

A Bootstrap Test of Stochastic Equality of Two Populations

Jenő REICZIGEL, Ildikó ZAKARIÁS, and Lajos RÓZSA

When comparing two variables with nonnormal distributions, application of the Wilcoxon-Mann-Whitney test (WMW) is a common choice. However, it is only valid to test the null hypothesis stating equality of the distributions. Sometimes the hypothesis of interest is $H_0: P(X < Y) = P(X > Y)$ against $P(X < Y) \neq P(X > Y)$, called stochastic equality and inequality. Here we propose a bootstrap test for this problem. Results of an extensive simulation study based on empirical distributions suggest that the new test is valid for a wide range of problems in parasitology and psychology, and the loss of power as compared to WMW is rather small in those cases when both tests are applicable.

KEY WORDS: Brunner-Munzel test; Mann-Whitney U -test; Rank Welch test; Stochastic equality; Wilcoxon rank sum test.

1. INTRODUCTION

Nonnormal distributions are typical in several biomedical applications. Count variables, which do not take negative values, but may take very large positive values, usually exhibit rather skewed distributions. For example, in parasitology, intensity of infection is a variable of this type, defined as the number of parasites living in an infected host individual (Bush, Lafferty, Lotz, and Shostak 1997). Nonnormality of parasite intensity distributions is impressively illustrated by Figure 1 (graph is made by QP 2.0, Reiczigel and Rózsa 2001). Similar problems are documented in many other application areas, for example, in cost analysis (Rascati, Smith, and Neilands 2001; Zhou, Gao, and Hui 1997; Zhou, Li, and Gao 2001) or in psychology (Micceri 1989; Delaney and Vargha 2002). For the analysis of nonnormal data, statistics textbooks recommend the use of nonparametric methods, such as the Wilcoxon-Mann-Whitney (WMW) test for the comparison of two independent samples (Wilcoxon 1945; Mann and Whitney 1947). However, the standard assumption of the WMW test cited in many textbooks, the so-called shift model, or shift alternative, usually does not hold in these cases. Under the shift model, it is assumed that the distributions to be compared have the same shape, allowing only for a poten-

tial shift of location, that is, the distributions to be compared are $F(x)$ and $G(x) = F(x + d)$, with the null hypothesis of $H_0: d = 0$. The distributions in Figure 1 seem to be quite far from satisfying such an assumption. Even theoretically, the shift alternative would imply, for example, that parasite infection intensity may take negative values (which is simply nonsense) or conversely may exclude low intensity values near 0 (which is quite unrealistic). Unfortunately, the level of the WMW test is seriously affected by the violation of the shift model even for large samples (Pratt 1964; Skovlund and Fenstad 2001). In fact, the distributions do not really need to have different shapes, a pure scale difference is enough to result in considerable alpha inflation (it can be observed, e.g., if two uniform distributions are compared with the same center of location but with different ranges).

Some authors concluded that, contrary to the textbook recommendations, it is still better to apply the usual parametric tests even in case of nonnormality. For example, Skovlund and Fenstad (2001), comparing the Type I error rates of three tests (Student- t , Welch- t , and WMW) under various circumstances, found that the Welch test (Welch 1938) performs best, as it produces practically no alpha inflation. However, the WMW test is not so much related to the comparison of means as to the hypothesis of $H_0: P(X < Y) = P(X > Y)$ against $P(X < Y) \neq P(X > Y)$, also called stochastic equality/inequality (Vargha and Delaney 1998, 2000). Although there are other generalizations of WMW keeping the original H_0 and testing it against the so-called stochastic ordering alternative—that is, $H_0: F(x) = G(x)$ against $F(x) \leq G(x)$ with strict inequality for some x values (Kocher 1978; Deshpande and Kocher 1980; Ahmad 1996; Priebe and Cowen 1999)—here we focus on the above hypothesis test. There are procedures developed just for this particular problem (Zimmerman and Zumbo 1993;

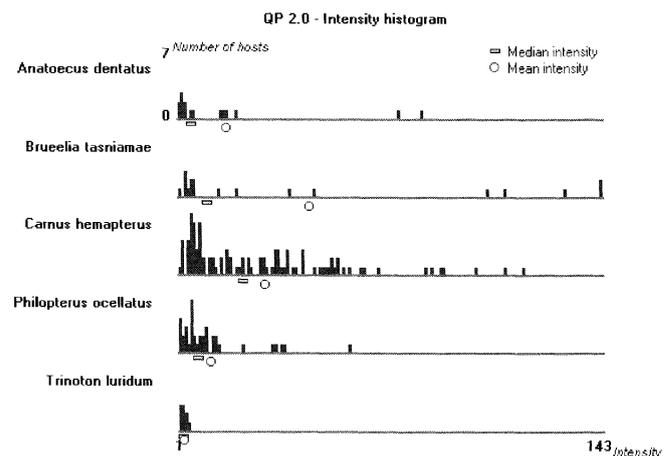


Figure 1. Empirical frequency distributions of parasite infection intensity.

