Qual Life Res (2014) 23:5–7
DOI 10.1007/s11136-013-0437-2

COMMENTARY

Significance, truth and proof of *p* values: reminders about common misconceptions regarding null hypothesis significance testing

Mathilde G. E. Verdam · Frans J. Oort · Mirjam A. G. Sprangers

Accepted: 14 May 2013 / Published online: 23 May 2013

Our statistics should not become substitutes instead of aids to thought
After Bakan [1]

Null hypothesis significance testing has successfully reduced the complexity of scientific inference to a dichotomous decision (i.e., 'reject' versus 'not reject'). As a consequence, *p* values and their associated statistical significance play an important role in the social and medical sciences. But do we truly understand what statistical significance testing and *p* values entail? Judging by the vast literature on controversies regarding their application and interpretation, this seems questionable. It has even been argued that significance testing should be abandoned all together [2]. We seek to extend Fayer's [3] paper on statistically significant correlations and to clarify some of the controversies regarding statistical significance testing by explaining that (1) the *p* value is not the probability of the null hypothesis; (2) rejecting the null hypothesis does not prove that the alternative hypothesis is true; (3) not rejecting the null hypothesis does not prove that the alternative hypothesis is false; (4) statistical significance testing is not necessarily an objective evaluation of results; and (5) the *p* value does not give an indication of the size of the effect.

We note that this article does not raise new issues (see [4] for an extensive overview), but rather serves as a reminder of our responsibility as researchers to be knowledgeable about the methods we use in our scientific endeavors.

M. G. E. Verdam (✉) · F. J. Oort
Department of Child Development and Education, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, The Netherlands
e-mail: m.g.e.verdam@uva.nl; m.g.e.verdam@amc.uva.nl

M. G. E. Verdam · F. J. Oort · M. A. G. Sprangers
Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands

The *p* value is not the probability of the null hypothesis

Understanding what the null hypothesis significance test assesses can reduce the risk of misinterpreting the *p* value. Imagine we want to answer the question: do women experience a lower level of health-related quality of life (HRQL) than men? As we are not able to measure HRQL in the entire population, we need to select a sample of men and women. Statistical significance tests are used to infer whether an observed difference reflects a 'real' difference (i.e., a difference in the population) or one that is merely due to random sampling error (i.e., chance fluctuation). The null hypothesis is often chosen to be a 'nil hypothesis' (e.g., no relationship between variables, no difference between groups, or no effect of treatment). For the calculation of the *p* value, it is assumed that the null hypothesis is true, for example, that in reality, there is no difference in HRQL between men and women. Under this assumption, the statistical test will tell us the probability that we find a difference in our sample of the observed magnitude or larger. If this probability is very small (even smaller than the chosen level of significance), we can conclude 'given that HRQL of men and women is equal, the probability that we find the observed difference (or larger) is very small'. However, the calculated probability is often misinterpreted as 'given the observed difference, the probability is very small that in reality, HRQL of men and women is equal'. In symbols, the former corresponds to the following:

$P(D|H_0)$ The probability (P) that we find a difference of the observed magnitude (or larger) (D), given that the null hypothesis (H_0) is true.

While the latter corresponds to the following:

$P(H_0|D)$ The probability that the null hypothesis is true, given a difference of the observed magnitude (or larger).

That $P(D|H_0)$ and $P(H_0|D)$ are not the same is clarified by the following example: What is the probability of death (D), given a fatal heart attack (H_0); that is, what is $P(D|H_0)$? Obviously, it will be very high. Now, what is the probability that a person had a fatal heart attack (H_0), given that the person is dead (D); that is, what is $P(H_0|D)$? This probability is of course much lower. Therefore, one should not mistake $P(D|H_0)$ for $P(H_0|D)$ when interpreting the test of significance [2].

