Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Combining the *t* test and Wilcoxon's rank-sum test

Markus Neuhäuser

To cite this article: Markus Neuhäuser (2015) Combining the *t* test and Wilcoxon's rank-sum test, Journal of Applied Statistics, 42:12, 2769-2775, DOI: 10.1080/02664763.2015.1070809

To link to this article: https://doi.org/10.1080/02664763.2015.1070809

Journal of Applied Statistics, 2015
Vol. 42, No. 12, 2769–2775, http://dx.doi.org/10.1080/02664763.2015.1070809

Taylor & Francis
Taylor & Francis Group

Combining the *t* test and Wilcoxon's rank-sum test

Markus Neuhäuser*

Department of Mathematics and Technology, Koblenz University of Applied Sciences, RheinAhrCampus, Joseph-Rovan-Allee 2, 53424, Remagen, Germany

(*Received 17 September 2014; accepted 6 July 2015*)

In the two-sample location-shift problem, Student's *t* test or Wilcoxon's rank-sum test are commonly applied. The latter test can be more powerful for non-normal data. Here, we propose to combine the two tests within a maximum test. We show that the constructed maximum test controls the type I error rate and has good power characteristics for a variety of distributions; its power is close to that of the more powerful of the two tests. Thus, irrespective of the distribution, the maximum test stabilizes the power. To carry out the maximum test is a more powerful strategy than selecting one of the single tests. The proposed test is applied to data of a clinical trial.

Keywords: asymptotic relative efficiency; maximum test; non-normal data; permutation test; two-sample problem

1. Introduction

In this paper, the two-sample location-shift, or constant effect, model is considered. Hence, it is assumed that the population distributions from which the two samples are selected are identical except for a possible change in location. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ denote the two random samples. The observations within each sample are independent and identically distributed, and independence between the two samples is assumed. Let F_1 and F_2 be the distribution functions, possibly non-continuous, corresponding to the populations of X and Y values, respectively. In the location-shift model, we have $F_1(t) = F_2(t-\theta)$ for every t. The null hypothesis is H_0: $\theta = 0$, whereas the alternative H_1 states $\theta \neq 0$.

If F_1 and F_2 were normal distributions Student's two-sample *t* test is the uniformly most powerful unbiased test for H_0 versus H_1. However, in practical applications, distributions usually deviate more or less from a normal distribution. In the literature, there is no agreement how severe a deviation has to be before the *t* test should be replaced [22]. The most popular non-parametric alternative to Student's *t* test is the Wilcoxon rank-sum test which is equivalent to the

*Email: neuhaeuser@rheinahrcampus.de

Mann–Whitney U test [15]. Under non-normality, the rank-sum test can be (much) more powerful than the t test. Apart from the asymptotic relative efficiency (ARE) of the Wilcoxon test to the t test, which can exceed 1 [10], the better power characteristics were shown in computer simulations (e.g. [2–4,14]; or [19] for small-sample cases).

The article is organized as follows: a maximum test is introduced in Section 2 and applied to data of a placebo-controlled clinical trial in Section 3. Section 4 presents the results of a simulation study, whereas the ARE is discussed in Section 5. A concluding discussion is given in Section 6.

2. The proposed maximum test

Student's t statistic is defined as $t = \sqrt{nm/(n+m)}(\bar{X} - \bar{Y})/S$, where $\bar{X}$ and $\bar{Y}$ are the sample means and S is the square root of the pooled-sample variance estimator. One might compute Student's t statistic on the ranks, the resulting statistic is denoted by t_R. Let R_i be the ranks from 1 to N ($N = n + m$), and W the sum of the ranks of the X's. Then, t_R can be computed as [7]

$$t_R = \frac{1}{n}W - \frac{1}{m}\left(\frac{N(N+1)}{2} - W\right) \div \left[\left(\sum_{i=1}^{N} R_i^2 - \frac{1}{n}W^2 - \frac{1}{m}\left(\frac{N(N+1)}{2} - W\right)^2\right)\frac{N}{nm(N-2)}\right]^{1/2}.$$

Conover and Iman [7] showed that

$$t_R = \frac{W_{\text{std}}}{\sqrt{(N-1)/(N-2) - (1/(N-2))W_{\text{std}}^2}},$$

where W_{std} is the standardized form of the rank sum W with the adjustment for ties incorporated. Therefore, t_R is a monotonically increasing function of W_{std}, and the Wilcoxon rank-sum test can be performed using t_R as test statistic.