Jenő Reiczigel is Senior Research Fellow, Szent István University, Faculty of Veterinary Science, Budapest, Hungary (E-mail: Reiczigel.Jeno@aotk.szie.hu). Ildikó Zakariás is Statistician, Johan Béla National Center for Epidemiology, Budapest, Hungary. Lajos Rózsa is Senior Research Fellow, Animal Ecology Research Group of the Hungarian Academy of Sciences, Hungarian Natural History Museum, Budapest, Hungary. This research was supported by the Hungarian National Research Fund, OTKA T035150. Some parts of the work were made when the first author was on a NATO research fellowship at the Philipps-University, Marburg, Germany. The authors are indebted to two anonymous referees and an associate editor for their valuable suggestions.

Delaney and Vargha 2002; Brunner and Munzel 2000). Here we propose a bootstrap test for the same purpose.

Section 2 briefly describes the WMW, the rank Welch test (RW), and the Brunner-Munzel test (BM), and their relation to testing for stochastic equality. Section 3 introduces the new bootstrap test. Section 4 presents a simulation study on the level and power of the new test, comparing it to the other tests. Sections 5 and 6 explain the results and summarize the conclusions.

2. THE WMW TEST AND RELATED PROCEDURES

Several equivalent formulations of the WMW test can be found in the literature, expressed in terms of various models, hypotheses, and test statistics. Here the variables to be compared are denoted by X and Y , their cdf's by $F(x)$ and $G(x)$, and the samples by $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_m)$, respectively. The test statistic can be written as

$$U = \sum_{x_i < y_j} 1 + \sum_{x_i = y_j} 1/2.$$

The distribution of U under the null hypothesis of the shift model, $H_0 : F = G$ can be determined using a permutation argument, that is, that all possible orderings of the values in the pooled sample are equally likely. For sample sizes $m, n > 10$, the null distribution of U is approximately normal with mean $\mu_U = nm/2$ and variance

$$\sigma_U^2 = \frac{nm(n+m+1)}{12}$$

(Mann and Whitney 1947). (In case of many ties, the formula for the variance needs an additional correction factor.) Thus, the statistic

$$z = \frac{U - \mu_U}{\sigma_U}$$

is approximately standard normal. The relation of the WMW test to the hypothesis of stochastic equality was defined by Vargha and Delaney (1998)—that is, the hypothesis of $P(X < Y) = P(X > Y)$ is explained by the fact that (U/nm) is an estimator of $P(X < Y) + 1/2P(X = Y)$. (Note that in case of continuous variables the second term vanishes.) Thus, the WMW test is sensitive only to such differences between F and G , which involve stochastic inequality of the two distributions. Several attempts have been made to replace the original $H_0 : F = G$ with a less restrictive one. Certainly the most desirable one would be the general hypothesis of stochastic equality, that is, $H_0 : P(X < Y) = P(Y < X)$, as this is the natural question of interest in many practical situations (McGraw and Wong 1992; Vargha and Delaney 2000). Conover and Iman (1981) pointed out that the WMW test is equivalent to a Student- t test applied to the ranks of the sample elements in the pooled sample. Based on this, Zimmerman and Zumbo (1993) proposed applying the Welch test to the ranks in the hope that the test remains valid for the general $H_0 : P(X < Y) = P(Y < X)$. Their results suggest that RW is really superior to WMW, but it may also show some alpha inflation in certain cases (Delaney and Vargha 2002).

To define the rank Welch test, let r_{1j} ($j = 1, \dots, n_1$) and r_{2k} ($k = 1, \dots, n_2$) denote the ranks of values of \mathbf{x} (sample 1) and \mathbf{y} (sample 2) in the pooled sample. Furthermore, let \bar{r}_1, \bar{r}_2 ,

and s_1^2, s_2^2 denote the means and variances of r_{1j} and r_{2k} , respectively. With this notation, the test statistic of the rank Welch test can be written as

$$t_{RW} = (\bar{r}_2 - \bar{r}_1) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The null distribution of t_{RW} is approximated by a Student- t distribution with degrees of freedom

$$df_{RW} = \frac{(n_2 s_1^2 + n_1 s_2^2)^2}{\frac{(n_2 s_1^2)^2}{n_1 - 1} + \frac{(n_1 s_2^2)^2}{n_2 - 1}}.$$

