J. Parasitol., 90(4), 2004, pp. 689–691

COMPARING PARASITE NUMBERS BETWEEN SAMPLES OF HOSTS

Markus Neuhäuser and Robert Poulin*

Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, Hufelandstrasse 55, D-45122 Essen, Germany. *e-mail: markus.neuhaeuser@medizin.uni-essen.de*

ABSTRACT: The comparison of parasite numbers or intensities between different samples of hosts is a common and important question in most parasitological studies. The main question is whether the values in one sample tend to be higher (or lower) than the values of the other sample. We argue that it is more appropriate to test a null hypothesis about the probability that an individual host from one sample has a higher value than individual hosts from a second sample rather than testing hypotheses about means or medians. We present a recently proposed statistical test especially designed to test hypotheses about that probability. This novel test is more appropriate than other statistical tests, such as Student's *t*-test, the Mann–Whitney *U*-test, or a bootstrap test based on Welch's *t*-statistic, regularly used by parasitologists.

According to Rózsa et al. (2000), intensity and abundance are among the most important measures used to quantify parasite loads in a host sample or population. Intensity is the number of conspecific parasites living in or on an infected host, whereas abundance is the number of conspecific parasites living in or on any host. Often, 2 (or more) different samples of hosts are compared, and the question is whether individual hosts in 1 sample are more parasitized than hosts in the other sample. This question appears in 1 form or another at the core of most parasitological studies, in which host samples represent either different subsets of the host population, different host populations, or different groups of laboratory hosts subjected to different experimental treatments.

Rózsa et al. (2000) presented a fictitious but typical example consisting of 2 samples of 10 infected hosts, each with the following intensities:

Sample A: 1, 1, 1, 1, 1, 1, 1, 1, 2, 50 and

Sample B: 1, 1, 2, 2, 2, 2, 3, 3, 4, 10.

These parasite distributions are aggregated and skewed because, as usual, a high proportion of parasites is concentrated on a few host individuals. Therefore, parametric tests such as Student's *t*-test should be avoided (Sawilowsky and Blair, 1992). Parasitologists often prefer nonparametric tests like the Mann–Whitney *U*-test or the Fisher–Pitman permutation test (also known as randomization test). However, these tests can give a significant result for a test at the 5% level with much more than 0.05 probability when the population means are identical but the population variances differ (Boik, 1987; Hayes, 2000; Kasuya, 2001). Different variances are common for parasite intensities, the data set of Rózsa et al. (2000) being an extreme example with standard deviations of 15.5 in sample A and 2.6 in sample B. Because of the heterogeneous variances, Student's *t*-test after a log transformation is not appropriate either (Zhou et al., 1997).

Hence, choosing an appropriate statistical test is not easy. Rózsa et al. (2000) recommended a bootstrap test based on Welch's *t*-statistic (Efron and Tibshirani, 1993, p. 222–224) for the comparison of mean intensities. However, this test compares mean values, and these are highly dependent on a few, extremely high, individual intensities. Means are not very useful descriptors for skewed distributions (see also Gould, 1996). Rózsa et al. (2000) quoted Margolis et al. (1982), who wrote that ''in some cases, median intensity or modal intensity will be appropriate substitutes for mean intensity.'' But irrespective of the measure for the central location of a sample, the main question is whether the values in one sample tend to be larger (or smaller) than the values of the other sample. Based on this question, the natural measure for a difference between 2 samples is the relative effect defined as

$$p = \Pr(X < Y) + (1/2)\Pr(X = Y),$$

where X is an observation from group 1 and Y is one from group 2 (Brunner and Munzel, 2000).

