Journal of Modern Applied Statistical Methods

Volume 5 | Issue 2 Article 4

11-1-2005

Maximum Tests are Adaptive Permutation Tests

Markus Neuhäeuser
Koblenz University of Applied Sciences, Remagen, Germany, neuhaeuser@rheinahrcampus.de

Ludwig A. Hothorn
Leibniz University of Hannover

DOI: 10.22237/jmasm/1162353780

Journal of Modern Applied Statistical Methods
November, 2006, Vol. 5, No. 2, 317-322

1538 – 9472/06/$95.00

Maximum Tests are Adaptive Permutation Tests

Markus Neuhäuser
Koblenz University of Applied Sciences

Ludwig A. Hothorn
Leibniz University of Hannover

In some areas, e.g., statistical genetics, it is common to apply a maximum test, where the maximum of several competing test statistics is used as a new statistic, and the permutation distribution of the maximum is used for inference. Here, it is shown that maximum tests are special cases of adaptive permutation tests. The 30-year old idea of adaptive statistical tests is more flexible than previously thought when permutation tests are used, and the selector statistic is calculated for every permutation. Because the independence between the selector and the test statistics is no longer needed, the test statistics themselves can be used as selectors. Then, the maximum tests fit into the concept of adaptive tests. In addition to the gained flexibility, maximum tests can be more powerful than classical adaptive tests.

Key words: Adaptive test, maximum test, maximum efficiency robust test, nonparametric statistics

Introduction

In this article the two-sample location problem is considered. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ denote 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 corresponding to populations 1 and 2, respectively. In the location-shift model the distribution functions are the same except perhaps for a change in their locations; that is, $F_1(t) = F_2(t - \theta)$ for every t. The null hypothesis is H_0: $\theta = 0$, whereas the alternative states $\theta \neq 0$.

Markus Neuhäuser is professor of statistics at the Koblenz University of Applied Sciences, RheinAhrCampus Remagen. His main research interests are nonparametric methods and application in life sciences. Ludwig Hothorn is professor for biostatistics at the natural science faculty of Leibniz University of Hannover in Germany. His research interests are multiple comparison procedure and order restricted inference with bio-medical applications. He is council member of the International Biometric Society.

Often, a normal assumption for F_1 and F_2 is not tenable. In this case, a nonparametric test can be performed using a linear rank statistic $T = \sum_{i=1}^{N} g(i)\, V_i$, where $g(i)$ are real valued scores, and $V_i = 1$ when the i-th smallest of the $N = n + m$ observations is from the first sample and $V_i = 0$ otherwise. There is a variety of different scores and, consequently, it is difficult for the practicing statistician to select a test statistic. A powerful test exists for every distribution, but the real distribution is usually *a priori* unknown and, consequently, one needs a test that has high relative power across the different possible distributions which may be difficult for small sample sizes.

In order to solve this dilemma Randles and Hogg (1973) and Hogg (1974) introduced adaptive statistical tests as a new dimension in distribution-free inference. The basic idea is that the value of a selector statistic decides which of some possible test statistics is applied. To be precise, the concept is based on the following lemma:

(i) Let $\mathcal{F}$ denote the class of distributions under consideration. Suppose that each of k tests $T_1, \ldots, T_k$ is distribution-free over $\mathcal{F}$, that is, $\Pr_{H_0}(T_i \in C_i) \leq \alpha$ for each $F \in \mathcal{F}$, $i = 1, \ldots, k$.

(ii) Let S be some statistic (called a selector statistic) that is, under H_0, independent of T_1, …, T_k for each $F \in \mathcal{F}$. Suppose we use S to decide which test T_i to conduct. Specifically, let M_S denote the set of all values of S with the following decomposition: $M_S = D_1 \cup D_2 \cup \cdots \cup D_k$, $D_h \cap D_j = \varnothing$ for $h \neq j$, so that $S \in D_i$ corresponds to the decision to use test T_i.

The overall testing procedure is then defined by: If $S \in D_i$ then reject H_0 if $T_i \in C_i$. This two-staged adaptive test is distribution-free under H_0 over the class $\mathcal{F}$, i.e., it maintains the level α for each $F \in \mathcal{F}$.

The proof of this lemma was given e.g. by Randles and Wolfe (1979, p. 388). Usually, tests based on ranks were used together with a selector statistic that depends on the combined ordered sample (Büning, 1991). The reason is that under the null hypothesis and in case of a continuous distribution, the rank vector is independent of the order statistics (Randles & Wolfe, 1979).

