Psychologica Belgica
2011, 51-2, 109-127

HOW TO STATISTICALLY SHOW THE ABSENCE OF AN EFFECT

Etienne QUERTEMONT[1]
University of Liège

In experimental studies, the lack of statistical significance is often interpreted as the absence of an effect. Unfortunately, such a conclusion is often a serious misinterpretation. Indeed, non-significant results are just as often the consequence of an insufficient statistical power. In order to conclude beyond reasonable doubt that there is no meaningful effect at the population level, it is necessary to use proper statistical techniques. The present article reviews three different approaches that can be used to show the absence of a meaningful effect, namely the statistical power test, the equivalence test, and the confidence interval approach. These three techniques are presented with easy to understand examples and equations are given for the case of the two-sample *t*-test, the paired-sample *t*-test, the linear regression coefficient and the correlation coefficient. Despite the popularity of the power test, we recommend using preferably the equivalence test or the confidence interval.

State of the problem: absence of evidence is not evidence of absence

At the end of a study, it is not unusual to obtain a non-significant result. When this occurs, the lack of statistical significance is often interpreted as the absence of an effect (for example no difference between two groups, no relationship between two variables, …). Unfortunately, such a conclusion is often a serious misinterpretation of non-significant findings and not warranted. Such misinterpretation of non-significant results is quite common despite a substantial and established literature warning against it (e.g., Altman & Bland, 1995; Dunnett & Gent, 1977). A very simple example can illustrate the problem. Suppose a study in which reaction times to a specific category of stimuli were recorded with a computer. Ten participants were recruited and randomly divided into two groups. The experimental group was submitted to a specific experimental treatment before the recording of reaction times, while the control group was tested without the experimental treatment. Figure 1 shows the mean reaction times for both groups. In this simple experiment, as there is no evidence of a serious violation of the normality assumption, the difference between the means of the two groups is tested with a two-sample

1. Etienne Quertemont, PhD, Psychologie quantitative, University of Liège.
The author wishes to thank Serge Brédart, Marc Brysbaert, Christian Monseur and Francis Pérée for helpful discussions and advice on this article.
Correspondence concerning this article should be addressed to Etienne Quertemont, Psychologie quantitative, Université de Liège, Boulevard du Rectorat 5 / B32, 4000 Liège. E-mail: equertemont@ulg.ac.be

t-test, which provides the following non-significant difference: $t(8) = 2.07$; $p = 0.07$.

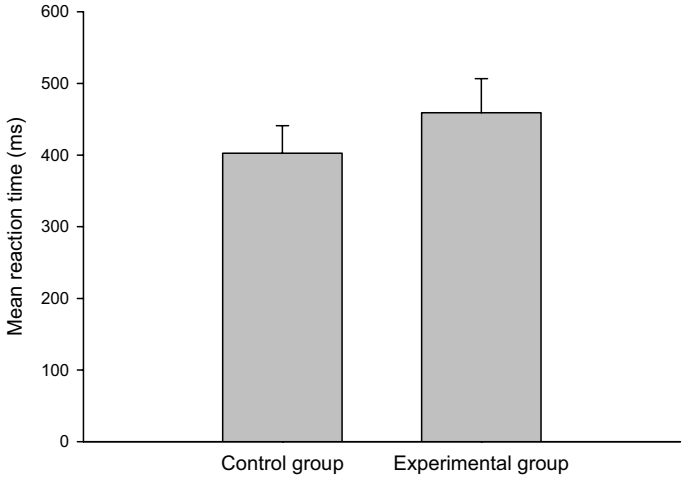

Figure 1

Results of a fictitious study comparing the reaction times to specific stimuli in a control and an experimental group (mean ± SD).

In the example, it would be a serious mistake to conclude that the experimental treatment had no effect on the reaction times. As shown on Figure 1, the mean difference between the two groups is 56 ms, which is quite substantial even if not statistically significant. Had the same difference (with a similar variance) occurred in a study with 20 participants in each group, the effect of the experimental treatment would have been highly significant ($p = 0.00019$). Under the assumption of no effect at the population level, the *p*-value of 0.07 only means that there is a 7% chance of getting a larger than 56 ms difference between the experimental and control group with two samples of 5 participants. In this example, the correct interpretation of the data, therefore, is a lack of sufficient evidence to conclude for an effect of the treatment. As absence of positive evidence of a difference is not positive evidence of equivalence, we are not allowed to conclude that the control and experimental groups are equivalent.

In any study, non-significant results can occur for three different reasons:

1. Mistakes have been made during the collection or encoding of the data, which mask otherwise significant results. This also includes measurement error (imprecision).