Let us consider the special case of normally distributed data with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. Then,

$$p = \Phi[(\mu_2 - \mu_1)/\sqrt{\sigma_1^2 + \sigma_2^2}]$$

where Φ is the cumulative distribution function of the standard normal distribution (Reiser and Guttman, 1986). Since $\Phi(t) = 1/2$ if and only if $t = 0$, $\mu_1 = \mu_2$ is equivalent to $p = 1/2$, but the standard deviations σ_1 and σ_2 can differ. Furthermore, p is smaller than 1/2 if $\mu_1 > \mu_2$ and larger if $\mu_1 < \mu_2$. In general, the observations in group 1 tend to be larger in comparison with those of group 2 if $p < 1/2$. In the case of $p > 1/2$, the observations in group 1 tend to be smaller than those of group 2. If $p = 1/2$, neither group generally has larger values than the other and $\Pr(X > Y) = \Pr(X < Y)$ holds (Delaney and Vargha, 2002). The parameter p is well established in the nonparametric analysis of clinical trials (e.g., Munzel and Hauschke, 2003, for references), but to the best of our knowledge, it has not yet been used in parasitology.

In this paper, we give a recently proposed statistical method based on the relative effect p, and we argue that the novel test is preferable not only to Student's *t*-test and the Mann–Whitney *U*-test but also to the bootstrap test recommended by Rózsa et al. (2000). In addition, we illustrate the method using the example data set mentioned above as well as hypothetical data sets.

THE BRUNNER AND MUNZEL (2000) TEST

Recently, Brunner and Munzel (2000) recommended a nonparametric test based on ranks for testing the null hypothesis H_0: $p = 1/2$ versus the 2-sided alternative $p \neq 1/2$ or versus a 1-sided alternative such as $p > 1/2$. The advantages of this test are that arbitrary distribution functions are acceptable (with the exception of a 1-point distribution, i.e., all data points have the

Received 7 October 2003; revised 7 December 2003; accepted 8 December 2003.

* Department of Zoology, University of Otago, P.O. Box 56, Dunedin, New Zealand.

TABLE I. Simulated power of the BM test for different negative binomial distributions (sample size per group: 10, significance level α = 0.05, 10,000 simulation runs).

Distribution	θ*†	MF = 1†‡	MF = 2
k§ = 1, mean 1.5	3	0.92	0.88
k = 1, mean 2.33	4	0.90	0.91
k = 1, mean 4	6	0.87	0.77
k = 2, mean 3	4	0.89	0.81
k = 2, mean 4.67	5	0.82	0.52
k = 2, mean 8	8	0.82	0.49

* θ, location shift.
† After simulating the data and transforming the distributions to have a median of 0, the values in 1 group were multiplied with MF and shifted by the amount θ. Consequently, in the case of MF = 1, there is a difference in location only. For MF = 2, there is an additional difference in variability.
‡ MF, multiplication factor.
§ Aggregation parameter (exponent) of the negative binomial distribution.

same value) and that a small sample approximation was given. These 2 points are especially important for the comparison of parasite numbers in samples of hosts. The test proposed by Brunner and Munzel (2000), hereafter the BM test, and the corresponding confidence interval are presented in the Appendix. The test can be carried out using a SAS program available at www.maths.otago.ac.nz/home/downloads/markus_neuhauser/BM_test.sas. Note that equivalence or reverse tests (Parkhurst, 2001) are also possible based on the relative effect *p* (Munzel and Hauschke, 2003).

The BM test compares well with alternative approaches suggested for testing hypotheses about *p* (Delaney and Vargha, 2002; Neuhäuser and Lam, 2004). Here, we show that the BM test is also powerful when the data come from negative binomial distributions (Table I). Note that, according to Crofton (1971), the distribution of parasite numbers can be adequately described by a negative binomial distribution. Moreover, the BM test is more appropriate than the bootstrap test recommended by Rózsa et al. (2000) because the focus is on *p* and not on the difference in means.