Brunner and Munzel (2000) proposed a test which is proven to be asymptotically valid. The test statistic is

$$t_{BM} = \frac{n_1 n_2 (\bar{r}_2 - \bar{r}_1)}{(n_1 + n_2) \sqrt{n_1 s_1^2 + n_2 s_2^2}},$$

where n_i ($i = 1, 2$) denote the sample sizes, \bar{r}_i ($i = 1, 2$) denote the mean rank of the i th sample within the pooled sample, and s_i^2 ($i = 1, 2$) is defined as follows:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(r_{ik} - w_{ik} - \bar{r}_i + \frac{n_i + 1}{2} \right)^2,$$

where r_{ik} and w_{ik} ($i = 1, 2; k = 1, 2, \dots, n_i$) are the rank of the k th measurement in the pooled sample and within its originating sample, respectively. The asymptotic distribution of t_{BM} is standard normal, but for small samples Brunner and Munzel (2000) proposed to use Student's- t distribution as an approximation, with degrees of freedom

$$df_{BM} = \frac{(n_1 s_1^2 + n_2 s_2^2)^2}{\frac{(n_1 s_1^2)^2}{n_1 - 1} + \frac{(n_2 s_2^2)^2}{n_2 - 1}}.$$

Other tests for stochastic equality were also proposed and evaluated by simulation; see, for example, Delaney and Vargha (2002) and references therein.

3. THE PROPOSED BOOTSTRAP TEST

The proposed method applies the bootstrap principle to testing $H_0 : P(X < Y) = P(X > Y)$. It is based on the rank Welch test statistic t_{RW} above. Following the usual method of bootstrap hypothesis testing (Efron and Tibshirani 1993), first the two samples \mathbf{x} and \mathbf{y} are transformed into \mathbf{x}' and \mathbf{y}' so as to satisfy the null hypothesis, that is, to be stochastically equal, possibly preserving all other characteristics of them. Then the null distribution of t_{RW} is obtained by resampling from the stochastically equal distributions \mathbf{x}' and \mathbf{y}' , that is, by drawing B bootstrap sample pairs of size n_1 and n_2 with replacement from \mathbf{x}' and \mathbf{y}' , and calculating the test statistic for each sample pair ($t_{RW}^{*(k)}, k = 1, 2, \dots, B$). Based on this simulated null distribution, the bootstrap p value is obtained as

$$p_1 = \frac{1}{B} \sum_{t_{RW} \geq t_{RW}^{*(k)}} 1$$

or

$$p_2 = \frac{1}{B} \sum_{t_{RW}^{*(k)} \leq t_{RW}} 1$$

in case of a one-tailed test, and $p = 2 \min\{p_1, p_2\}$ for a two-tailed test.

The transformation of \mathbf{x} and \mathbf{y} into \mathbf{x}' and \mathbf{y}' should reflect the characteristics of the distributions that are regarded as most important to preserve when adjusting the samples to satisfy the null hypothesis. Selection should always be based on considerations relevant to the specific application area as well as on statistical considerations (Davison and Hinkley 1997, p. 163). In fact, it is sufficient to transform just one of the samples, say \mathbf{y} , to make it stochastically equal to \mathbf{x} . In the present study, three potentially useful transformations were tried and a sensitivity analysis was made to compare the behavior of the bootstrap test with each of these transformations. The first one is the shift transformation (1), which is generally applied in relation with two-sample tests (Efron and Tibshirani 1993). The shift constant c_1 can be obtained as the median of the values $x_i - y_j$. For skewed distributions with a theoretical minimum $w \geq 0$, the stretch transformation (2) seems to be more natural, as it preserves the minimum w and the skewness. Note that in case of $w = 0$ the formula reduces to (2a). The appropriate stretch constant c_2 can be obtained as the median of the values $(x_i - w)/(y_j - w)$. In case of $w \geq 1$, another potentially useful transformation is the power transformation (3), in which the appropriate exponent c_3 can be obtained as the median of the values $\ln(x_i)/\ln(y_j)$.

$$y' = y + c_1, \quad (1)$$

$$y' = c_2(y - w) + w, \quad (2)$$

$$y' = c_2 y, \quad (2a)$$

$$y' = y^{c_3}. \quad (3)$$

Based on the sensitivity analysis, the usual shift transformation (1) was found to be most appropriate. Briefly, the other two transformations, while adjusting the distributions for stochastic equality, also change the ratio of the variances, which results in biased bootstrap null distributions. Computer programs for the bootstrap test procedure (executable for MS-Windows as well as S-Plus code) are available on request from the first author.

4. SIMULATION STUDY

The aim of the simulation study was to determine the level and power of the new test (bootstrap rank Welch test, BRW) and compare it to those of WMW, RW, and BM. Power was analyzed even in those cases when the assumptions of WMW held; given the power of the BRW is approximately the same as that of WMW, then BRW can substitute the other one without respect to the assumptions.

