g. substantial effect, trivial effect, harmful effect).

The question of statistical power. The notion of undetermination answers the question of statistical power: If a hypothesis is undetermined, it simply means that the sample is not large enough to let the test provide a clear answer. As with tests of significance, the power of a test can also be estimated in advance. However, in the case of confidence intervals, the *sensitivity* of the test might be a more appropriate term than its *power*. The relevant question now becomes either "How much cases are required to make sure that an observed effect of X passes as a corroboration

Figure 1: Evaluation of the subjective probabilities for a fictional example



(or as a falsification) of hypothesis Y at an alpha level of Z?" or "Assuming X cases, what effect size will be considered as a corroboration (or as a falsification) of hypothesis Y with an alpha level of Z". To answer one of these questions, one has simply to isolate either n or the observed effect size in the equation used to calculate the CI. As with traditional tests of significance, the standard deviation has to be estimated a priori for means.

Confidence intervals and sample size. The relation between sample size and statistical power is important to understand when dealing with confidence intervals. Since both tests of significance and confidence intervals use the same statistical principles, sample size has the same general impact on them. As sample size increases, both the p value (in a NHT approach) and the width of the corresponding CI (in a CI approach) become smaller. This is explained by the fact that sample size directly affects the standard error of a test (e.g. equations 3, 5, 8, 11, 14 and 21). Point hypotheses will thus always be rejected if the sample size is big enough, unless the effect size in the sample corresponds exactly to the point hypothesis. From a CI point of vue, it can be said that as sample size increases, the bayesian probability of concluding that an effect is undetermined decreases. Since point hypotheses cannot be corroborated (as discussed earlier), they automatically become falsified when sample size is big enough. However, when range hypotheses are tested instead of point hypotheses, the problem of automatic falsification with large sample sizes disappears. Since range hypotheses include many possible parametric values (instead of only one for point hypotheses), they can both totally include or totally exclude the small-width CI of a large-sample study. Consequently, range hypotheses do not share with point hypotheses the problem of automatic falsification for large samples.

A complete example. Using the example presented above, a researcher wants to evaluate the correlation between self-esteem and school performance in a high school sample (n = 200). The purpose of this study is to evaluate whether or not a large-scale self-esteem intervention could improve school performance in a substantial way. After an evaluation of the situation (i.e. cost and efficiency of the intervention, sensitivity of the scales,…), the value of a contextually interesting effect is fixed at $r > 0.30$ and the value of a theoretically interesting effect is fixed at $r > 0.10$. Under these conditions, the *sensitivity* (or *statistical power*) of the test can already be calculated. It is thus calculated that an observed correlation of $r > 0.23$ will be necessary to corroborate the presence of a theoretically interesting effect while an observed correlation of $r > 0.42$ will be necessary to corroborate a contextually interesting effect. The researcher finds a correlation of 0.37 which he could traditionally report as $r = 0.37$ ($p < 0.05$). Since he is not interested in the null hypothesis, the researcher decides to estimate the parametric value of the correlation with confidence intervals ($0.25 < r < 48$).

Because the confidence interval includes both values below and above the minimal value of the contextually interesting effect, the hypothesis of a contextually substantial effect is undetermined (i.e. $0.05 < p < 0.95$). As for the presence of a theoretically interesting effect ($r > 0.10$), it is corroborated since the confidence interval is completely included in the range of theoretically interesting values. If need be, the precise subjective probabilities can also be calculated for both the contextually interesting effect ($p_{(\text{substantial})} = 0.86$; $p_{(\text{trivial})} = 0.14$; $p_{(\text{harmful})} = 0.00$) and/or the theoretically interesting effect ($p_{(\text{substantial})} = 1.00$; $p_{(\text{trivial})} = 0.00$; $p_{(\text{harmful})} = 0.00$). In this case, even though the presence of a contextually interesting effect was not corroborated, decision makers could still go on with the intervention if they evaluate that an 86% chances of producing a beneficial effect is a fair risk (especially since the chances of producing a harmful effect is negligible). As demonstrated by this example, the main advantage of testing for substantial effects with CIs over a NHT approach is that it that it immediately translates data into meaningful answers to research questions.