The Wilcoxon test can be carried out as a permutation test, that is, based on the permutation null distribution of the rank sum. This is especially useful in case of small samples, in the presence of ties, or when treatments have been assigned to the experimental units at random [15].

A permutation test can also be performed with Student's t statistic, this test is often called Fisher–Pitman permutation test [15]. The ARE of the Fisher–Pitman permutation test to the t test is 1 [12]. In a simulation study presented by van den Brink [6], the powers of the two tests were nearly equal. Thus, there is no loss in power when the Fisher–Pitman permutation test is applied instead of Student's original t test. Note that there are several equivalent test statistics that can be used to carry out the Fisher–Pitman permutation test, for instance, $\bar{X} - \bar{Y}$ [15].

In practical applications, the underlying distribution is usually unknown and therefore it is hard to decide which test to apply. An alternative to choosing one single test statistic is the use of the maximum of the competing statistics as the test statistic. Therefore, we suggest using the maximum $\max(|t|, |t_R|)$ for a two-sided test. Here, permutation tests are applied with this statistic. Of course, one-sided tests are possible without using the absolute values.

3. Example

As an example, we consider a clinical trial presented by Sedlmeier and Renkewitz [18]. Twenty patients were randomized to an active drug and placebo, respectively. Table 1 presents the measured reaction times necessary to react on a visual signal.

When applying Student's t test, we get $t = 1.800$ and the two-sided p-value 0.0886 is obtained. The Fisher–Pitman permutation test gives a similar p-value: 0.0919. However, the Wilcoxon rank-sum test is significant at $\alpha = 5\%$: with the test statistic $t_R = 2.357$ the p-value of the exact permutation test is 0.0355. The proposed maximum test has a p-value close to this value: 0.0390.

Table 1. Reaction times in msec in a two-arm clinical trial.

Active drug (X values, $n = 10$)	Placebo (Y values, $m = 10$)
171, 172, 178, 179, 184, 185, 186, 194, 196, 223	154, 155, 158, 159, 161, 163, 177, 183, 192, 219

Data source: Sedlmeier and Renkewitz [18].

4. Simulation study

In a simulation study performed with R, we compared the maximum test with the two tests that were used to construct this new test. For each configuration, 10,000 simulation runs were performed. Results for the balanced sample sizes $n = m = 10$ are given in Figure 1 and Table 2, whereas results for the unbalanced case $n = 2m = 10$ are displayed in Table 3. We present results for the following distributions: standard normal, exponential with rate parameter $\lambda = 1$, uniform on (0, 1), Poisson with mean $\lambda = 5$, and a contaminated normal distribution which was simulated as follows: a value comes from a standard normal distribution with probability 0.7, and with probability 0.3 from a normal distribution with mean 5 and standard deviation 4. For group 2, these data were shifted by an amount ranging from 0 to 2, as specified in Figure 1 and Tables 2

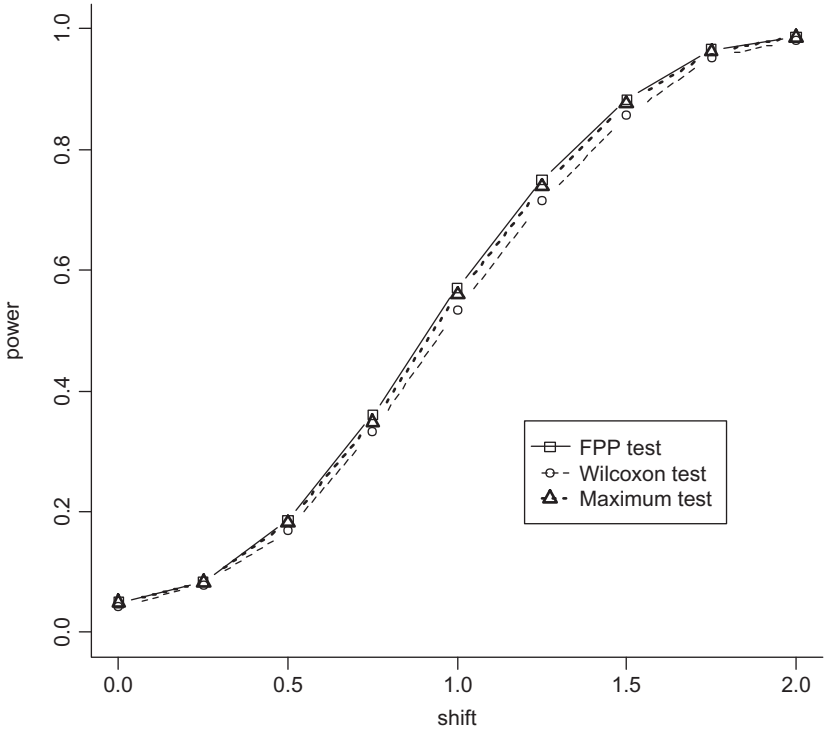

Figure 1. Simulation results for normally distributed data (group 1: standard normal distribution, group 2: normal distribution with mean *shift* and variance 1), for $n = m = 10$; FPP test = Fisher–Pitman permutation test.