The simulation study was based mainly on empirical distributions: beyond a few theoretical distributions (uniform, Gaussian, and bimodal, the latter represented by a mixture of two Gaussians), 15 empirical parasite intensity distributions, and 50 empirical distributions of variables of the Rorschach test were

included. Empirical distributions are considered to represent the extent of alpha inflation that may be expected if applying the WMW test in everyday practical situations in parasitological or psychological research. In fact, several authors pointed out that using realistic data and distributions in simulation studies is more convincing than results based on data from purely theoretical distributions (Bridge and Sawilowsky 1999; Micceri 1989). Parasite distributions are based on empirical distributions of avian lice (for the sources of data see Rózsa, Reiczigel, and Majoros 2000). Rorschach distributions are based on data of 359 normal subjects, published in the tables of the Hungarian Rorschach Standard (Vargha 1989). All variables are in the form of number of occurrences in the Rorschach protocol divided by the total response number and multiplied with 100. Results are reported for seven pairs of distributions, representing typical patterns experienced in the simulation study. The distribution pair Uniform-Uniform with $U(0, 100)$ and $U(40, 60)$ illustrate the potential alpha inflation of the tests purely due to the different variances. The pair Unimodal-Bimodal (Gaussian with $\mu = 3$, $\sigma = .3$, and 50–50% mixture of Gaussians with $\mu_1 = 2$, $\mu_2 = 3$, $\sigma_1 = \sigma_2 = .4$), as well as the pair Bimodal-Bimodal (29–71% mixture of Gaussians with $\mu_1 = 4$, $\mu_2 = 8$, $\sigma_1 = \sigma_2 = .6$, and 71–29% mixture of Gaussians with $\mu_1 = 2$, $\mu_2 = 6$, $\sigma_1 = \sigma_2 = .6$) illustrate the effects of serious shape differences between distributions. The first pair of parasite distributions represents infection intensity of two louse species (*Brueelia tashniamae* and *Philopterus ocellatus*) collected from rooks (*Carvus frugilegus*). The second pair represents intensity of two louse species (*Anatoecus dentatus* and *Trinoton luridum*) from mallards (*Anas platyrhynchos*). Concerning the Rorschach variables, results are reported for the following ones: human movements (HM%), associated movements (AssocM%), number of responses on card 2 (RN2%), and defect references to color or shading (Defect C+Sh%). Their density graphs (made by S-Plus 2000) are shown in Figure 2.

Each distribution was represented by 300 values, corresponding to the equidistant quantiles (1/300 quantiles) of the distribution. For level analysis, distributions were adjusted to be stochastically equal, while for power analysis, they were transformed to be stochastically unequal, so that $P(X < Y) = .3$ and $P(X > Y) = .7$. Theoretical and Rorschach distributions were adjusted by the shift transformation (1), while parasite distributions, in order to produce realistic distributions, were adjusted by the stretch transformation (2). As infection intensity is defined only for infected hosts (Bush, Lafferty, Lotz, and Shostak 1997), and so the minimum intensity is one parasite per host, $w = 1$ was used.

To demonstrate actual level and power for all nominal levels rather than just for a selected one, we report the empirical cdf of the p value estimated from 10,000 sample pairs by sampling with replacement from the above populations. Sample sizes are set to 10, 30, and 90, representing small and medium sample sizes, typical in practical situations. All four tests (WMW, RW, BM, and BRW) were applied to these samples. In BRW 1,000 bootstrap replications were used. Figures 3 and 4 display the results with respect to level for selected pairs of distributions (sample sizes of 90 gave quite similar results as sample sizes of 30, therefore not reported). The bootstrap test proved to be valid in the whole range of tested distributions at 5% (remain-