During the last 30 years several adaptive tests were introduced, not only for the two-sample location problem, but also for multi-sample problems and scale tests (Beier & Büning, 1997, Büning, 1991, 2000, 2002). Freidlin et al. (2003a) proposed a test where the selector and the test statistics are asymptotically uncorrelated only. Furthermore, the concept of adaptive tests was applied to parametric tests (Neuhäuser & Hothorn, 1997). However, this study focused on nonparametric two-sample location tests.

In 1995, Weerahandi wrote that, "until recently, most of the applications involving nonparametric tests were performed using asymptotic approximations" (p. 78). Therefore, most adaptive tests are constructed of asymptotic tests. Obviously, permutation tests (see e.g. Good, 2000) can be combined to an adaptive test, too, an example is the test introduced by O'Gorman (2001). The aim of this article is to show that permutation tests can offer a large flexibility to the concept of adaptive tests and that a maximum test is an adaptive permutation test.

The Combination of Permutation Tests

On the one hand, one can use the concept of adaptive tests in the classical way. That is, the selector is computed once and the chosen test is performed, now based on the permutation distribution. On the other hand, there is an alternative: the selector may be calculated for each permutation. In this case, a permutation test is carried out using the statistic $T_{P1} = \sum_{i=1}^{k} I(S \in D_i) T_i$, where $I(.)$ denotes the indicator function. With this statistic a permutation test can be performed, and neither the independence between S and the T_i nor the continuousness of the underlying distribution is necessary, in contrast to tests based on the lemma given in the introduction. Note that for a classical adaptive test the distributions have to be continuous for the independence between rank vector and order statistics. In practice, however, ties frequently occur in a variety of settings (see e.g. Coakley & Heise, 1996). Even when the underlying distribution is continuous rounding leads to ties. For example, reaction times may be measured with a time clock graduated in tenths or hundredths of a second. Moreover, it is an advantage of nonparametric rank tests that they can also be applied to ordered categorical data, but when continuousness has to be assumed, this advantage is lost.

Because the independence to the selector is no longer necessary one can use the (standardized) test statistics themselves as selectors. To be precise, one can perform a permutation test based on the statistic,

$$T_{P2} = \sum_{i=1}^{k} I\left(T_i = \max(T_1, \ldots, T_k)\right) I\left(T_i > T_j \ \forall\, j > i\right) T_i .$$

The second indicator function is needed because two statistics T_i and T_j (with $i \neq j$) could have an equal value for a given data set. Now, it is easy to see that $T_{P2} = \max(T_1, \ldots, T_k)$. Thus, a maximum test may be regarded as an adaptive permutation test.

The use of the maximum of several (standardized) statistics as a new test statistic is

common in rather different testing problems (e.g. Bretz & Hothorn, 2001, Chung & Fraser, 1958, Freidlin & Korn, 2002, Freidlin et al., 1999, 2002, 2003b; Gastwirth & Freidlin, 2000; Hirotsu, 1986; Marozzi, 2004a, 2004b, Neuhäuser & Hothorn, 1999, Neuhäuser et al., 2000, 2004, Zheng et al., 2002). The approach has the advantage that neither a selector statistic nor the specification of which test should be performed for which values of the selector is needed. Furthermore, a maximum test is possible for relatively small sample sizes. In contrast, a classical adaptive test needs a sample size of at least 20 per group to avoid too many misclassifications (Hill et al., 1988, Büning, 1991, p. 238).

Example

As an example, the class of all continuous and symmetric distributions is considered. In this case the following scores $g(i)$ may be useful:

Gastwirth test (short tails):

$$g(i)=\begin{cases} i-\dfrac{N+1}{4} & \text{for } i\le\dfrac{N+1}{4} \\ 0 & \text{for } \dfrac{N+1}{4}<i<\dfrac{3(N+1)}{4} \\ i-\dfrac{3(N+1)}{4} & \text{for } i\ge\dfrac{3(N+1)}{4} \end{cases}$$

Wilcoxon test (medium to long tails): $g(i)=i$

Median test (very long tails):

$$g(i)=\begin{cases} 1 & \text{for } i>\dfrac{N+1}{2} \\ 0 & \text{for } i\le\dfrac{N+1}{2}. \end{cases}$$

Above, in the parenthesis that type of distribution is indicated for which the test has high power (Büning, 1994). As a selector $\hat{Q}=\dfrac{\hat{U}_{0.05}-\hat{L}_{0.05}}{\hat{U}_{0.5}-\hat{L}_{0.5}}$ is chosen as a measure for tailweight (Hogg, 1974); $\hat{L}_\gamma$ and $\hat{U}_\gamma$ denote the average of the smallest and largest γN order statistics, respectively, in the combined sample. Fractional items are used when γN is not an integer. The longer the tails the greater is $\hat{Q}$. The adaptive test can be defined as follows:

If $\hat{Q}\le 2$, apply the Gastwirth test,

if $2<\hat{Q}\le 7$, apply the Wilcoxon test,

if $\hat{Q}>7$, apply the Median test.