2. The study did not have enough statistical power to prove the existence of an otherwise real effect at the population level. The result is a "false equivalence", due to a sampling error.
3. There is actually no real effect (or a negligible effect) at the population level. The result is a "true equivalence".

As most investigators would be adamant that no mistakes were made in the collection of their data (and are much more likely to double check for such mistakes when the results go against their expectations than when they confirm the predictions), we can discard the first explanation and focus on the last two possible origins of non-significant results. With most non-significant results, it is often impossible to know whether they indicate a true equivalence or a false equivalence. In other terms, it is very difficult to distinguish between the absence of an effect at the population level and an insufficient statistical power in the study.

Often, investigators mistakenly believe that a very small effect in a sample, for example a very small difference or even a lack of difference between two groups, necessarily means the absence of an effect at the population level. For example, many investigators would probably interpret a correlation of 0.00 in a sample of ten participants as strong evidence for the absence of a relationship between two variables. Unfortunately, they forget that random sampling error may lead to the absence of an effect in a sample even when there is a substantial effect at the population level. For instance, with a sample of 10 participants, a correlation of 0.50 at the population level leads to sample correlations of 0.00 or lower in about 9% of the cases. Even a population correlation of 0.80 sometimes leads to a sample correlation of 0.00 or lower (this would happen in 1.7% of the samples with $n = 10$). Thus, even the total absence of an effect in a sample does not necessarily prove the absence of an effect at the population level, especially with small sample sizes.

In order to conclude beyond reasonable doubt that there is no meaningful effect at the population level, i.e., to show that the study reports a true equivalence, it is necessary to use proper statistical techniques. Before describing these techniques, it is important to keep in mind that it is impossible to show the total absence of an effect in the population. What the techniques below can show is the likelihood that the size of an effect in the population is lower than some value considered to be too low to be useful. To come back to our reaction time example, the techniques we discuss cannot show that the experimental treatment has no effect on reaction time at all. However, they can demonstrate that the effect of the treatment on reaction times is lower than, for example, 15 ms, which may be considered as negligible and non relevant for practical purposes. As a consequence, all the techniques used to show the absence of an effect start with defining what would be considered as the

threshold between a negligible and a useful effect. It is important to keep in mind that this threshold is defined on theoretical grounds and not from a statistical point of view. In our example, we will therefore first have to decide what would be a negligible difference. For instance, we could decide that all differences equal to or lower than 15 ms are unlikely to be of practical use. In such a case, if we can prove that the effect of the experimental treatment is unlikely to be higher than 15 ms at the population level, we can conclude that the group difference is negligible. In other words, the effect of the treatment would be scientifically and practically trivial. It is clear that a decision about the threshold is disputable and somewhat arbitrary. Different investigators may have slightly different definitions of what is a negligible effect for a particular study. However, by making the estimate available we are much more open about what we conclude from a null-effect, and divergent assumptions can be tested.

In the following paragraphs, we will review three different approaches to test the absence of an effect, namely the statistical power test, the equivalence test, and the confidence interval approach[2]. In order to illustrate the calculations, we will use the two examples summarised in Table 1. The first example reports the results of the fictitious study depicted in Figure 1. This example illustrates a non-significant result that may reflect a real effect at the population level, i.e., a possible false equivalence. Note, however, that it is impossible to conclude from these results whether or not there is a substantial difference at the population level. As we will see below, the first example illustrates a study from which it is impossible to conclude anything, due to the small sample size. The right conclusion for this study is "statistical indeterminacy". The second example illustrates, using the same scenario, data that are more compatible with a true equivalence, i.e., a study in which it is reasonable to conclude that there is no substantial effect at the population level.

Statistical power test

The statistical power test is probably the most popular method to show the lack of an effect in the case of non-significant results. It is also the test most often asked by reviewers, although it is not the most straightforward in terms

2. There are other, more sophisticated, methods, e.g., Bayesian and resampling methods, to test for the absence of a meaningful effect at the population level. However, for the sake of simplicity and because the present paper is aimed at the largest possible audience, we have limited the discussion to statistical approaches that can be adapted from the traditional hypothesis testing and confidence interval approaches. We also limit the discussion to a difference between two conditions, as this is what most conclusions in scientific research boil down to.

of calculation and interpretation. The test is based on a line of reasoning that can be summarised as follows:

1. Define what would be the minimal value of a useful effect size in your study. This is the threshold to be used for the power calculation.
2. Calculate the *a priori* probability of rejecting the null hypothesis if the threshold value were the actual effect size at the population level. This will provide you with an estimate of the power of your test.
3. If the power calculated under 2 is sufficiently high (usually defined as 0.80 or higher), you can conclude that chances of a population effect larger than the threshold value are very small given the null result you obtained. Therefore, you can conclude that the true effect size at the population level is unlikely to be of practical value and, therefore, that there is no real difference between the conditions.