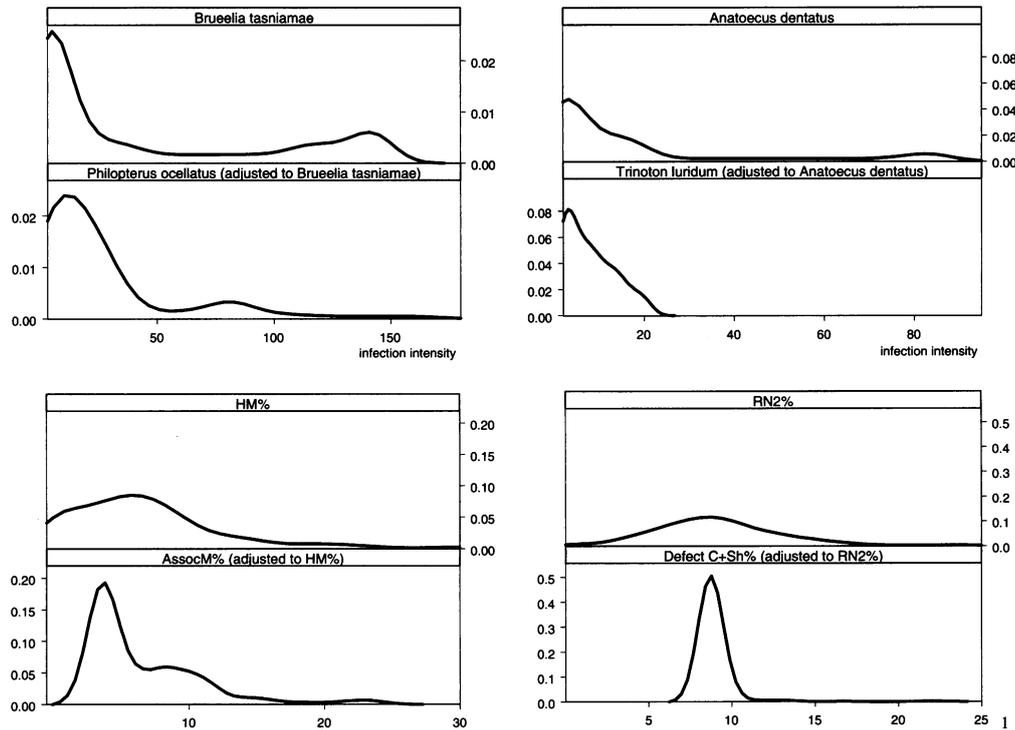


Figure 2. Density graphs of distributions derived from empirical parasite intensity distributions and Rorschach distributions and adjusted to be stochastically equal for level analysis. Descriptive statistics (mean, D , skewness, kurtosis): *Brueelia* (44.2, 53.0, .92, $-.85$), *Philopterus* (26.9, 31.6, 2.16, 4.67); *Anatoecus* (62.2, 163.6, 3.07, 8.75), *Trinoton* (4.22, 2.53, 0.67, $-.60$), HM% (6.81, 5.65, 1.47, 2.87), AssocM% (6.59, 4.30, 1.84, 3.69), RN2% (9.18, 3.94, .84, 1.92), Defect C + Sh% (8.91, 1.07, 8.37, 86.6).

ing below 6%), and slightly liberal at 1% (going up to about 1.5%). BM performed well for samples ≥ 30 and in case of not too extreme differences between the distributions also for $n = 10$, while WMW and RW showed considerable alpha inflation throughout. Simulation with unbalanced sample sizes (10–30 and 30–90) showed that alpha inflation of WMW was highest when the smaller sample was drawn from the distribution with greater variance, whereas for RW the converse occurred (results

not reported). Figure 5 illustrates power properties for two pairs of uniform distributions. Graphs are quite typical, that is, very similar results were obtained for other distributions.

5. DISCUSSION

It is known (Pratt 1964; Skovlund and Fenstad 2001) that the WMW test may produce considerable alpha inflation even for large samples if the shift alternative does not hold. One could

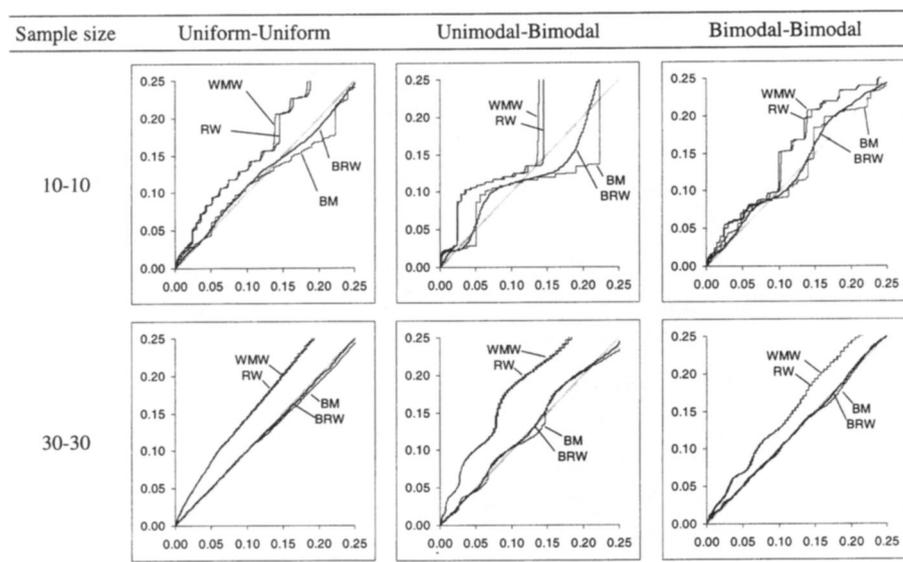


Figure 3. Graphs of the empirical cdfs of the p values under $H_0: P(X < Y) = P(X > Y)$ for selected distributions. A certain (x, y) point of a graph can be interpreted as the test has an actual alpha error rate of y at a nominal alpha of x . Each graph is based on 10,000 pairs of random samples drawn from the given populations. The line $y = x$ is also displayed (in gray) for information.