Table 2. Simulation results for $n = m = 10$.

Shift (in group 2)	FPP[a] test	Rank-sum test	Maximum test
Exponential distribution with rate parameter $\lambda = 1$ in group 1			
0	0.050	0.043[b]	0.050
0.5	0.25	0.32	0.32
1.0	0.63	0.72	0.73
1.5	0.88	0.91	0.92
Uniform distribution on (0, 1) in group 1			
0	0.048	0.043[b]	0.047
0.5/3	0.21	0.19	0.21
1/3	0.68	0.60	0.65
1.5/3	0.96	0.92	0.95
Contaminated normal distribution[c] in group 1			
0	0.053	0.043[b]	0.053
0.5/0.3	0.24	0.42	0.40
1/0.3	0.60	0.74	0.74
1.5/0.3	0.88	0.89	0.90
Poisson distribution with mean $\lambda = 5$ in group 1			
0	0.040	0.048	0.047
0.5/0.4	0.21	0.29	0.27
1/0.4	0.64	0.65	0.65
1.5/0.4	0.94	0.90	0.93

Notes: [a]Fisher–Pitman permutation test.
[b]The actual type I error rate can be exactly determined for a rank test in case of a continuous distribution.
[c]A value comes from a standard normal distribution with probability 0.7, and with probability 0.3 from a normal distribution with mean 5 and standard deviation 4.

and 3. For the shift 0, the null hypothesis is true, thus, the actual type I error rate is estimated by the simulation, for a shift $\neq 0$ the power is estimated.

Exact permutation tests guarantee that the actual type I error rate does not exceed the nominal one. However, the rank-sum test is conservative for small-sample sizes. The Fisher–Pitman permutation test is known to be much less conservative, at least for continuous distributions [15]. The new maximum test is also less conservative in comparison to the Wilcoxon test. Some simulated actual type I error rates are slightly larger than $\alpha = 0.05$; this is caused by the simulation error. In these cases, the actual size is close to 0.05 and the simulated estimate varies around this true value (which cannot be above α in a permutation test).

Sometimes, there is no big difference in power, as for instance in the case displayed in Figure 1 where the three curves are very close to each other. However, bigger differences in power occur (see Tables 2 and 3). Most power values based on the maximum test are slightly smaller than or equal to the better of the two single tests, but, irrespective of the distribution, the power of the maximum test is close to that of the better of the two single tests. If the underlying distribution was known a priori, one could choose the best (single) test, in this case, the maximum test offers no gain. However, the usual situation in practice is that the underlying distribution is not known. Thus, when choosing a test, one cannot know which single test is more powerful. In this common situation, that is, when there is no a priori reason to select either the t or the rank-sum test, the maximum test is a good choice because it stabilizes the power. The maximum test performs well over a broad spectrum of distributions.

The maximum test can even be slightly more powerful than the best single test, especially in cases where there is no big difference in power between the Fisher–Pitman permutation test and the Wilcoxon test. Note that it is one of the advantages of a maximum test that it can be even more powerful than the competing univariate tests [16].

Table 3. Simulation results for $n = 10$ and $m = 5$.

Shift (in group 2)	FPP[a] test	Rank-sum test	Maximum test
Standard normal distribution in group 1			
0	0.049	0.040[b]	0.048
0.5	0.13	0.11	0.12
1.0	0.39	0.34	0.38
1.5	0.72	0.66	0.71
2.0	0.93	0.89	0.92
Exponential distribution with rate parameter $\lambda = 1$ in group 1			
0	0.051	0.040[b]	0.047
0.5	0.19	0.18	0.19
1.0	0.48	0.51	0.53
1.5	0.75	0.79	0.80
2	0.90	0.92	0.93
Uniform distribution on (0, 1) in group 1			
0	0.050	0.040[b]	0.048
0.5/3	0.16	0.14	0.16
1/3	0.48	0.41	0.45
1.5/3	0.83	0.74	0.81
2/3	0.99	0.95	0.98
Contaminated normal distribution[c] in group 1			
0	0.053	0.040[b]	0.051
0.5/0.3	0.20	0.26	0.26
1/0.3	0.46	0.56	0.56
1.5/0.3	0.73	0.73	0.76
2/0.3	0.91	0.87	0.90
Poisson distribution with mean $\lambda = 5$ in group 1			
0	0.044	0.047	0.047
0.5/0.4	0.16	0.20	0.19
1/0.4	0.46	0.45	0.47
1.5/0.4	0.81	0.73	0.79
2/0.4	0.96	0.96	0.96