When applied to the above-mentioned example of Rózsa et al. (2000), the estimate for the relative effect is $\hat{p} = 0.78$. The 95% confidence interval for *p* is 0.54 to 1.0. Furthermore, we have $W_{BF} = 2.5077$ and the 2-sided *P*-value based on the *t*-distribution with df = 16.06 is 0.0233. Hence, the result is significant at the 0.05 level, and we can conclude that the values of the 2 samples are not equal. In contrast, the bootstrap test is far from being significant, with the *P*-value based on 100,000 bootstrap samples being 0.59.

DISCUSSION

Which sample of hosts contains individuals that are more parasitized? To answer this question, it does not always make sense to focus only on average or median differences. Instead, the whole distribution should be considered; consequently, one can focus on the relative effect *p* as a natural measure for the difference between the samples' distributions. We presented the recently proposed BM test that is especially designed to test hypotheses about *p*. The null hypothesis $p = 1/2$ is equivalent to the equality of means and medians only if the underlying distributions are symmetric. However, parasite distributions are usually skewed. Thus, the BM test seems to be very appropriate for these parasite distributions. Because mean comparisons ignore what is happening with individuals, it also has advantages in comparison with the bootstrap test that Rózsa et al. (2000) recommended. In particular, it shifts the emphasis away from comparisons between means or medians and places it on the probability that an individual host from 1 sample has a higher intensity value than individual hosts from a second sample. Although this is not necessarily a better way in general, it seems to be more meaningful than comparing central tendencies of parasite distributions.

Often, more than 2 groups are compared within a study. In the case of $k \geq 3$ groups, the relative effect of group *i* can be defined as

$$p_i = \frac{1}{k} \sum_{j=1}^{k} [\Pr(X_j < X_i) + 0.5\Pr(X_j = X_i)],$$

where X_j is an observation from group *j* ($1 \leq j \leq k$). The observations from group *i* tend to be larger than those from group *j* if $p_i > p_j$. In the case of $p_i < p_j$, the observations from group *i* tend to be smaller than those from group *j*. The observations tend to be equal if $p_i = p_j$. Further details about the analysis of more than 2 groups can be found in Brunner and Munzel (2002, p. 24–25 and 124–128).

Comparing infection levels among host samples is one of the most common statistical procedures used by parasitologists. Obtaining an accurate result is critical, especially in the context of studies assessing the effectiveness of anthelmintics or other parasite control methods. Yet, the results are often entirely dependent on the statistical method chosen by the investigators. The test we propose in this study is, statistically speaking, the most appropriate test among the simple tests currently available, and we recommend it instead of the now widely used parametric and nonparametric alternatives.

ACKNOWLEDGMENTS

R.P. is supported by a James Cook Research Fellowship from the Royal Society of New Zealand. We thank Edgar Brunner for comments on an earlier version of the manuscript.

LITERATURE CITED

BOIK, R. J. 1987. The Fisher-Pitman permutation test: A non-robust alternative to the normal theory *F* test when variances are heterogeneous. British Journal of Mathematical and Statistical Psychology **40:** 26–42.

BRUNNER, E., AND U. MUNZEL. 2000. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small sample approximation. Biometrical Journal **42:** 17–25.

———, AND ———. 2002. Nichtparametrische Datenanalyse. Springer, Berlin, Germany, 312 p.

CROFTON, H. D. 1971. A quantitative approach to parasitism. Parasitology **62:** 179–193.

DELANEY, H. D., AND A. VARGHA. 2002. Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. Psychological Methods **7:** 485–503.

EFRON, B., AND R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman & Hall, New York, 436 p.

GOULD, S. J. 1996. Full house: The spread of excellence from Plato to Darwin. Harmony Books, New York, 244 p.

HAYES, A. F. 2000. Randomization tests and the equality of variance assumption when comparing group means. Animal Behaviour **59:** 653–656.

KASUYA, E. 2001. Mann-Whitney *U* test when variances are unequal. Animal Behaviour **61:** 1247–1249.

MARGOLIS, L., G. W. ESCH, J. C. HOLMES, A. M. KURIS, AND G. A. SHAD. 1982. The use of ecological terms in parasitology (report of an ad hoc committee of the American Society of Parasitologists). Journal of Parasitology **68:** 131–133.