Table 1
Results of two fictitious studies

	Example 1 **Possible false equivalence**	**Example 2** **True equivalence**
Number of subjects	n = 5 / group	n = 150 / group
Mean (± SD) for control group	402.61 (± 38.42)	402.98 (± 41.79)
Mean (± SD) for experimental group	459.09 (± 47.53)	407.24 (± 41.77)
Comparison of the two groups	$t(8) = 2.07$; $p = 0.07$	$t(298) = 0.88$; $p = 0.38$
Power	0.077	0.873
Equivalence test	$t(8) = 1.52$; $p = 0.92$	$t(298) = –2.23$; $p = 0.013$
CI for the difference between groups	–6.55 to 119.51	–5.23 to 13.76
Conclusion of the study	Not enough evidence to draw any conclusion	Experimental treatment has a trivial effect of no practical importance

CI: 95 % Confidence Interval

To illustrate the power calculation test, we apply it to the two examples in Table 1. We have defined above that we consider a difference in reaction times of 15 ms to be negligible and not relevant for practical purposes. This will be our minimal threshold value. Now we can calculate the power, i.e., the probability of rejecting the null hypothesis if the difference is postulated as 15 ms at the population level. Most of the common statistical software packages provide power calculation tests. You can also freely use power calculators on the internet. For instance, you can use Russ Lenth's power calculator at http://www.stat.uiowa.edu/~rlenth/Power/ (checked on April 7, 2011; use a search robot if the link no longer works). This website looks as shown in Figures 2 and 3 (make sure your computer supports Java).

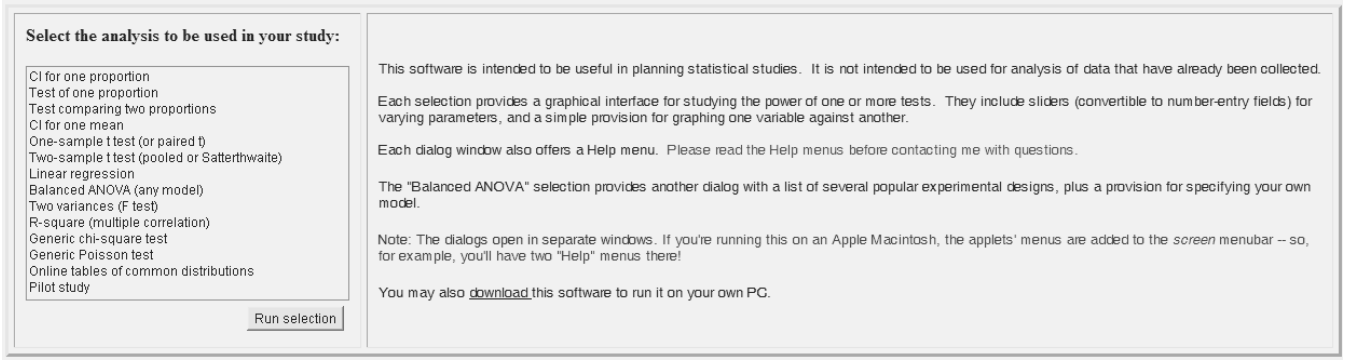

Figure 2

Front page of Russ Lenth's power calculator at http://www.stat.uiowa.edu/~rlenth/Power/

Select the option "two-sample *t*-test" in Figure 2. This gives you the panel of Figure 3, in which you can enter all values (click on the small rectangles next to each variable to enter numeric values if you do not want to work with the sliding rulers):