Notes: [a]Fisher–Pitman permutation test.
[b]The actual type I error rate can be exactly determined for a rank test in case of a continuous distribution.
[c]A value comes from a standard normal distribution with probability 0.7, and with probability 0.3 from a normal distribution with mean 5 and standard deviation 4.

5. Asymptotic relative efficiency

Let N_1 and N_2 be the total sample sizes required by two tests T_1 and T_2 to achieve the same power with an identical significance level. Then, the ARE according to Pitman of test T_1 to test T_2 is [12]

$$\text{ARE} = \lim \frac{N_2}{N_1} \text{ as } N_1, N_2 \to \infty.$$

A more formal definition is given, for example, by Neuhäuser [15].

The ARE of the Wilcoxon rank-sum test to Student's t test was calculated by Hodges and Lehmann [10]. When this ARE is smaller than 1, the t test is better, if ARE > 1, the Wilcoxon test is better. For example, for a normal distribution, the ARE is $3/\pi = 0.955$, for an exponential distribution, we have ARE $= 3$ [10,12].

As mentioned above, the ARE of the FPP test to the t test is 1 for essentially all distributions [12], thus the ARE of the Wilcoxon rank-sum test to the FPP test is identical to the ARE of the Wilcoxon rank-sum test to the t test.

Table 4. Convergence of the maximum test statistic $\max(|t|, |t_R|)$.

Sample size (per group)[a]	Average difference between maximum and t[b]	Average difference between maximum and t_R[c]
Standard normal distribution in group 1, shift (in group 2): 0.5		
10	0.0978	0.1076
100	0.0602	0.1131
500	0.0392	0.1497
1000	0.0264	0.1866
5000	0.0035	0.3581
10,000	0.0004	0.4986
Exponential distribution with rate parameter $\lambda = 1$ in group 1, shift (in group 2): 0.5		
10	0.4100	0.0657
50	1.0740	0.0031
75	1.3341	0.0005
100	1.5445	0.0001
150	1.9125	0.0000
200	2.2288	0.0000

Notes: [a]Balanced design.
[b]Average of the difference $|\max(|t|, |t_R|) - |t|\,|$ from 10,000 simulations.
[c]Average of the difference $|\max(|t|, |t_R|) - |t_R|\,|$ from 10,000 simulations.

When $N_1, N_2 \to \infty$, the test statistic of the proposed maximum test, $\max(|t|, |t_R|)$, goes to $|t|$ if the t test is better, and to $|t_R|$ if the Wilcoxon test is better, respectively. Table 4 demonstrates this convergence for two distributions. Hence, asymptotically there is no difference between the maximum and the better one of the two single test statistics.

6. Discussion

The use of a maximum test is quite common, for instance in statistical genetics (for references, see [11,16]). In some cases, when the correlation between the different test statistics is known, the (asymptotic) distribution of the maximum can be used for inference; an example is a multiple contrast test [5]. However, permutation tests may be preferable even in these cases [1,13]. Freidlin and Korn [9] presented an example where the approximation using the asymptotic distribution of a maximum can be poor even when all univariate test statistics are asymptotically normal.

We proposed to carry out the maximum test as an exact permutation test. Therefore, there is no need to restrict the application of the maximum test to continuous distributions. Exact permutation tests can be carried out for discrete data as well. When sample sizes are large, an approximate permutation test can be performed using a simple random sample of permutations; often 10,000 permutations are used. SAS and R programs to implement permutation tests can be found, for example, in Zieffler, Harring, and Long [21] and Neuhäuser [15].

As in Tippett's combination procedure, the minimum p-value may be used instead of the maximum test statistic [20]. There are further combination functions that could also be used; for an overview of nonparametric combination methodology, see Pesarin [17]. However, using the maximum is a common procedure, it is easily interpretable, and a maximum test has 'useful diagnostic properties' [8]. In the example, the much higher value of t_R in comparison to t indicates that these data are far from being normal.

