Journal of Clinical Trials

Rapid Communication Open Access

Two-Tailed P-Values Calculation in Permutation-Based Tests: A Warning Against "Asymptotic Bias" in Randomized Clinical Trials

Fernando Pires Hartwig*

Postgraduate Program in Epidemiology, Department of Social Medicine, Faculty of Medicine, Federal University of Pelotas, Brazil

Abstract

The rapid developments in computer science worldwide are enabling the routine use of computation-intensive methods for statistical applications. Among such methods, the permutation method is of particular interest because it allows robust calculation (regarding test assumptions) of the P-value based on an empirical null distribution. Moreover, this approach fits well with the general design and rationale of randomized clinical trials, indicating the potential of such method for studies with this design. In this commentary, a discussion to clarify the inadequacy of applying asymptotic reasoning for calculating two-sided P-values for permutation-based tests is considered, since such mistake can be observed in modern teaching literature and is of great concern for cases when the empirical null distribution is asymmetric and/or the P-value is close to the pre-defined α level. Moreover, the suitability of permutation-based tests to analyze results from randomized clinical trials indicates that such mistake has to be stressed to the medical research community in order to avoid both incorrect analyses and misinterpretations of such studies.

Keywords: Permutation; Randomized clinical trial; Two-sided P-value; Asymptotic reasoning

Introduction

Parallels between permutation and randomized clinical trials

Statistics has been applied in medical research for several years at an increasingly rate, and its importance to this field is hardly arguable. The core of statistical thinking has been developed in times when powerful computational tools were not available, and its admirable how brilliant the early statisticians were in developing theoretical (asymptotic) probability distributions to which approximations could be done, thus allowing the estimation of the probabilities that the observed results occurred due to mere sampling variation (i.e., under the null hypothesis). Such measure, usually called the P-value, although sometimes over-rated as a threshold of "statistical significance" (with its meaning being more interpretable when associated with respective confidence intervals), has a central role in hypothesis testing [1].

The rapid development of computer science and access of computational resources worldwide provided additional tools for statisticians to better calculate central concepts such as the P-value. Among such tools there are the so-called resampling methods, which allow the calculation of empirical distributions (through computation-intensive methods) to obtain, for example, confidence intervals for a given statistic (based on an empirical sampling distribution obtained by bootstrapping) or calculate P-values (based on an empirical null distribution) [2]. Perhaps the main advantage of this approach is the calculation of the "true null distribution" for the data at hand, which results in better estimates regarding a given exposure-outcome association. Focusing on the calculation of P-values, the rationale underlying the randomized clinical trial design is particularly illustrative: under the null, the values of a quantitative variable (i.e., outcome) that indicates the efficacy of the treatment are independent on whether or not that given experimental unit (in this case, the patient) received the treatment or the placebo. Then, the values of such quantitative variable can be shuffled, and the relevant statistic (any statistic, even ones not previously described in the literature - provided its adequacy) calculated. By repeating this process several times, an empirical distribution of the statistic is obtained and can be used to calculate the P-value of the observed statistic, calculated using the original (i.e., unshuffled) data.

The possibility of calculating null distributions that better fit the data is a strong motivation for the use of permutation (normally based on an approximate permutation process, such as Monte Carlo) tests. Moreover, since the permutation rationale fits well with the general design of randomized clinical trials, it is important to consider this statistical method as an option to be applied for robust (regarding departures of the data from assumptions of a given statistical test) inference in such important studies, where it is critical to prevent (given their relevance for causal inference) errors as much as possible.

Two-sided P-values for permutation-based tests: avoiding "asymptotic bias"

It is interesting to note that several concepts from asymptotic statistics are applicable to resampling-based statistics. In fact, some fundamental concepts, such as confidence intervals calculation for a given statistic, were originated from hypothesizing a theoretical re-sampling process from the underlying population [3]. However, such applicability is not necessarily true for the calculation of two-sided (or two-tailed) P-values, which is normally the case for the majority of statistical analysis (excluding hypothesis testing based on statistics such as F or chi-squared). This probability is obtained by calculating the area under the probability distribution curve referent to values equally or more "extreme" (that is, more distant from the null) than the calculated statistic for the observed data at both ends or tails of the distribution (since the two-sided P-value is open to more than one – normally two – alternative hypotheses). For asymptotic distributions that are symmetrical (e.g., standard normal or T distributions), this calculation can be simplified by calculating a one-sided P-value and multiplying

***Corresponding author:** Fernando Pires Hartwig, Postgraduate Program in Epidemiology, Department of Social Medicine, Faculty of Medicine, Federal University of Pelotas, Brazil, Tel: (5553)81347172; E-mail: fernandophartwig@gmail.com

Received July 19, 2013; **Accepted** September 27, 2013; **Published** September 30, 2013

Citation: Hartwig FP (2013) Two-Tailed P-Values Calculation in Permutation-Based Tests: A Warning Against "Asymptotic Bias" in Randomized Clinical Trials. J Clin Trials 3: 145. doi:10.4172/2167-0870.1000145

this value by two. This works due to the symmetry of asymptotic distributions, but this might very well not be the case for empirical distributions calculated by a permutation-based method, especially when the dependent variable, in the case of a continuous outcome, does not approximate the normal distribution (and, ironically, such cases are when using permutation is more justifiable, since normality is one of the main assumptions among asymptotic statistical tests) [4].

