



Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: examples from *Animal Behaviour*

ROGER MUNDRY & JULIA FISCHER

Institut für Verhaltensbiologie, Freie Universität Berlin

(Received 5 December 1997; initial acceptance 20 January 1998;
final acceptance 9 February 1998; MS. number SC-1124)

In recent years, the use of software for the calculation of statistical tests has become widespread. For many nonparametric tests, a number of statistical programs calculate significance levels based on algorithms appropriate for large samples only ('asymptotic testing'). In behavioural sciences, however, small samples are common. In nonparametric tests, this requires the use of the 'exact' variant of the respective statistical test. Using the asymptotic variant of a nonparametric test with small sample sizes usually yields an incorrect P value, and consequently, this may lead to a false acceptance or rejection of the null hypothesis. With the frequent application of statistical packages, the inappropriate use of the asymptotic variant has unfortunately become quite common. We examined the results of nonparametric tests with small sample sizes published in a recent issue of *Animal Behaviour* and found that in more than half of the articles concerned, the asymptotic variant had apparently been inappropriately used and incorrect P values had been presented. Before describing this analysis of papers published in *Animal Behaviour*, we provide a short overview of the differences between asymptotic and exact testing and discuss the consequences of the inappropriate use of asymptotic tests.

Exact and Asymptotic Test Procedures

Almost any statistical test is based on the same idea: in a first step, a specific figure, the so-called test statistic is calculated, for example, χ^2 , U , or W in the Mann-Whitney U test, or T^+ in the Wilcoxon signed-ranks test for matched pairs. In a second step, the calculated test statistic is compared with a critical value. If the value of the test statistic is smaller or larger than the critical value (depending on the test applied), the null hypothesis can be rejected. The critical values are usually determined by

Correspondence: R. Mundry, Institut für Verhaltensbiologie, Freie Universität Berlin, Haderslebener Str. 9, 12163 Berlin, Germany (email: rmundry@biologie.fu-berlin.de).

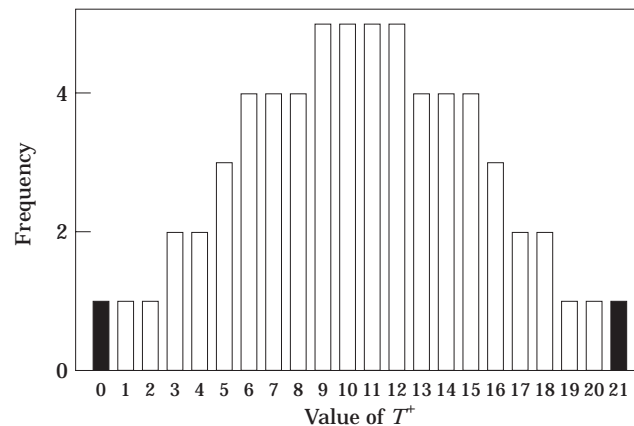


Figure 1. Distribution of T^+ values for the Wilcoxon signed-ranks test with $N=6$, under the assumption that the null hypothesis is true. The distribution was obtained by calculating the value of T^+ for each of the 64 possible combinations of '+' and '-' ranks. The two-tailed probability of obtaining a T^+ as extreme as 0 or 21 (black columns) is $2/64=0.03125$. With these values of T^+ , the null hypothesis can be rejected with $P<0.05$. The two-tailed probability of obtaining a T^+ at least as extreme as 1 or 20 is $4/64=0.0625$. In these cases, the null hypothesis cannot be rejected with $P<0.05$.

cutting off the most extreme 5% of the theoretical frequency distribution of the test statistic (Fig. 1; Siegel & Castellan 1988).

There are two ways of estimating the theoretical frequency distribution of a test statistic for many nonparametric tests: one for small and one for large samples. When the sample size is small, the exact probability of obtaining the calculated value of the test statistic or any less likely value has to be determined. The sum of these probabilities is the exact P value of the test statistic. This procedure to determine the exact probability of a specific value of a test statistic (and the less likely ones) is more or less complicated, but has fortunately been worked out for the common nonparametric tests. The results are

presented in tables included in most statistical books, and calculated values of the test statistic can be compared with the tabulated critical values. Since this procedure is based on the calculation of the exact probability of a given test statistic, it is called 'exact' testing (Siegel & Castellan 1988; Bortz et al. 1990).

With increasing sample size, the frequency distribution of a test statistic often asymptotically approaches a normal or a chi-square distribution. Then, the P value can be obtained by transforming the test statistic and looking up the transformed value in a table of Z or chi-square values. As in the exact test, the test statistic is first calculated. If the distribution of the test statistic approaches a normal distribution, the theoretical mean (also referred to as expected value of the test statistic) and the standard deviation of this normal distribution are calculated from the parameters of the data set to be analysed (e.g. the sample size N and the number of samples k). Then, a Z value is calculated by subtracting the theoretical mean from the empirical test statistic and dividing the result by the standard deviation of the theoretical distribution (Sokal & Rohlf 1987; Siegel & Castellan 1988; Bortz et al. 1990). The resulting value then becomes the test statistic and can be directly compared with the critical values of the standard normal distribution which are given in most statistical books. In some tests the difference between the empirical test statistic and its expected value has to be corrected for continuity before dividing by the standard deviation (Siegel & Castellan 1988). If the frequency distribution of the test statistic approaches a chi-square distribution, the test statistic can simply be compared with the critical values given in a chi-square table. These procedures are called 'asymptotic' testing (Bortz et al. 1990).