Non-normal data are common in practical applications. Then, the presented maximum test can be applied. As a consequence, there is no need to select one of the two tests, Student's t test or Wilcoxon's rank-sum test. Nevertheless, the applied maximum test has a power very similar

to that of the better test. It should be noted that at least interval measurements of the variable being studied is required for the maximum test since the test utilizes the observed numerical values, whereas the Wilcoxon test can be applied for ordinal data as well. However, the question which of the two tests to apply does not exist in case of ordinal data as Student's *t* test, or the Fisher–Pitman permutation test, cannot be applied. Therefore, there is no need for the presented maximum test if these data were ordinal categories.

Disclosure statement

No potential conflict of interest was reported by the author.

References

[1] V.W. Berger, *A Socratic dialogue*, J. Mod. Appl. Stat. Methods. 8 (2009), pp. 316–321.
[2] R.C. Blair and J.J. Higgins, *The power of t and Wilcoxon statistics: A comparison*, Eval. Rev. 4 (1980), pp. 645–656.
[3] R.C. Blair and J.J. Higgins, *A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions*, J. Educ. Stat. 5 (1980), pp. 309–335.
[4] R.C. Blair, J.J. Higgins, and W.D.S. Smitley, *On the relative power of the U and t tests*, Br. J. Math. Stat. Psychol. 33 (1980), pp. 114–120.
[5] F. Bretz and L.A. Hothorn, *Detecting dose–response using contrasts: Asymptotic power and sample size determination for binomial data*, Stat. Med. 21 (2002), pp. 3325–3335.
[6] W.P. van den Brink and S.G.J. van den Brink, *A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem*, Br. J. Math. Stat. Psychol. 42 (1989), pp. 183–189.
[7] W.J. Conover and R.L. Iman, *Rank transformation as a bridge between parametric and nonparametric statistics*, Amer. Statist. 35 (1981), pp. 124–129.
[8] D.R. Cox, *The role of significance tests*, Scand. J. Statist. 4 (1977), pp. 49–70.
[9] B. Freidlin and E.L. Korn, *A testing procedure for survival data with few responders*, Stat. Med. 21 (2002), pp. 65–78.
[10] J.L. Hodges and E.L. Lehmann, *The efficiency of some nonparametric competitors of the t-test*, Ann. Math. Statist. 27 (1956), pp. 324–335.
[11] T. Hothorn, K. Hornik, M.A. van de Wiel, and A. Zeileis, *A lego system for conditional inference*, Amer. Statist. 60 (2006), pp. 257–263.
[12] E.L. Lehmann, *Parametric versus nonparametrics: Two alternative methodologies*, J. Nonparametr. Stat. 21 (2009), pp. 397–405.
[13] J. Ludbrook, and H. Dudley, *Why permutation tests are superior to t and F tests in biomedical research*, Amer. Statist. 52 (1998), pp. 127–132.
[14] H.R. Neave and C.W.J. Granger, *A Monte Carlo study comparing various two-sample tests for differences in means*, Technometrics. 10 (1968), pp. 509–522.
[15] M. Neuhäuser, *Nonparametric Statistical Tests: A Computational Approach*, CRC Press, Boca Raton, FL, 2012.
[16] M. Neuhäuser and L.A. Hothorn, *Maximum tests are adaptive permutation tests*, J. Mod. Appl. Stat. Methods. 5 (2006), pp. 317–322.
[17] F. Pesarin, *Multivariate Permutation Tests*, Wiley, New York, NY, 2001.
[18] P. Sedlmeier and F. Renkewitz, *Forschungsmethoden und Statistik in der Psychologie*, Pearson, Munich, Germany, 2008.
[19] H. Tanizaki, *Power comparisons of nonparametric tests: Small-sample properties from Monte Carlo experiments*, J. Appl. Stat. 24 (1997), pp. 603–632.
[20] M. Weichert, and L.A. Hothorn, *Robust hybrid tests for the two-sample location problem*, Comm. Statist. Simulation Comput. 31 (2002), pp. 175–187.
[21] A.S. Zieffler, J.R. Harring, and J.D. Long, *Comparing Groups: Randomization and Bootstrap Methods Using R*, Wiley, Hoboken, NJ, 2011.
[22] D.W. Zimmerman and B.D. Zumbo, *The relative power of parametric and nonparametric statistical methods*, in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, G. Keren and C. Lewis, eds., Lawrence Erlbaum, Hillsdale, NJ, 1993, pp. 481–517.