Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 30

5-1-2010

An Equivalence Test Based on *n* and *p*

Markus Neuhäeuser
Koblenz University of Applied Sciences, Remagen, Germany, neuhaeuser@rheinahrcampus.de

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 304-307

BRIEF REPORTS
An Equivalence Test Based on *n* and *p*

Markus Neuhäeuser
Koblenz University of Applied Sciences,
Remagen, Germany

An equivalence test is proposed which is based on the *P*-value of a test for a difference and the sample size. This test may be especially appropriate for an exploratory re-analysis if only a non-significant test for a difference was reported. Thus, neither a confidence interval is available, nor is there access to the raw data. The test is illustrated using two examples; for both applications the smallest equivalence range for which equivalence could be demonstrated is calculated.

Key words: Equivalence; *P*-value; re-analysis; reverse test.

Introduction

Two or more groups are often compared in applied research, thus begging the question: What should be done in the case of a non-significant difference between the groups? Concluding that the null hypothesis of no difference is true without any further support is not correct. Here, it is shown that an equivalence test can be performed without access to raw data if the sample size and the *P*-value of a test for a difference are known, and if the test statistic is at least approximately normally distributed. This allows any reader to perform a re-analysis and it is possible to determine the smallest difference for which equivalence can be established.

A procedure sometimes performed in case of a non-significant difference is a retrospective power analysis, but such a retrospective power analysis has logical flaws and shortcomings (Hoenig & Heisey, 2001; Nakagawa & Foster, 2004). When the aim is to demonstrate the absence of a relevant difference it is necessary to reverse the traditional hypotheses in an equivalence test (McBride, 1999; Hoenig & Heisey, 2001). The null hypothesis will then state that there is a relevant difference, whereas there is essentially no difference – that is, a negligible difference only – under the alternative. Defining the effect size Cohen's *d*, calculated as $d = \frac{\mu_1 - \mu_2}{\sigma}$ (Cohen, 1988) where μ_i denotes the population mean of group *i* and σ the population standard deviation, results in

$$H_{0,\text{ equiv.}}: d \le -\theta \text{ or } d \ge \theta$$

vs.

$$H_{1,\text{ equiv.}}: -\theta < d < \theta \text{ (with } \theta > 0).$$

When the appropriate confidence interval for *d* is completely included within the equivalence range –θ to θ, the equivalence test's null hypothesis $H_{0,\text{ equiv.}}$ can be rejected (Steinijans, et al., 2000). Hence, the alternative $H_{1,\text{ equiv.}}$ cannot be $d = \theta$ only, the entire confidence interval has to be consistent with $H_{1,\text{ equiv.}}$.

Parkhurst (2001) suggested performing such an equivalence test whenever a classical test with a no-effect hypothesis has failed to yield a significant difference and he introduced the term reverse test for an equivalence test applied in this context. Parkhurst's suggestion has not become common practice. However, reporting a confidence interval would allow a

Markus Neuhäeuser is Professor of Statistics at the Koblenz University of Applied Sciences, RheinAhrCampus Remagen. His main research interests are nonparametric methods and their application in life sciences. Email: neuhaeuser@rheinahrcampus.de.

reader to check whether the interval lies within an assumed equivalence range and therefore to judge the biological importance of a result. Unfortunately, reporting confidence intervals is also not commonplace, although it is often recommended (see Nakagawa & Foster, 2004, and references therein). By contrast, two measures are almost always given when a null hypothesis of no difference is tested: the *P*-value *p* and the sample size *n*. It is the aim of this article to demonstrate how an equivalence test can be carried out based on *n* and *p* only.

The Proposed Equivalence Test

It is assumed that the test statistic is at least approximately normally distributed, which is true for a wide variety of commonly applied tests. Under the null hypothesis of no difference, the one-sided *P*-value has a uniform distribution over the interval [0, 1] regardless of the sample size *n*. Under the alternative hypothesis, that is, under the assumption that there is a difference, the probability for a small *P*-value increases. In this case, the *P*-value's distribution depends on *n* and *d* (Hung, et al., 1997).

First, consider a one-sample test with H_0: $\mu = 0$ vs. the one-sided alternative H_1: $\mu > 0$. If the effect size is defined as $d = \frac{\mu}{\sigma}$, then the distribution function of the *P*-value *p* is

$$G_d(p) = 1 - \Phi(Z_p - \sqrt{n}d), \qquad (1)$$

where Φ denotes the distribution function of the standard normal distribution and Z_p the $(1-p)^{\text{th}}$ percentile of that distribution, i.e. $\Phi(Z_p) = 1 - p$ (Hung, et al., 1997).

The *P*-value, *p*, of the test for a difference, can be used as the test statistic for the equivalence test. The critical region of the resultant equivalence test is [$G_\theta^{-1}(1-\alpha)$, 1], that is, whenever *p* lies within this interval equivalence can be concluded. The equivalence test's *P*-value is $p_{\text{equiv.}} = 1 - G_\theta(p)$.

When two samples with m_1 and m_2 observations, respectively, are compared H_0: $\mu_1 = \mu_2$ may be tested vs. the one-sided alternative H_1: $\mu_1 > \mu_2$. With the effect size $d = \frac{\mu_1 - \mu_2}{\sigma}$ and $n = \frac{m_1 m_2}{m_1 + m_2}$ the above-mentioned formulas for the one-sample scenario can be used (Hung, et al., 1997).

The formulas discussed apply to one-tailed tests. In the case that a two-tailed *P*-value is reported, a one-tailed *P*-value of the test for a difference can be calculated because the original test statistic is assumed to be at least approximately normally distributed (George & Mudholkar, 1990).