This distinction between an exact procedure for small samples and an asymptotic procedure for large samples is made for many nonparametric tests, for example the Wilcoxon test, the binomial test, the sign test, the Mann-Whitney U test (also referred to as Wilcoxon-Mann-Whitney test), and the test of significance of the Spearman rank-order correlation coefficient. With increasing sample size, the distributions of the corresponding test statistics approach a normal distribution so that ' Z ' can be used as the test statistic. In other tests, such as the Kruskal-Wallis one-way analysis of variance by ranks or the Friedman two-way analysis of variance by ranks, the distribution of test statistics approaches a chi-square distribution (Siegel & Castellan 1988; Bortz et al. 1990).

In the following, we use the example of the Wilcoxon signed-ranks test for matched pairs as described by Siegel & Castellan (1988) to illustrate the difference between the exact and the asymptotic procedure: either way, the test statistic T^+ is first calculated. If the sample size is smaller than 16, the critical value for T^+ should be taken from the table 'Critical values of T^+ for the Wilcoxon signed-ranks test' in the appendix of the book. If the sample size is at least 16, the asymptotic test can be performed. First, the mean and standard deviation of T^+ are calculated. These describe the theoretical frequency distribution of T^+ values, expected under the assumption that the null hypothesis is true, in sufficient approximation. Then, a Z value is calculated and compared with the critical values

in, for instance, the table 'Selected significance levels for the normal distribution' in the appendix of the book. Note that other authors may disagree on the sample size required to use the asymptotic variant: Sokal & Rohlf (1987), for example, suggest the use of the asymptotic variant only when $N > 50$. In ambiguous cases we recommend the use of the exact variant.

Consequences of the Inappropriate Use of the Asymptotic Test

Use of the asymptotic variant when the sample size is smaller than the threshold value can lead to a false decision, that is, incorrect rejection of the null hypothesis or false acceptance. In fact, for the Wilcoxon signed-ranks tests with sample sizes smaller than 16 and a significance level of 5%, use of Z as the test statistic can lead to a false rejection of the null hypothesis with nine sample sizes in a one-tailed test and with four sample sizes in a two-tailed test. For instance, with $N=6$ and $T^+=20$ the exact two-tailed Wilcoxon test yields an exact probability of 0.0625 (see Siegel & Castellan 1988), and thus, the null hypothesis cannot be rejected (Fig. 1). The Z transformation of T^+ results in $Z=1.99$, a value that exceeds the critical value of $Z=1.96$ and consequently leads to a rejection of the null hypothesis with $\alpha < 5\%$ (two-tailed). Therefore, in this case, the actual probability for a type I error (false rejection of the null hypothesis) is increased from 5 to 6.25%.

We systematically examined the difference between the exact and the asymptotic variant for the Wilcoxon signed-ranks test, and found that the exact variant was more conservative than the asymptotic one. In other words, the asymptotic test sometimes led to a rejection of the null hypothesis whereas the exact test led to its acceptance. We also checked the difference between the results of the asymptotic and the exact procedure for a few data sets for some other tests: in the Mann-Whitney U test, the use of the asymptotic variant often led to lower P values and thus to an increase in the probability of a type I error, as was the case for the Wilcoxon signed-ranks test. In contrast, the asymptotic test of significance for the Spearman rank-order correlation coefficient usually led to higher P values and thus increased the probability of a type II error (false acceptance of the null hypothesis). In the binomial and the sign test, respectively, asymptotic P values could be higher or lower than exact P values, but differences were always small and in none of the tested cases led to an incorrect acceptance or rejection of the null hypothesis. These differences between asymptotic and exact variants of tests refer to a significance level $\alpha < 5\%$ and the procedures as described by Siegel & Castellan (1988). Using different procedures for calculating asymptotic P values may have different effects on the probability of type I and type II errors.