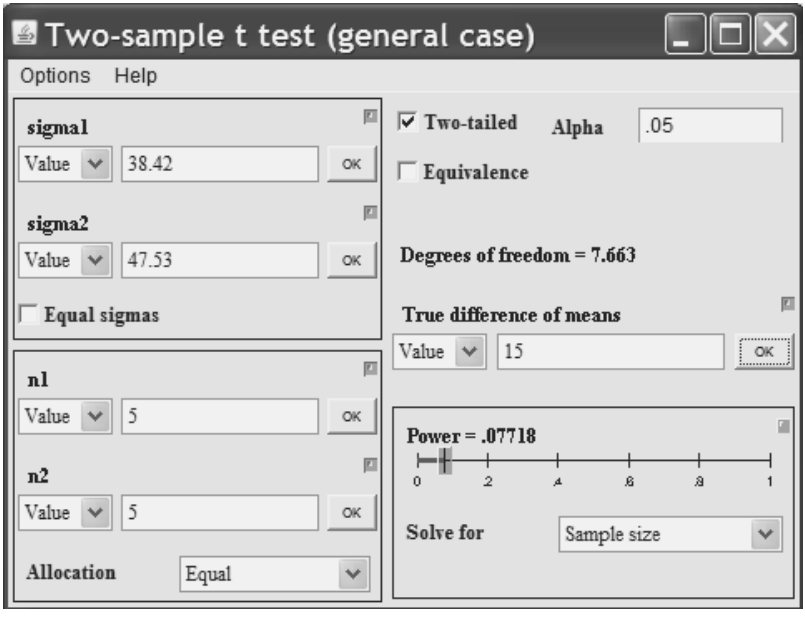

Figure 3
Two-sample t-test on Russ Lenth's power calculator at http://www.stat.uiowa.edu/~rlenth/Power/

From the panel in Figure 3 we can read the *a priori* chances of obtaining a statistically significant effect for a difference of 15 ms given standard deviations of 38.42 and 47.53 and sample sizes of 5. These chances were .077 or 7.7%.[3] We should never have run this study given its abysmal power to detect differences as low as 15 ms. Such a low power clearly indicates that we cannot conclude to the equivalence of the two groups on the basis of our study and, therefore, to the absence of an effect at the population level.

The same power analysis for the second example returns a value of .8727, meaning that *a priori* our study was expected to return a significant test result in 87.3% of the cases if the difference between the groups was 15 ms and the sample sizes were 150. Finding a null result in this situation is much more

3. The calculations presented here are based on a two-sided t test for independent samples.

informative about the likely effect size at the population level. Since the study led to non-significant results, we can reasonably conclude that the true population difference is below 15 ms, a difference we would interpret as negligible and not relevant for practical purposes.

Equivalence test

Although the statistical power test is best known[4], it is also possible to show the likelihood of trivially small effect sizes more easily within the traditional hypothesis testing approach (Parkhurst, 2001; Rogers, Howard, & Vessey, 1993). When we compare two groups with a traditional *t*-test, the aim is to show that there is a difference between the groups, and the null hypothesis posits that the group means are equivalent at the population level. An alternative is to define the null hypothesis in such a way that the difference between the groups is expected to be at the threshold value or above, while the alternative hypothesis asserts that the difference is below the threshold value. Such a test is called an *equivalence test*.

For our example, assuming that the threshold value for a practical difference in reaction times is 15 ms, the null and alternative hypotheses would be[5]:

$$H_0: \left| \mu_1 - \mu_2 \right| \geq 15 \text{ ms}$$
$$H_1: \left| \mu_1 - \mu_2 \right| < 15 \text{ ms}$$

where μ_1 is defined as the population mean for the experimental group and μ_2 as the population mean of the control group.

As in traditional hypothesis testing, the aim of our equivalence test is to reject the null hypothesis in order to conclude that the alternative hypothesis is demonstrated beyond reasonable doubt. In our example, if we can reject the null hypothesis defined above (i.e., that the difference between the two groups is higher than or equal to 15 ms), we are allowed to conclude that the mean difference between the groups at the population level is below 15 ms, a value defined as trivial. Therefore, the conclusion would be that the effect of the experimental treatment is practically or scientifically unimportant. In contrast, if the null hypothesis is not rejected, the conclusion would be that we do

4. At least in terms of its name, given that very few people really know how to calculate the power values.
5. Equivalence tests are usually based on the definition of an equivalence interval. In our example, we have defined a symmetrical interval from –15 ms to 15 ms. With such an interval, we define the two groups as equivalent if the difference between their means at the population level is lower than 15 ms and higher than –15 ms. However, it is possible to define different lower and upper limits for this interval, for instance an equivalence interval defined as –10 ms to 15 ms. In that case, the single test as described above would have to be replaced by two equivalence tests. See Rogers et al. (1993) for more details.

not have enough evidence to conclude that the effect of the treatment is unimportant.

In equivalence tests, the formula and calculations are identical to those in the traditional hypothesis tests, only the formulations of the null and alternative hypotheses change. For the *t*-test for independent samples, the formula is:

$$t = \frac{\left|(\overline{X_1} - \overline{X_2})\right| - \delta}{\sqrt{S_P^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}, \qquad df = N_1 + N_2 - 2$$

where $\overline{X_1}$ and $\overline{X_2}$ are the sample means of the two groups, N_1 and N_2 the sample sizes of the two groups, S_P^2 is the pooled variance of the two groups, and δ is the minimal threshold value as defined above. Note that the test works with the absolute value of the difference between the means, because it is a non-directional test. Indeed, equivalence means that neither of the groups is significantly higher than the other.