The maximum test is constructed of the same three statistics. However, because the two-sided alternative $\theta \neq 0$ is considered, the maximum of the absolute values of the standardized statistics is used. Under H_0, expectation and variance of a linear rank statistic T are

$$E(T)=\frac{m}{N}\sum_{i=1}^{N}g(i)$$

and

$$Var(T)=\frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}g^2(i)-\left(\sum_{i=1}^{N}g(i)\right)^2\right]$$

(Büning & Trenkler, 1994, pp. 127-130). Let ST_G, ST_W, and ST_M denote the standardized statistic $\dfrac{T-E(T)}{\sqrt{Var(T)}}$ using Gastwirth, Wilcoxon, and Median scores, respectively. Then, the test statistic of the maximum test considered here is $T_{\max}=\max\left(\left|ST_G\right|,\left|ST_W\right|,\left|ST_M\right|\right)$. Inference is based on the permutation distribution of this maximum.

Table 1 shows type I error rates and powers of the univariate tests, the adaptive test and the maximum test. According to these results, the maximum test is less conservative than the other tests (for α = 0.05). This finding also holds for other maximum tests (see e.g.

Table 1. Type I error rates (simulated for the adaptive test) and simulated powers of different permutation tests, the adaptive test and the corresponding maximum test ($n = m = 10$, $\alpha = 0.05$, 10,000 simulation runs for each configuration)

Distribution	θ	Gastwirth test	Wilcoxon test	Median test	Adaptive test	Maximum test
Uniform on (0, 1)	0	0.042	0.043	0.023	0.042	0.049
	0.4	0.880	0.751	0.365	0.758	0.854
Standard normal	0	0.042	0.043	0.023	0.044	0.049
	1.5	0.755	0.854	0.625	0.834	0.835
Cauchy	0	0.042	0.043	0.023	0.039	0.049
	3	0.264	0.683	0.711	0.684	0.742

Neuhäuser et al., 2004). According to Table 1, the maximum test is more powerful than the adaptive test. Moreover, an important point is that the maximum test can be more powerful than the best univariate test, as it is here in the case of the Cauchy distribution. In contrast, the power of an adaptive test is always a weighted average of the powers of the univariate tests, i.e. the power of the adaptive test is always between the best and worst of the powers of the univariate tests.

Conclusion

The use of the maximum of several competing univariate statistics is quite common nowadays, especially in statistical genetics (see the references given above). Here, it is demonstrated that such maximum tests can be integrated within the 30-year old theory of adaptive tests. The distribution of the maximum can be determined by generating all possible permutations. The p-value of the resultant exact permutation test is the proportion of permutations yielding a statistic as supportive or more supportive of the alternative than the originally observed test statistic. When sample sizes are large and/or a multi-sample problem is considered, permutation tests can be performed using a simple random sample from all the possible permutations (see e.g. Good, 2000).

In some applications the correlation between the different statistics is known and the (asymptotic) distribution of the maximum is available, an example is a multiple contrast test where the maximum is multivariate *t*-distributed (see e.g. Hothorn et al., 1997, Genz & Bretz, 2002). However, to use a standard distribution is not generally a better way than using the permutation distribution. Instead, permutation tests may be preferable for several applications (Ludbrook & Dudley, 1998). Note that an approximation using the asymptotic distribution of a maximum statistic can be poor even when all univariate statistics are asymptotically normal (Freidlin & Korn, 2002).

Some decades ago, permutation tests were “almost never quick … seldom practical, and often … not even feasible” (Bradley, 1968, p. 84). Thus, maximum tests based on the permutation distribution could not be carried out. As an alternative method to univariate tests the concept of maximin efficiency robust tests (MERT) was introduced in order to obtain a single robust test statistic from a set of possible statistics (Gastwirth, 1966, 1970). The MERT idea is to maximize the minimum asymptotic efficiency over the possible tests.

Recently, MERTs were compared with the corresponding maximum tests (Freidlin et al., 1999, 2002, 2003b; Freidlin & Korn, 2002, Gastwirth & Freidlin, 2000, Neuhäuser &

Hothorn, 1999, Neuhäuser et al., 2004, Zheng et al., 2002). Such a comparison depends on the minimum correlation ρ^* between two of the univariate tests. When this correlation is small the maximum test is often preferable to the MERT, in particular in case of $\rho^* \leq 0.5$. For $\rho^* \geq 0.7$ there was, however, virtually no difference in their powers (Freidlin et al., 1999, 2002; Freidlin & Korn, 2002, Gastwirth & Freidlin, 2000, Neuhäuser et al., 2004, Zheng et al., 2002). Other linear combinations than the MERT are further alternatives to the maximum test, see e.g. Chi and Tsai (2001).