Use of Statistical Packages

Apparently, many of the commonly used versions of statistical programs calculate P values using asymptotic

Table 1. Threshold values for sample sizes (N) and number of samples (k) above which asymptotic testing is permitted according to Siegel & Castellan (1988)

Wilcoxon signed-ranks test	$N > 15$
Mann-Whitney U test	$N_1 = 3$ or 4 and $N_2 > 12$; $N_1 > 4$ and $N_2 > 10$
Kruskal-Wallis H test	$k > 3$ and all $N > 5$
Friedman ANOVA	$k = 3$: $N > 13$; $k = 4$: $N > 8$; $k = 5$: $N > 5$; $k > 5$
Sign test	$N > 35$
Test for significance of Spearman's rank-order correlation coefficient	$N > 20$ to 25
Binomial test	$N > 25$, with $p, q = 0.5$

p, q = probability of the two categories.

procedures regardless of the sample size and therefore give incorrect P values when the sample size is small. Whether a program has used an asymptotic or exact procedure can easily be tested by generating a small data set, and comparing the P value given by the program with the P value obtained by looking up the test statistic calculated by hand in a printed table. As an example, a Wilcoxon signed-ranks test with five matched pairs in which all the values in the second condition are higher gives the test statistic $T^+ = 15$, which yields an exact two-tailed $P = 0.0626$ (see Siegel & Castellan 1988). The asymptotic procedure, however, gives a two-tailed $P = 0.0434$. We emphasize that programs may use different asymptotic procedures from those given in Siegel & Castellan (1988) and thus may yield different P values from those mentioned here.

Assessing the results of a test calculated with the aid of a statistical package is further complicated by the fact that some programs may provide the exact test statistic, but do not specify whether the P value is based on the exact or asymptotic procedure. Programs can also be inconsistent, using an exact procedure in one nonparametric test, but not in another. Therefore, such programs should be used with care when sample sizes are small. When the software program does not indicate whether an exact or asymptotic procedure was used, the program should be checked as described above. If the program provides the test statistic for the exact procedure (e.g. T^+), this obviates calculating the statistic by hand, but a check should be made that the program calculates the test statistic in the same way as for the tabulated values.

Examples from *Animal Behaviour*

The problem of the inappropriate use of asymptotic procedures with small samples is not a trivial one: we examined the application of exact and asymptotic tests to nonparametric data in one volume of *Animal Behaviour* (1997, vol. 53) for the tests mentioned above. We found 51 publications in which sample sizes required the use of exact tests according to Siegel & Castellan (1988) (Table 1). In 18 of these publications, an asymptotic Z or chi-square statistic was used inappropriately in at least one, and often all, of the tests applied on small samples that required exact testing. This concerned Wilcoxon signed-ranks, Mann-Whitney U and Friedman tests. In an additional 17 publications it was unclear in at least one case whether an exact procedure or an asymptotic

procedure was used. These mainly concerned Spearman rank-order correlation coefficients and binomial tests where the test statistic was either not given, or did not allow us to decide whether an asymptotic or exact procedure was applied. Finally, there were also a number of Wilcoxon and Mann-Whitney U tests in which the authors did not present a test statistic at all.

The appropriate exact test statistic (e.g. T^+ , U) was indicated for every test made that required an exact procedure in only 16 out of the 51 papers. However, in several of these publications, the P value given by the authors was presumably obtained from a Z transformation, because it did not correspond to that given by the exact procedure using tables (Siegel & Castellan 1988), but rather to the respective asymptotic procedure. Two examples will illustrate this. $P < 0.05$ was given for a Mann-Whitney U test performed on $N_1 = N_2 = 3$ with the test statistic $W = 15$, without indicating whether a one-tailed or a two-tailed test was used. However, the correct exact two-tailed probability is $P = 0.1$. We assume that a program was used that conducted a Z transformation (without correction for continuity) yielding $Z = -1.964$ and a two-tailed P of 0.0495. Another example was a Wilcoxon test with $T = 0$ (equivalent to $T^+ = 15$), $N = 5$ and $P = 0.04$. Again, it was not clear whether this was a one- or two-tailed P value. In fact, the exact P is 0.0313 (one-tailed) or 0.0625 (two-tailed). The indicated P value was probably obtained from a Z transformation yielding $Z = 2.02$ and a two-tailed $P = 0.0434$. Finally, our analysis of the use of exact and asymptotic test procedures was limited by the fact that in some further publications it was in at least one case impossible to identify either the test procedure used or the sample size.

In conclusion, inappropriate asymptotic tests and corresponding significance levels were used regularly when the sample size was below the threshold for doing so. As demonstrated above, this can lead to a false rejection or acceptance of the null hypothesis and accordingly to a misinterpretation of the results of a study. Since the output of statistical programs is often likely to be misunderstood, we strongly recommend using such programs with care and checking that small sample sizes are tested appropriately. Finally, we encourage anyone to examine statistical programs for the correct calculation of test statistics (i.e. exact and asymptotic P values, correction for ties, correction for continuity), and would be pleased to receive reports on this issue. Eventually, we aim to

compile a database that might be made available on the Internet.

We thank Henrike Hultsch, Kate Lessells, Marc Naguib, Lars Schrader, Hans Slabbekoorn and an anonymous referee for helpful comments on the manuscript. R.M. was supported by the Berlin-Brandenburgische Akademie der Wissenschaften.

References

- Bortz, J., Lienert, G. A. & Boehnke, K. 1990. *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer.
- Siegel, S. & Castellan, N. J. 1988. *Nonparametric Statistics for the Behavioural Sciences*. New York: McGraw-Hill.
- Sokal, R. R. & Rohlf, F. J. 1987. *Introduction to Biostatistics*. New York: W. H. Freeman.