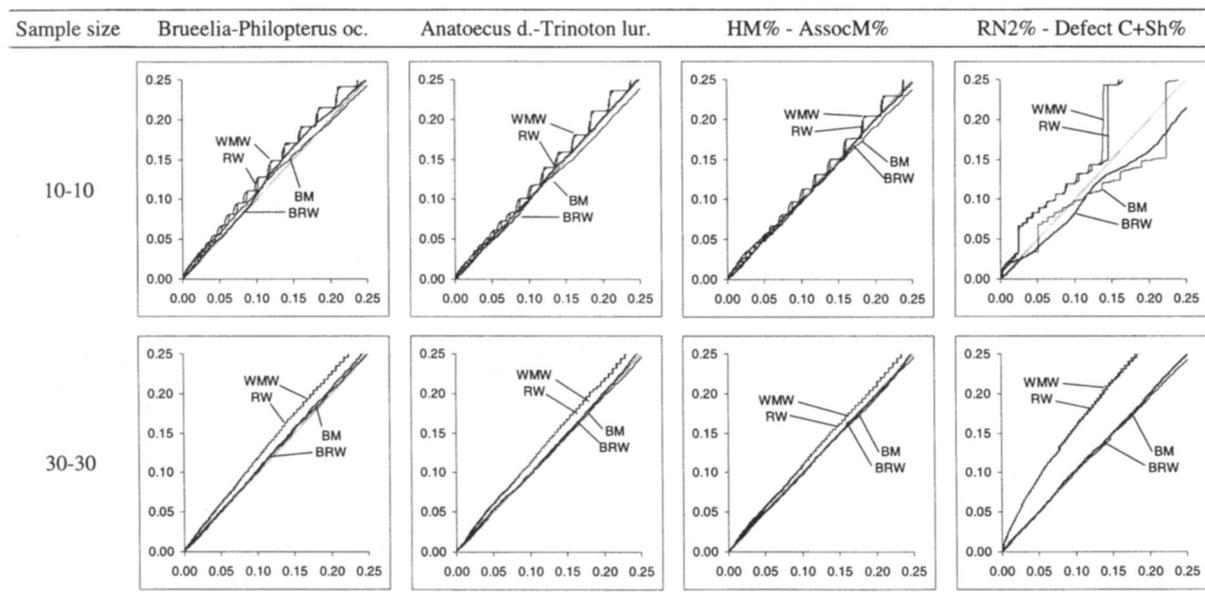


Figure 4. Graphs of the empirical cdfs of the p values under $H_0: P(X < Y) = P(X > Y)$ for selected pairs of parasite intensity distributions and Rorschach distributions.

think, however, that this is of theoretical interest only, and does not cause serious problems with real data in everyday practical situations. Contrary to this presumption, our simulations with realistic parasite and Rorschach distributions demonstrated that the actual alpha error rate may be about 10% instead of the nominal 5%. This implies that one should avoid using the WMW test in such cases. RW showed almost the same poor behavior, thus it should not be used either. Results are in line with those by Delaney and Vargha (2002). The proposed new bootstrap test had alpha level about the nominal (although slightly liberal at 1%). BM also performed well for sample sizes ≥ 30 . Similar results were obtained in case of one-tailed testing (results not reported here).

Simulations showed that WMW and RW, unlike BM, have seriously inflated alpha error rates even for large samples ($n = 90$). For $n < 30$ smaller or bigger jumps of the p value cdf make BM also rather unsatisfactory. The location and size of these jumps depend on the distributions and sample sizes, resulting in a serious dependence of the alpha error rate on sample size (Table 1). It is embarrassing to see that drawing samples of 15 or 16 from the same populations and performing the same test on them result in quite different alpha error rates. Similar jumps can be observed in relation with power as well (Figure 5). Such strong dependence on sample size may cause controversy between simulation studies using different sample sizes, or may lead to misleading results as round sample sizes like 10 or 15 are more often used than 11 or 16. The existence of the jumps is a consequence of calculating the test statistic from the ranks, because the assumptions “all possible samples occur equally likely” and “all possible rank orderings occur equally likely” may well differ for some distributions. For example, drawing samples of two with replacement from populations $A = \{1, 4\}$ and $B = \{2, 3\}$, there are 16 possible sample pairs but only six possible orderings: $P(AABB) = 4/16$, $P(ABAB) = 0$, $P(ABBA) = 8/16$, $P(BAAB) = 0$, $P(BABA) = 0$, $P(BBAA) = 4/16$. If the test statistic is calculated from the ranks, then its values inherit the probabilities of the orderings, resulting in jumps of the cdf of the