Rejecting the null hypothesis does not prove that the alternative hypothesis is true

The rejection of the null hypothesis should not be taken as proof that the alternative hypothesis is true. This formal fallacy is known as the *error of confirming the consequent*. For example, when we theorize that lung cancer has a different effect on HRQL in men than women, the fact that we find a statistically significant difference in HRQL does not necessarily prove that this can be attributed to a differential gender effect of the disease. Alternative explanations should be excluded (e.g., women report more symptoms than men in general), and more rigorous support is needed (e.g., substantive theorizing, replication of findings) to be able to draw conclusions about the probability that the alternative hypothesis is true. Therefore, the rejection of the null hypothesis does not give direct evidence that the alternative hypothesis is valid.

Not rejecting the null hypothesis does not prove that the alternative hypothesis is false

Similarly, not rejecting the null hypothesis does not prove that the alternative hypothesis is false. Instead, not rejecting the null hypothesis might be a consequence of insufficient statistical power (i.e., the probability of rejecting the null hypothesis when in fact, the alternative hypothesis is true; in symbols: $P(D|H_A)$). The calculation of statistical power requires the specification of the alternative hypothesis. This may be difficult as it requires the specification of what 'a difference' entails, that is, how much of a difference makes a difference? Such specifications are especially important in clinical practice, that is, when ensuring sufficient statistical power to detect minimal (clinically) important differences [5]. Thus, one should determine the probability of rejecting the null hypothesis when in reality, the effect of interest exists. Only when this probability is high (i.e., high statistical power) and we do not reject the null hypothesis, the tenability of the alternative hypothesis can be rejected.

Statistical significance testing is not necessarily an objective evaluation of results

The objectivity associated with statistical significance testing has led to a perceived objectivity of its application in evaluating results. However, the process of statistical significance testing requires several subjective decisions of the researcher: the determination of the level of significance (i.e., alpha), whether the test is one or two-tailed, and the number of observations. These decisions influence the chance of getting a statistical significant result, that is, with a higher alpha, one-tailed test, and large number of observations, the chance of finding statistically significant results increases. Consequently, the result of significance testing should not automatically be regarded as an objective way of interpreting results.

The *p* value does not give an indication of the size of the effect

Statistically significant does not mean the same as *clinically significant*, that is, important. When only significant p values are considered, important but statistically insignificant effects can be overlooked. Conversely, small effect sizes may turn out to be statistically significant with large sample sizes. Therefore, the use of effect sizes with confidence intervals has been persuasively recommended by many researchers [6]. In contrast to the p value, an effect size does give an indication of the magnitude of the effect, and the associated confidence interval provides information on the precision of the estimate. It can also provide information on the statistical significance of the estimate (i.e., a 95 % confidence interval reflects a significance level of 0.05). Furthermore, in clinical practice, the effect size estimate can be related to the assessment of minimal important differences. Norman and colleagues [7] suggested that anchor-based (e.g., patient-rated, clinician-rated) and distribution-based (e.g., effect size) estimates of minimal important differences in the area of HRQL consistently appear to be half a standard deviation, which corresponds to a medium effect size as indicated by Cohen [8]. Therefore, the effect size estimate, rather than the p value, may provide an answer to the question of how much of a difference was found and whether the difference matters [9].

Conclusion: what to do?

To cite Cohen [10]: 'Don't look for a magic alternative to null hypothesis significance testing […]. It doesn't exist.' (p. 1001). One should be aware of the limitations of statistical significance testing and use it only to support rather than replace (or make up for the absence of) theoretical and substantive foundations of the research. In addition, for substantive interpretation of results, one should turn to effect sizes and their confidence intervals rather than *p* values.

In conclusion, null hypothesis significance testing and *p* values should not lead us to think that inductive inference can be reduced to a simple, objective, dichotomous decision (i.e., 'reject' versus 'not reject'). Instead, we should remember that the significance of our results is determined by the informed judgement in planning and conducting our research, as well as in interpreting its findings, rather than by finding statistical significance.

References

1. Baken, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 1–29.
2. Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378–399.
3. Fayers, P. M. (2008). The scales were highly correlated: $p = 0.001$. *Quality of Life Research, 17*, 651–652.
4. Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301.
5. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*, 102–109.
6. Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594–604.
7. Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41*, 582–592.
8. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
9. Glaser, D. N. (1999). The controversy of significance testing: Misconceptions and alternatives. *American Journal of Critical Care, 8*, 291–296.
10. Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologists, 49*, 997–1003.