MUNZEL, U., AND D. HAUSCHKE. 2003. A nonparametric test for proving noninferiority in clinical trials with ordered categorical data. Pharmaceutical Statistics **2:** 31–37.

NEUHÄUSER, M., AND F. C. LAM. 2004. Nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *In* Conferences in research and practice in information technology, Vol. 29, Y.-P. P. Chen (ed.). Australian Computer Society, Adelaide, Australia, p. 139–143.

PARKHURST, D. F. 2001. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. Bioscience **51:** 1051–1057.

REISER, B., AND I. GUTTMAN. 1986. Statistical inference for $\Pr(Y < X)$: The normal case. Technometrics **28:** 253–257.

RÓZSA, L., J. REICZIGEL, AND G. MAJOROS. 2000. Quantifying parasites in samples of hosts. Journal of Parasitology **86:** 228–232.

SAWILOWSKY, S. S., AND R. C. BLAIR. 1992. A more realistic look at the robustness and type II error properties of the *t* test to departures from population normality. Psychological Bulletin **111:** 352–360.

ZHOU, X.-H., S. GAO, AND S. L. HUI. 1997. Methods for comparing the means of two independent log-normal samples. Biometrics **53:** 1129–1135.

APPENDIX

To calculate the test statistic W_{BF}, the ranks have to be determined first. Let R_{ij} denote the combined-samples rank of X_{ij}, where X_{ij} is the jth observation from the ith group ($i = 1, 2$; $j = 1, \ldots, n_i$). When there are ties, the usual way of dealing with these values is to assign average ranks, that is, we give tied observations the average of the ranks for which these observations are competing. Let $\overline{R}_1$ and $\overline{R}_2$ denote the mean of the ranks in groups 1 and 2, respectively. Then, W_{BF} is defined as

$$W_{BF} = \sqrt{\frac{n_1 n_2}{N}} \cdot \frac{\overline{R}_2 - \overline{R}_1}{\hat{\sigma}_{BF}},$$

where n_1 and n_2 are the sample sizes of the 2 groups ($N = n_1 + n_2$). Furthermore,

$$\hat{\sigma}^2_{BF} = \sum_{i=1}^{2} \frac{N\tilde{S}_i^2}{N - n_i} \quad \text{and}$$

$$\tilde{S}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(R_{ij} - R_{ij}^{(i)} - \overline{R}_i + \frac{n_i + 1}{2} \right)^2,$$

where $R_{ij}^{(i)}$ is the (within) rank of the observation X_{ij}, i.e., the rank among the n_i observations within group i. The statistic W_{BF} is asymptotically standard normal. However, for $\min(n_1,n_2) < 20$, one should use the t-distribution to compute a P-value. To be precise, the t-distribution with

$$df = \frac{\left(\sum_{i=1}^{2} \frac{\tilde{S}_i^2}{N - n_i} \right)^2}{\sum_{i=1}^{2} \frac{[\tilde{S}_i^2/(N - n_i)]^2}{n_i - 1}}$$

is appropriate (Brunner and Munzel, 2000, 2002, Chapter 2.1).

An unbiased and consistent estimator of p is $\hat{p} = (1/N)(\overline{R}_2 - \overline{R}_1) + (1/2)$. The corresponding 95% confidence interval for p is

$$\hat{p} \pm \frac{t_{df;0.975}}{n_1 n_2} \sqrt{\sum_{i=1}^{2} n_i \tilde{S}_i^2},$$

where $t_{df;0.975}$ denotes the 0.975 quantile of the t-distribution. In case of large sample sizes, the 0.975 quantile of the standard normal distribution (i.e., 1.96) may be used instead (Brunner and Munzel, 2002, Chapter 2.1). Confidence intervals with other coverage probabilities can be computed in the same way.