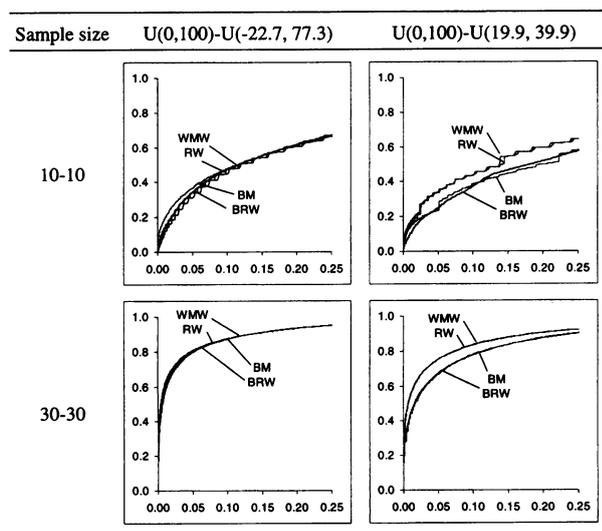


Figure 5. Graphs of the empirical cdfs of the p values under $H_1: P(X < Y) = .3, P(X > Y) = .7$ assuming uniform distributions. Note that although WMW and RW are shown in the second column too for information, these tests are inapplicable in that case.

Table 1. Type I Error Rates of a Nominal 5% BM Test for Different Sample Sizes and Distributions (estimated from 10,000 replications)

Sample sizes	Distributions	
	Unimodal-bimodal	Bimodal-bimodal
15-15	.0432	.0675
15-16	.0433	.0558
16-15	.0737	.0539
16-16	.0670	.0509

test statistic. If the test statistic is compared to a continuous reference distribution as null distribution, jumps also appear in the cdf of the p value. Jumps occur where there are blocks of values from one distribution without values from the other distribution between them. This is likely to happen if in a particular interval the density of one distribution is high relative to the other one, which occurs easily with multimodal distributions. BRW can avoid such jumps due to resampling, because the test statistic is compared from sample to sample to another reference distribution, but some waving of the p value cdf can be observed here too.

For discrete variables with just a few distinct values BRW may fail, since it may be impossible to adjust the samples to be stochastically equal, and resampling from stochastically unequal distributions results in a biased null distribution. The following pair of samples serves as an example of this: Sample A : 1, 2, 2, 3, 3, 3; Sample B : 2, 2, 2, 3, 3, 4. Here $U = \sum_{A_i < B_j} 1 + \sum_{A_i = B_j} 1/2 = 21/36 = .583$. If adjusting Sample B downwards by a small shift, say, B' : 1.9999, 1.9999, 1.9999, 2.9999, 2.9999, 3.9999, U makes a big step downwards far beyond .5: $U = \sum_{A_i < B_j} 1 + \sum_{A_i = B_j} 1/2 = 15/36 = .417$. Here it is the discreteness of the sample, that is, the many ties, rather than the shift adjustment that is responsible for this; any strictly monotone transformation would produce the same result. (Although transforming the distribution, that is, the frequencies of the values, rather than transforming the values themselves might help.)

Other procedures, such as variants of the Student- t test, the median test, and so on, are often considered as potential alternatives of the WMW test. However, we must emphasize that these procedures test quite different hypotheses, for example, the equality of means or medians, which is not at all equivalent to $H_0 : P(X < Y) = P(X > Y)$ (Hart 2001; Rózsa et al. 2001). One should test that particular hypothesis, which fits to the specific biological question of interest, and should not change it for purely statistical reasons (Thompson and Barber 2000; Zhou et al. 2001). Though we admit that making such fine distinctions is still not typical in everyday practice, there is another, rather pragmatic argument against substituting the WMW by variants of the t test or the median test. The WMW test, if applicable, has considerably higher power than these alternative tests. Note that the proposed bootstrap test both preserves this advantage and has wider applicability.

6. CONCLUSION

Using WMW or RW to test for stochastic equality results in considerable alpha inflation and therefore should be avoided. Alpha inflation can be demonstrated even for typical, everyday data in parasitology and psychology. Simulation results based on empirical distributions suggest that BRW maintains the alpha level well. As the loss of power compared to WMW is small when both tests are applicable, the strategy of always using BRW is also reasonable. For moderate and large samples (≥ 30) also BM performs well.

[Received April 2003. Revised November 2004.]