The precise calculations for various situations are given in the appendix. For example 1, we get the following result: $t(8) = 1.52$; $p = 0.92$. So, the null hypothesis cannot be rejected. Given the observed difference of 56 ms, we do not have enough evidence to conclude that the mean difference in reaction times between the experimental and the control group at the population level is unimportant, i.e., below 15 ms. For example 2, the results are quite different. We get $t(298) = -2.23$; $p = 0.013$. This allows us, on the basis of our observed difference of 4.3 ms, to reject the null hypothesis and to conclude that the mean difference between the groups at the population level is unlikely to be larger than 15 ms. Therefore, there is enough evidence to say that the effect of the experimental treatment is too small in terms of reaction times to have practical consequences.

The confidence interval approach

In line with the hypothesis testing approach, the aim of the above tests was to allow us to make yes/no decisions (Is the test powerful enough to find evidence for the threshold value? Is the evidence strong enough to reject a difference between the groups beyond the threshold value?). The main weakness of those approaches is that people may question the threshold value determined (i.e., Is a difference of 15 ms between the groups really trivial?).

An alternative approach consists of providing readers with the range of effect sizes that are likely to exist at the population level, given the outcome of the study. It involves computation of the confidence interval around the effect size obtained in the sample. This immediately gives readers information about the highest difference likely to exist at the population level. If the entire confidence interval falls below a sensible threshold value, it will be

possible to conclude that the true effect size at the population level is negligible.

A further advantage of the confidence interval approach is that nearly all statistical software packages compute confidence intervals for means, differences between means, correlations, regression coefficients, etc… (although one should be careful about repeated measures designs, as the confidence intervals of these often are calculated incorrectly; see Brysbaert, 2011; or Masson & Loftus, 2003). It is therefore very easy to use the confidence interval approach to examine whether a null-effect could be of practical consequence or not.

Table 1 shows that the 95% confidence interval for the difference between the control and experimental groups is –6.55 to 119.51 ms for example 1[6] (see the appendix for the calculations). Therefore, we are 95% confident that the true difference between the groups at the population level is between –6.55 ms (the mean reaction time of the experimental group is about 7 ms below that of the control group) and 119.51 ms (the mean reaction time of the experimental group is about 120 ms above that of the control group). As is clear from this confidence interval, it is dangerous to conclude on the basis of this evidence that there is no effect at all at the population level. Indeed, the confidence interval indicates that a difference as high as 119 ms could exist between both groups at the population level. This is well above the 15 ms we have defined as the threshold for a non-negligible effect size. At the same time, the confidence interval informs us that we do not have sufficient evidence to conclude that the experimental treatment has a real effect. Indeed, a difference of 0 ms is included within the confidence interval as well, which indicates that it is possible that there is actually no difference in the mean reaction time between the control and the experimental group. Clearly, this large confidence interval indicates that we cannot conclude anything from the results of example 1.

In contrast, the 95% confidence interval of example 2 is much smaller: –5.23 to 13.76 ms. Furthermore, both limits of the confidence interval are below the threshold value of 15 ms, which we defined as the minimum requirement for a meaningful effect. This means that we are allowed to conclude that the experimental treatment had no effect of practical importance. With our 95% confidence interval, the estimated difference between the two groups is at best 13.76 ms at the population level, which remains below the value we have defined as the threshold for practical or scientific importance.

6. To be operationally identical to the equivalence test, however, $1 - 2\alpha$ (not $1 - \alpha$) confidence interval should be computed. In the present case, a 90% confidence interval should be computed instead of a 95% confidence interval. See Schuirmann (1987) or Rogers et al. (1993) for explanations. However, this minor point does not notably change the general discussion of the confidence interval approach.

Relative to the other two approaches described above, the confidence interval approach has several advantages. The underlying reasoning is easier to understand than the reasoning behind the equivalence test and, most certainly, the reasoning behind the power test. Confidence intervals are also easy to compute as most statistical software packages have options to compute them. In contrast, many simple statistical software packages do not allow easy computation of equivalence tests. Finally, another main advantage of the confidence interval approach is that the investigators and their audience do not have to agree on the precise value of the threshold for a minimal effect size. In the power and equivalence tests, a threshold value must be introduced into the calculations and the outcome depends on the value entered. This threshold value can often be disputed and, therefore, seems like an arbitrary decision to the audience. With the confidence interval approach, readers can form their own opinion about whether the interval limits are small enough to be of no practical significance. The confidence interval is interpretable throughout its range, whereas the probabilities of the power test and the equivalence test critically depend on the threshold value chosen (with which the reader may disagree). For example 2, somebody defining the minimal threshold value at 10 ms would obtain the same confidence interval, but would reach a different conclusion, namely that the evidence is not strong enough to determine that there is no effect of practical value at the population level.