Although the "multiplying a one-sided P-value by two" approach (typical of asymptotic statistics) not being necessarily applicable for hypothesis testing under the permutation framework, recommendations to use such approach in this context can be seen in modern and qualified teaching materials. Two recent books [5,6] are examples of that (it is important to note that the quality of the entire books is not being assessed). In both of them, the authors correctly explain the logic of one-sided tests in the context of permutation, but seem to be somehow "influenced" or "biased" towards the asymptotic reasoning regarding the calculation of two-sided P-values. This (in a particular perspective) suggests that the problem is not lack of statistical knowledge, but rather a bias caused by transposing the reasoning used in asymptotic hypothesis testing for calculating two-sided P-values into the permutation context, where such reasoning might not apply. The mentioned "bias of asymptotic reasoning" might be such that the authors discuss even the calculation of a two-tailed P-value (using the "asymptotic approach") based on a skewed empirical null distribution, which is clearly incorrect since the symmetry of the null distribution is what makes the use of the "multiplying a one-sided P-value by two" approach correct. What should be done in place is to calculate the P-value for each side of the distribution (i.e., calculate the two one-sided P-values individually) and sum them. An alternative (and perhaps simpler) approach would be to calculate a one-sided P-value (relative to the null statistics equal to or greater than the observed statistic) using a null distribution of absolute values of the statistic of interest and the absolute value of the observed statistic, since this would put both positive and negative values of a given statistic in the same side (>0) of the distribution.

Discussion and Remarks

It is difficult to verify whether or not the pointed mistake in teaching literature has influenced the analyses and/or interpretation of published studies, since it is usually not described how the "two-tailing" has been done. As an example, an interventional study published more than 10 years ago in the prestigious The New England Journal of Medicine is considered; in the statistical analyses, the use of a "two-sided permutation t-test" is mentioned [7]. This manuscript is a particularly illustrative example since it indicates that permutation-based methods have been employed in high-quality studies for some time (which can be verified in the archives of this and other medical journals). Although the study itself (including the adequacy of the statistical analyses) is not being assessed here, the description of the permutation test as it is in the manuscript is not sufficient to clarify which P-value calculation approach (asymptotic-based or actually two-sided) was used. This consideration is even more important for studies that reported P-values in the borderline of the a priori defined significance threshold (i.e., the α level, normally defined as 5%), where the use of one approach or the other may interfere with the interpretations (that is, the P-value is < α if one approach is used, but ≥ α for the other) and/or conclusions of the results by the authors and by the scientific community (since a great deal of attention – arguably more than it should in some cases – is normally putted on the dichotomy significant/not significant).

The issue regarding the statistical concept underlying the calculation of two-tailed P-values in permutation-based tests is relevant to medical research (and to the scientific community in general, as well as other users of such information) for three main reasons: the first is the relevance that is generally given to the P-value, which is, indeed, an important concept for statistical inference; the second is the increasing popularity that permutation-based approaches have been receiving due to the almost universal access to powerful computational tools and to the appropriateness of such approach to the general design of randomized clinical trials,, which is commonly regarded as the gold-standard for causal inference in medical research; the third is the origin of the pointed equivoque: it is (or seems to be) the use of the wrong rationale rather than lack of knowledge of the topic, indicating that the misuse of asymptotic thinking for calculating two-sided P-values in the context of permutation can be easily avoided. It has be to stressed to the medical researcher that the bias (possibly) related to thinking under the asymptotic statistical framework has to be avoided when working with (possible) asymmetric empirical null distributions originated from the permutation process by thinking what a two-sided P-value actually means and, then, proceed on calculating it.

References

1. Kirkwood BR, Sterne JAC (2003) Essential Medical Statistics. Blackwell Publishing 71-79.
2. Kabacoff RI (2011) R in Action. Manning Publications Co 291-310.
3. Kirkwood BR, Sterne JAC (2003) Essential Medical Statistics. Blackwell Publishing 50-57.
4. Kirkwood BR, Sterne JAC (2003) Essential Medical Statistics. Blackwell Publishing 42-49.
5. Chiara LM, Tim HC (2011) Mathematical Statistics with Resampling and R. John Wiley & Sons 35-75.
6. Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R (2008) The Practice of Business Statistics: Using Data for Decisions. WH Freeman 18.4-18.65.
7. Trollfors B, Taranger J, Lagergård T, Lind L, Sundh V, et al. (1995) A placebo-controlled trial of a pertussis-toxoid vaccine. N Engl J Med 333: 1045-1050.