Discussion

The purpose of this article was both to demonstrate why an approach based on confidence intervals and range hypotheses is an interesting alternative to an approach based on NHT and to provide the conceptual tools necessary to make the transition from a NHT logic to a logic of confidence intervals and substantial effects. The research community is becoming more and more aware of the limits of NHT: CIs are now officially preferred over NHT by most APA journals and some scientific journals are now reluctant to publish traditional NHT studies unless CIs or standardized effect sizes are also presented (e.g. Memory & Cognition). Transforming your NHT habits into CI habits is relatively easy since both approaches are based on the same statistical model. Most of what you learned in your statistical courses is still relevant. And after reading this article, you already have all the tools you need to immediately start using CIs. In addition, a CI approach expands the possibilities of a NHT approach and improves the potential impact of a study because it allows you to easily test scientifically relevant hypotheses instead of automatically testing a single, false and unrealistic hypothesis (i.e. the null hypothesis). Such an approach will simply give you an edge in the "publish or perish" challenge as soon as you will adopt it.

Advices for an easy and enjoyable transition to a CI approach

Because most popular statistical softwares were created in a pro-NHT era, they always provide *p* values, they do not always provide their CI counterpart and they rarely provide a way to test range hypotheses. However, it is rarely a big problem since many solutions are easy to implement. A first solution is simply to copy the CI equations you want on a spreadsheet (e.g. Excel) and keep your spreadsheet for further use. You can then take the information you need from your statistical output (e.g. mean, standard deviation,…) and instantly translate your NHT result into a CI. A second solution is to find a spreadsheet or a CI calculator on the internet. CIs are becoming more and more popular and many researchers provide downloadable spreadsheets or online calculators to calculate them (e.g. Hopkins, 2006; Beaulieu-Prévost, 2006). These calculators can generally be found with an internet search using keywords such as "confidence intervals", either "spreadsheet" or "calculator", and a keyword describing the type of CI you want to calculate (e.g. "correlations"). Naturally, as with any information downloaded from the internet, you have to evaluate the quality of your calculator from the author's credibility or at least test your calculator with known data.

If you know the upper and lower boundaries of your CI but have no information about the standard error used in the calculations (as might happen with some statistical softwares), you might believe that you cannot calculate the precise subjective probability of a hypothesis. However, as long as you know the formula used to calculate the CI, the standard error can be calculated by isolating it in the CI formula. Using the general formula (see equation 1) as a model, the formula to calculate the standard error is:

$$SE = \frac{UB - ES}{V_C} \tag{24}$$

where SE is the standard error; UB is the upper boundary of the confidence interval; ES is the size of the effect in the sample; and V_C is the critical value for the specified alpha level.

Finally, if you are interested to calculate confidence intervals that were not presented in this article (e.g. CIs for linear regressions), you can either verify if your statistical software provides these confidence intervals, browse the internet for calculators or look for the formula in a statistical manual. Remember that for each test of significance, there is a corresponding CI. As soon as you have your CI, you can test your hypotheses using the approach presented in this article.

Towards more precise confidence intervals

The formulae presented in this article represent the basic CI formulae used in social sciences. It is important to realize that when CIs are built for complex distributions such as the *t* distribution, the CI calculated with these formulae represents an approximation. Recently, a more exact method called *noncentrality interval estimation* (Steiger & Fouladi, 1997) has been proposed to improve the precision of CIs for these complex distributions. However, such a method is mathematically demanding and user-friendly softwares that can handle these calculations are still rare. Other methods such as the bootstrap procedure are also proposed to improve the precision of CIs (Kline, 2004). We can probably expect these new algorithms to be included in our future statistical softwares as the CI approach gradually grows in popularity in social sciences.