Conclusion

We have discussed three procedures to investigate whether an effect is negligible at the population level. Although they often lead to similar conclusions, the confidence interval approach is probably the easiest to implement and understand. Despite its popularity, the power test involves a convoluted reasoning that is frequently misunderstood, leading to misinterpreted results (Hoenig & Heisy, 2001). Furthermore, there is evidence that the power approach may be inferior to the other approaches (e.g., Schuirmann, 1987). We therefore recommend the use of either the equivalence test or the confidence interval to show the lack of a real effect at the population level. The procedures reviewed above were illustrated by an example in which the aim was to compare the means of two independent groups. These procedures can easily be extended to other statistical analyses (correlations, regression coefficients, analysis of variance …). Some examples are given in the Appendix (see also Goertzen & Cribbie, 2010 for equivalence tests for correlations).

To conclude, we would like to point to one important limitation of these three techniques. Similarly to most statistical procedures trying to demonstrate a small effect size, they usually require high (and sometimes huge) sample sizes to demonstrate the lack of a real effect. In our example 2, we needed

150 subjects in each group to show that the mean difference in reaction times between the groups was below 15 ms at the population level. If we had defined a lower threshold for negligible effects, e.g., 10 or 5 ms, the required sample size would have been even higher. To illustrate the importance of the sample size, Table 2 shows the required number of participants to obtain a power of 0.80 with various thresholds for negligible correlations. For example, if we define 0.1 as the threshold of a negligible correlation at the population level, we need 782 subjects to have a statistical power of 0.8 and therefore to convince the audience that there is only a trivial relationship between our variables in case of non-significant results. Such sample sizes are quite unusual in psychology studies. In many cases, a study in which the results are not significant will, therefore, not allow the researcher to conclude that there is a lack of an effect. In such conditions, the investigators have to recognise statistical indeterminacy and suspend their judgment as there is no evidence for or against anything (Tryon, 2001). The study provided no evidence of any kind and additional data will be required to conclude something. Unfortunately, such a conclusion of statistical indeterminacy is expected to be the default in studies reporting non-significant results based on small numbers of observations. Researchers should be aware that the sample sizes required for demonstrating the lack of a real effect may be prohibitively large in many cases.

Table 2

Minimal sample sizes required for a statistical power of 0.8 with various population correlations that might be defined as trivial

Population correlation	Sample size
0.05	3137
0.10	782
0.15	346
0.20	193

Note: the calculations are based on a two-sided *t*-test.

In the present article, we have adopted a "post-hoc" approach, in which equivalence tests were implemented at the end of a study when the investigator found non-significant results. However, when the aim of the study is to show the absence of an effect, it is also possible, and even recommendable, to plan an equivalence approach before the start of the study. In such cases, the investigator will be able to define the sample size properly, which often will mean testing a large number of participants. Note that in such studies with very large sample sizes, it is quite possible to obtain a significant effect with the traditional hypothesis testing approach together with a significant equivalence test. This would be interpreted as showing an effect at the population level, but an effect that is too small to be of practical value.

References

Altman, D.G., & Bland, J.M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 311*, 485-485.

Brysbaert, M. (2011). *Basic statistics for psychologists.* Basingstoke, UK: Palgrave Macmillan.

Dunnett, C.W., & Gent, M. (1977). Significant testing to establish equivalence between treatments with special reference to data in the form of 2 x 2 tables. *Biometrics, 33*, 593-602.

Goertzen, J.R., & Cribbie, R.A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*, 527-537.

Hoenig, J.M., & Heisy, D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*, 19-24.

Lenth, R.V. (2009). Java applets for power and sample size [Computer software]. Retrieved April 8, 2011, from http://www.stat.uiowa.edu/~rlenth/Power.

Masson, M.E.J., & Loftus, G.R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57*, 203-220.

Parkhurst, D.F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *Bioscience, 51*, 1051-1057.

Rogers, J.L., Howard, K.I., & Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553-565.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657-680.

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.

Appendix
Equations and calculations

The following paragraphs show the equations and calculations that are used to test for equivalence and to calculate the confidence intervals in the case of (1) the *t*-test for two independent means, (2) the *t*-test for paired samples, (3) simple linear regressions, and (4) Pearson's correlation coefficient. Example 1 from Table 1 is used to illustrate the calculations for the comparison of independent means, assuming a threshold value (δ) of 15 ms (see main text). New examples are introduced for the paired samples, and the correlation coefficient. As is usual for parametric tests, all tests discussed rely on the assumptions of normality and homogeneity of variances.