REFERENCES

Ahmad, I. A. (1996), "A Class of Mann-Whitney-Wilcoxon Type Statistics,"

- The American Statistician*, 50, 324–327.
- Bridge P. D., and Sawilowsky, S. S. (1999), "Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t Test and Wilcoxon Rank-Sum Test in Small Samples Applied Research," *Journal of Clinical Epidemiology*, 52, 229–235.
- Brunner, E., and Munzel, U. (2000), "The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation," *Biometrical Journal*, 42, 17–25.
- Bush, A. O., Lafferty, K. D., Lotz, J. M., and Shostak, A. W. (1997), "Parasitology Meets Ecology on its Own Terms: Margolis et al. revisited," *Journal of Parasitology*, 83, 575–583.
- Conover, W. J., and Iman, R. L. (1981), "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics," *The American Statistician*, 35, 124–129.
- Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge, UK: Cambridge University Press.
- Delaney, H. D., and Vargha, A. (2002), "Comparing Several Robust Tests of Stochastic Equality with Ordinally Scaled Variables and Small to Moderate Sized Samples," *Psychological Methods*, 7, 485–503.
- Deshpande, J. V., and Kochar, S. C. (1980), "Some Competitors of Tests Based on Powers of Ranks for the Two-Sample Problem," *Sankhya B*, 42, 236–241.
- Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.
- Hart, A. (2001), "Mann-Whitney Test is Not Just a Test of Medians: Differences in Spread can be Important," *British Medical Journal*, 323, 391–393.
- Kochar, S. C. (1978), "A Class of Distribution-Free Tests for the Two-Sample Slippage Problem," *Communications in Statistics, Part A—Theory and Methods*, 7, 1243–1252.
- Mann, H. B., and Whitney, D. R. (1947), "On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other," *Annals of Mathematical Statistics*, 18, 50–60.
- McGraw, K. O., and Wong, S. P. (1992), "A Common Language Effect Size Statistic," *Psychological Bulletin*, 111, 361–365.
- Micceri, T. (1989), "The Unicorn, the Normal Curve, and Other Improbable Creatures," *Psychological Bulletin*, 111, 361–365.
- Pratt, J. W. (1964), "Robustness of Some Procedures for the Two-Sample Location Problem," *Journal of the American Statistical Association*, 59, 665–680.
- Priebe, C. E., and Cowen, L. J. (1999), "A Generalized Wilcoxon-Mann-Whitney Statistic," *Communications in Statistics, Part A—Theory and Methods*, 28, 2871–2878.
- Rascati, K. L., Smith, M. J., and Neilands, T. (2001), "Dealing with Skewed Data: An Example Using Asthma-Related Costs of Medicaid Clients," *Clinical Therapeutics*, 23, 481–498.
- Reiczigel, J., and Rózsa, L. (2001), "Quantitative Parasitology 2.0," Budapest, distributed by the authors (free download from <http://bio.univet.hu>).
- Rózsa, L., Reiczigel, J., and Majoros, G. (2000), "Quantifying Parasites in Samples of Hosts," *Journal of Parasitology*, 86, 228–232.
- Skovlund, E., and Fenstad, G. U. (2001), "Should We Always Choose a Nonparametric Test When Comparing Two Apparently Nonnormal Distributions?" *Journal of Clinical Epidemiology*, 54, 86–92.
- Thompson, S. G., and Barber, J. A. (2000), "How Should Cost Data in Pragmatic Randomised Trials be Analysed?" *British Medical Journal*, 320, 1197–1200.
- Vargha, A. (1989), "The Tables of the Hungarian Rorschach Standard" (in Hungarian), *Tankönyvkiadó*, Budapest, Hungary.
- Vargha, A., and Delaney, H. D. (1998), "The Kruskal-Wallis Test and Stochastic Homogeneity," *Journal of Educational and Behavioral Statistics*, 23, 170–192.
- (2000), "A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong," *Journal of Educational and Behavioral Statistics*, 25, 101–132.
- Welch, B. L. (1938), "The Significance of the Difference Between Two Means When the Population Variances are Unequal," *Biometrika*, 29, 350–362.
- Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics*, 1, 80–83.
- Zhou, X. H., Gao, S. J., and Hui, S. L. (1997), "Methods for Comparing the Means of Two Independent Log-Normal Samples," *Biometrics*, 53, 1129–1135.
- Zhou, X. H., Li, C. M., and Gao, S. J. (2001), "Methods for Testing Equality of Means of Health Care Costs in a Paired Design Study," *Statistics in Medicine*, 20, 1703–1720.
- Zimmerman, D. W., and Zumbo, B., D. (1993), "Rank Transformations and the Power of the Student t -Test and Welch t -Test for Nonnormal Populations with Unequal Variances," *Canadian Journal of Experimental Psychology*, 47, 523–539.