Limits of a CI approach

As some readers might have realized, the article did not discussed CI equivalents of ANOVAs, i.e. situations in which more than two groups or two conditions are involved. CI equivalents of ANOVAs do exist but are very difficult to interpret because the metric used in these tests is not directly related to meaningful unstandardized units. Indeed, this problem is already acknowledged by NHT researchers: As soon as more than two groups or conditions are included in an ANOVA, a statistically significant result can only indicate whether or not there is at least one (statistically significant) difference. One has to use additional post-hoc tests to find out exactly where is this difference and, consequently, to adequately interpret the results. It is thus generally useless to use CI equivalents of ANOVAs. However, CI equivalents of different post-hoc tests are available and can be used as any other CI.

Conclusions

Null hypothesis testing systematically quantifies the plausibility of a "known-to-be false" hypothesis (H_0) to evaluate the validity of an unfalsifiable "known-to-be-true" alternate hypothesis (H_1). It is thus, at least, a problematic procedure if one wants to evaluate the validity of a scientific hypothesis. A CI approach that uses range hypotheses avoids these pitfalls and offers a very interesting alternative. Using such an approach, researchers can easily evaluate relevant scientific hypotheses by operationalizing them as falsifiable range hypotheses, estimating the minimal value of a substantial effect and constructing confidence intervals from their data. As demonstrated in this article, it is relatively easy to change your NHT habits into CI habits since both approaches are based on the same statistical model. As a last argument, remember that if you are more interested by the substantial significance of your results than by their statistical significance, such a change will improve the impact of your publications. Whether you see it as an alternative to a NHT approach or simply as an additional statistical tool, this approach can give you an edge over researchers using exclusively a traditional NHT approach!

You will find on the journal's web site an Excel spreadsheet with the simulated data used in the school performance example above on the first worksheet and a confidence intervals calculator for correlations on the second worksheet.

References

Beaulieu-Prévost, D. *Projet mémoire,* [Online]. http://www.projetmemoire.info/materiel.htm (Page visited march 1st, 2006).

Beaulieu-Prévost, Dominic. (in press). Statistical decision and falsification in science: Going beyond the null hypothesis! In Benoît Hardy-Vallée (Ed). *Cognitive decision-making: Empirical and foundational issues*. Cambridge Scholar Press.

Brown, L. B., Cai, T. T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science, 16*, 101-133.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* London: Oliver & Boyd.

Hopkins, Will G. *New view of statistics: Confidence limits*, [Online]. http://www.sportsci.org/resource/stats/generalize.html (Page visited march 1st, 2006).

Kline, R. B. (2004). *Beyond significance testing.* Washington: American Psychological Association.

Meelh, P. E. (1990). Why Summaries of Research on Psychological Theories Are Often Uninterpretable. *Psychological Reports, 66*, 195-244.

Newcombe, R. G. (1998a). Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine, 17*, 857-872.

Newcombe, R. G. (1998b). Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. *Statistics in Medicine, 17*, 873-890.

Popper, K. R. (1959). *The Logic of Scientific Discovery.* London, UK: Hutchison

Rogers, J. L., Howard, K. I. and Vessey, J. (1993). Using Significance Tests to Evaluate Equivalency Between Two Experimental Groups. *Psychological Bulletin, 113 (3)*, 553-565.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russel Sage Foundation.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for the Holm on the range. *Journal of Experimental Education, 61*, 350-360.

Steiger, J. H., and Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik and J.H. Steiger (Eds), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

Vollset, S. E. (1993). Confidence Intervals for a Binomial Proportion. *Statistics in Medicine, 12*, 809-824.

West, R. W. *Cybergnostic applets,* [Online]. http://www.stat.sc.edu/~west/applets (Page visited march 1st, 2006).

Wilkinson, A. and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Submitted March 8th, 2006
Corrections received April 25th, 2006
Accepted April 27th, 2006.