1. Comparison of two independent means – two-sample t-test

1.1. Equivalence test

H_0: $|\mu_1 - \mu_2| \geq \delta$
H_1: $|\mu_1 - \mu_2| < \delta$

$$t = \frac{\left|(\overline{X_1} - \overline{X_2})\right| - \delta}{\sqrt{S_P^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \qquad df = N_1 + N_2 - 2$$

where μ_1 is defined as the population mean of the first group and μ_2 as the population mean of the second group, $\overline{X_1}$ and $\overline{X_2}$ are the sample means of the two groups, N_1 and N_2 the sample sizes of the two groups, S_P^2 is the pooled variance of the two groups and δ is the threshold value for a non trivial difference.

The estimated pooled variance is obtained by the following formula:

$$S_P^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

where S_1^2 and S_2^2 are the sample variances of the two groups, N_1 and N_2 the sample sizes of the two groups and S_P^2 is the pooled variance.

For example 1 (Table I) this gives,

$$S_P^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2} = \frac{(4 \times 38.42^2) + (4 \times 47.53^2)}{5 + 5 - 2} = 1867.60$$

$$t = \frac{\left|(\overline{X_1} - \overline{X_2})\right| - \delta}{\sqrt{S_P^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} = \frac{|459.09 - 402.61| - 15}{\sqrt{1867.60\left(\frac{1}{5} + \frac{1}{5}\right)}} = \frac{41.48}{27.33} = 1.52$$

To calculate the *p*-value associated with the *t*-value obtained, we must determine the probability of observing a lower *t*-value, which is $p(t(8) < 1.52) = .92$.[7] This gives us the probability for rejecting the null-hypothesis that the effect size at the population level is equal to or larger than 15.

1.2. 95% confidence interval

$$IC_{0,95} = (\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2}\sqrt{\frac{S_P^2}{N_1} + \frac{S_P^2}{N_2}}$$

where $\overline{X_1}$ and $\overline{X_2}$ are the sample means of the two groups, N_1 and N_2 the sample sizes of the two groups, S_P^2 is the pooled variance of the two groups, and $t_{\alpha/2}$ is the critical two-tailed *t*-value for $p < \alpha$ and $N_1 + N_2 - 2$ degrees of freedom (e.g., for a *t*-test with df = 8, the critical $t_{.05}$ value is 2.306, as any statistical handbook will tell you).

For example 1,

$$IC_{0,95} = (\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2}\sqrt{\frac{S_P^2}{N_1} + \frac{S_P^2}{N_2}}$$

$$= (459.09 - 402.61) \pm 2.306\sqrt{\frac{1867.60}{5} + \frac{1867.60}{5}}$$

$$= 56.48 \pm 63.03$$

The confidence interval therefore is –6.55 to 119.51 ms.

2. Testing the mean difference – paired-sample t-test

To illustrate the calculations we use an example in which 5 participants rated their mood, going from 1 (very bad) to 5 (very good), in the morning and the evening

7. To obtain the *p*-value, you can use a calculator on the internet, or a built-in function in your computer. For instance, Excel contains a function to calculate *p*-value of *t*-tests. By using a one-tailed distribution, and subtracting the *p*-value from 1, you get what you need. Just try it. Open an Excel sheet and write in a cell **=1-TDIST(1.52,8,1)** or **=1-LOI.STUDENT(1,52;8;1)** if you use the French version of Excel. Do you get the expected rounded-off value of .92?

	Morning	Evening	D
p1	3	1	2
p2	4	4	0
p3	4	1	3
p4	5	4	1
p5	4	5	-1

2.1. Equivalence test

H_0: $|\mu_D| \geq \delta$
H_1: $|\mu_D| < \delta$

$$t = \frac{|\overline{D}| - \delta}{\frac{S_D}{\sqrt{N}}}, \qquad df = N - 1$$

where μ_D is defined as the mean population difference, $\overline{D}$ is the mean sample difference, S_D is the standard deviation of the difference scores, N is the number of difference scores and δ is the minimal threshold value for a non trivial difference.

For the example: $\overline{D} = 1$, $S_D = 1.58$, $\delta = .5$, $t = \frac{1 - .5}{\frac{1.58}{\sqrt{5}}} = .71$, df = 4, $p = .74$

2.2. 95% confidence interval

Confidence interval for the mean difference at the population level:

$$IC_{0,95} = \overline{D} \pm t_{\alpha/2}\frac{S_D}{\sqrt{N}}$$

where $\overline{D}$ is the mean sample difference, S_D is the standard error for the difference scores, N is the number of difference scores and $t_{\alpha/2}$ is the critical value of the t distribution for a two-tailed test with $p < \alpha$ and N – 1 degrees of freedom (e.g., the critical value you would use for a two-sided t-test with $\alpha = .05$).