Applications

Scantlebury, et al. (2006) investigated the energy expenditure of the Damaraland mole-rat (*Cryptomys damarensis*). No significant change in body mass during the experimental period was found for any category of animal and condition. Consider frequent workers during dry conditions; in that case, $n = 21$ and Student's one-sample *t* test gives a one-tailed *P*-value $p = 0.18$.

When assuming that a moderate effect, $d = 0.5$, corresponds to a negligible change in body mass, the equivalence range is any effect size between −0.5 and 0.5. The critical region of the resulting equivalence test with $\alpha = 0.05$ is [0.259, 1], hence equivalence cannot be concluded in this case because 0.18 < 0.259. The equivalence test's *P*-value is $p_{\text{equiv.}} = 0.084$. The equivalence test with $\alpha = 0.05$ could demonstrate equivalence if an effect size with an absolute value of 0.559 or smaller would be regarded as a negligible difference. Thus, 0.559 is the smallest value of θ for which equivalence can be demonstrated.

Richdale (1957) observed yellow-eyed penguins (*Megadyptes antipodes*) from different colonies on New Zealand's South Island. He compared the number of days the birds were ashore as chicks between $m_1 = 27$ that were subsequently seen as juveniles or later, and $m_2 = 58$ chicks that were not seen again. Student's *t* test gives a one-tailed *P*-value of 0.300. Again, equivalence cannot be concluded if the range is any absolute value of the effect size smaller than a moderate effect of $d = 0.5$. The critical region of the resultant equivalence test with $\alpha = 0.05$ is [0.308, 1]. Here, 0.505 is the smallest value of θ for which equivalence can be demonstrated.

Is $\theta = 0.505$ a negligible effect in this example? Richdale (1957) reported means and standard deviations: 106.4 days (± 5.1) for the chicks not seen again and 105.8 days (± 4.4) for the other group, the estimated common standard deviation is 4.89. Hence, a mean difference of approximately 4 days would be a large effect of $d = 0.8$. A mean difference of approximately 2.5 days would give an effect of $d = 0.505$ for which equivalence can be demonstrated. Compared with the observed range of 97 to 118 days (Richdale, 1957) this difference appears to be negligible.

Conclusion

For any equivalence test the equivalence range has to be specified. Several proposals describe how to choose an equivalence range (see Ng, 2001 and references therein). Here, equivalence ranges based on the effect size *d* are used. According to Cohen (1988) $d = 0.2$ is a small effect, $d = 0.5$ a medium effect, and $d = 0.8$ a large effect. These values may be used although they depend on the variance, in particular because the equivalence test is used here with an exploratory intention. Different researchers may favour different equivalence ranges; in this case, Parkhurst (2001) recommended calculating the minimum value for which equivalence can be concluded. A SAS program to compute this value, given in the applications described herein, is available by request.

A large difference between μ_1 and μ_2 is possible even when the test for a difference gives a large one-tailed *P*-value. This is the case when the observed difference is in the opposite direction than specified by the one-sided alternative hypothesis; in this situation it is not useful to decide for equivalence. Therefore a conservative approach is warranted: the smaller one of the two possible one-tailed *P*-values for the equivalence test should be used. Note that this was done in the examples analysed, because the *P*-values of the test for difference were both ≤ 0.5.

When the equivalence test is performed as a reverse test after a non-significant test for difference, a multiple test problem occurs. It may be argued that the error rates of the entire procedure are not under control. However, the procedure is proposed here as a more exploratory means to allow a reader to gain additional information. When the aim of a study is to demonstrate equivalence of two treatments in a confirmatory manner an equivalence test must be performed as the first and main analysis. In this context it should be mentioned that Parkhurst (2001) recommended the reverse test particularly for basic science.

Finally, it should be noted that the idea of an original *P*-value-based equivalence test is not entirely new. Donahue (1999) mentioned that the temptation may exist to use the *P*-value in order to test for equivalence; however, he did not consider this idea any further because other equivalence tests exist. The situation considered herein is that, for a re-analysis, there is no access to raw data and no reporting of confidence intervals, hence, the equivalence test based on *p* and *n* may be the only choice. However, sometimes the *P*-value is not specified. If, instead of the *P*-value, a lower limit, such as $p > 0.45$ (e.g. Brown, et al., 2005), is specified the boundary can be used rather than the unknown *p* for the then conservative equivalence test.

References

Brown, W. M., et al. (2005). Dance reveals symmetry especially in young men. *Nature*, *438*, 1148-1150.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (*2nd Ed.*). Hillsdale: Lawrence Erlbaum Associates.

Donahue, R. M. J. (1999). A note on information seldom reported via the P value. *American Statistician*, *53*, 303-306.

George, E. O., & Mudholkar, D. S. (1990). P-values for two-sided tests. *Biometrical Journal*, *32*, 747-751.

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

Hung, H. M. J., O'Neill, R. T., Bauer, P., & Köhne, K. (1997). The behavior of the P-value when the alternative hypothesis is true. *Biometrics*, *53*, 11-22.

McBride, G. B. (1999). Equivalence tests can enhance environmental science and management. *Australian & New Zealand Journal of Statistics*, *41*, 19-29.

Nakagawa, S., & Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, *7*, 103-108.

Ng, T. H. (2001). Choice of delta in equivalence testing. *Drug Information Journal*, *35*, 1517-1527.

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

Scantlebury, M., et al. (2006). Energetics reveals physiologically distinct castes in a eusocial mammal. *Nature*, *440*, 795-797.

Steinijans, V. W., Neuhäuser, M., & Bretz, F. (2000). Equivalence concepts in clinical trials. *European Journal of Metabolism and Pharmacokinetics*, *25*, 38-40.

Richdale, L. E. (1957). *A population study of penguins*. Oxford: Clarendon Press.