For the example $1.0 \pm 2.776\ (1.58/\sqrt{5}) = 1.0 \pm 1.96$ (i.e., the interval goes from –.96 to + 2.96).

3. Testing for the slope of a regression – simple linear regression

3.1. Equivalence test

H_0: $|\beta| \geq \delta$
H_1: $|\beta| < \delta$

$$t = \frac{|b| - \delta}{S_b}, \qquad df = N - 2$$

where β is defined as the regression slope at the population level, b is the sample regression coefficient, S_b is the standard error for the regression coefficient and δ is the minimal threshold value for a non trivial regression slope[8].

3.2. 95% confidence interval

Confidence interval for the regression slope at the population level:

$$IC_{0,95} = b \pm t_{\alpha/2} S_b$$

where b is the sample regression coefficient, S_b is the standard error for the regression coefficient and $t_{\alpha/2}$ is the critical value on the t distribution for α and N – 2 degrees of freedom.

4. Testing for a correlation – Pearson's correlation coefficient

4.1. Equivalence test

H_0: $|\rho| \geq \delta$
H_1: $|\rho| < \delta$

$$Z = \frac{\frac{\ln\left(\frac{1 + |r|}{1 - |r|}\right)}{2} - \frac{\ln\left(\frac{1 + \delta}{1 - \delta}\right)}{2}}{\sqrt{\frac{1}{N - 3}}}$$

where ρ is defined as the population correlation coefficient, r is the sample correlation coefficient, N is the sample size and δ is the minimal threshold value for a non trivial correlation.

8. For the simple linear regression, the minimal threshold value has to be defined in terms of regression coefficients, i.e., what is the definition of a trivial effect in terms of the changes in the dependent variable for a one unit change on the independent variable.

Note that $\ln\left(\frac{1+|r|}{1-|r|}\right)$ transformations are used to correct for the well known bias in the standard error of the correlation test statistic, which then refers to the standard normal distribution.

Suppose a sample with N = 50, r = .1, and δ = .2. Then

$$Z = \frac{\frac{\ln\left(\frac{1+|r|}{1-|r|}\right)}{2} - \frac{\ln\left(\frac{1+\delta}{1-\delta}\right)}{2}}{\sqrt{\frac{1}{N-3}}} = Z = \frac{\frac{\ln\left(\frac{1+|.1|}{1-|.1|}\right)}{2} - \frac{\ln\left(\frac{1+.2}{1-.2}\right)}{2}}{\sqrt{\frac{1}{50-3}}} = -0.70$$

$p(Z < -0.70) = .24$.

4.2. 95% confidence interval

Confidence interval for the correlation coefficient at the population level:

$$IC_{0,95}\, for \frac{\ln\left(\frac{1+\rho}{1-\rho}\right)}{2} = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2} \pm Z_{\alpha/2}\frac{1}{\sqrt{N-3}}$$

where ρ is defined as the population correlation coefficient, r is the sample correlation coefficient, N is the sample size and $Z_{\alpha/2}$ is the critical value on the standard normal distribution for the defined probability of type 1 error (α) (i.e., 1.96 for α = .05).
These lower and upper confidence limits, L_1 and L_2, may then be transformed to ρ values with the following equation:

$$\rho = \frac{e^{2L}-1}{e^{2L}+1}$$

For example, suppose a sample with N = 50 and r = .10,

$$IC_{0,95}\, for \frac{\ln\left(\frac{1+\rho}{1-\rho}\right)}{2} = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2} \pm Z_{\alpha/2}\frac{1}{\sqrt{N-3}}$$

$$= \frac{\ln\left(\frac{1+0.10}{1-0.10}\right)}{2} \pm 1.96\frac{1}{\sqrt{50-3}}$$

$$= 0.10 \pm 0.286$$

$$L_1 = 0.10 - 0.286 = -0.186 \qquad L_2 = 0.10 + 0.286 = 0.386$$

$$\rho_1 = \frac{e^{2\times(-0.186)} - 1}{e^{2\times(-0.186)} + 1} = -0.18 \qquad \rho_2 = \frac{e^{2\times 0.386} - 1}{e^{2\times 0.386} + 1} = 0.37$$

The confidence interval for ρ is therefore –0.18 to 0.37.

Received April 2, 2011
Revision received April 9, 2011
Accepted April 12, 2011