Instead to use the maximum test statistic one may use the minimum p-value (see e.g. Weichert & Hothorn, 2002). Such a procedure is essentially Tippett's combination, although the latter was introduced for independent tests. However, other combination functions could be used as well, see Pesarin (2001) for an overview of nonparametric combination methodology which is outside the scope of this article. However, irrespective of the method used to combine the different tests, it is often difficult to select them. This is, of course, also the case for the classical adaptive test. On the one hand, statistics with low correlation may be suitable because they focus on different areas of the alternative hypothesis. On the other hand, the penalty for using more than one statistic may also depend on the correlation as the comparison maximum test versus MERT does. Hence, there seems to be no general principle to select the test statistics, but, in contrast to adaptive tests, a maximum test neither needs a selector statistic nor the specification of which test should be performed for which values of the selector.

References

Beier, F. & Büning, H. (1997). An adaptive test against ordered alternatives. C*omputational Statistics and Data Analysis 25*, 441-452.

Bradley, J. (1968). *Distribution-free statistical tests*. Englewood Cliffs: Prentice-Hall.

Bretz, F. & Hothorn, L. A. (2001). Testing dose-response relationships with a priori unknown, possibly non-monotonic shapes. *Journal of Biopharmaceutical Statistics 11*, 193-207.

Büning, H. (1991). *Robuste und adaptive Tests*. Berlin: De Gruyter.

Büning, H. (1994). Robust and adaptive tests for the two-sample location problem. *OR Spektrum 16*, 33-39.

Büning, H. (2000). Robustness and power of parametric, nonparametric, robustified and adaptive tests – the multi-sample location problem. *Statistical Papers 41*, 381-407.

Büning, H. (2002). An adaptive distribution-free test for the general two-sample problem. *Computational Statistics 17*, 297-313.

Büning, H. & Trenkler, G. (1994). *Nichtparametrische statistische Methoden* (2nd Ed.). Berlin: De Gruyter.

Chi, Y. & Tsai, M.-H. (2001). Some versatile tests based on the simultaneous use of weighted logrank and weighted Kaplan-Meier statistics. *Communications in Statistics – Theory and Methods 30*, 743-759.

Chung, J. H. & Fraser, D. A. S. (1958). Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association 53*, 729-735.

Coakley, C. W. & Heise, M. A. (1996). Versions of the sign test in the presence of ties. *Biometrics 52*, 1242-1251.

Freidlin, B. & Korn, E. L. (2002). A testing procedure for survival data with few responders. *Statistics in Medicine 21*, 65-78.

Freidlin, B., Miao, W. & Gastwirth JL (2003a): On the use of the Shapiro-Wilk test in two-stage adaptive inference for paired data from moderate to very heavy tailed distributions. *Biometrical Journal 45*, 887-900.

Freidlin, B., Podgor, M. V. & Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics 55*, 883-886.

Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J.L. (2002): Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human Heredity 53*, 146-152.

Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J. L. (2003b). Efficiency robust tests for mapping quantitative trait loci using extremely discordant sib pairs. *Human Heredity 55*, 117-124.

Gastwirth, J. L. (1966). On robust procedures. *Journal of the American Statistical Association 61*, 929-948.

Gastwirth, J. L. (1970). On robust rank tests. In: Puri, M. L. (ed.) *Nonparametric techniques in statistical inference*. Cambridge University Press, Cambridge, pp. 89-109.

Gastwirth, J. L. & Freidlin, B. (2000). On power and efficiency robust linkage tests for affected sibs. *Annals of Human Genetics 64*, 443-453.

Genz, A. & Bretz, F. (2002). Comparison of methods for the computation of multivariate *t* probabilities. *Journal of Computational and Graphical Statistics 11*, 950-971.

Good, P. I. (2000): *Permutation tests*. Springer, New York (2nd edition).

Hill, N. J., Padmanabhan, A. R. & Puri, M. L. (1988). Adaptive nonparametric procedures and applications. *Applied Statistics 37*, 205-218.

Hirotsu, C. (1986). Cumulative chi-squared statistic or a toll for testing goodness of fit. *Biometrika 73*, 165-173.

Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association 69*, 909-927.

Hothorn, L. A., Neuhäuser, M. & Koch, H.-F. (1997). Analysis of randomized dose finding studies: Closure test modifications based on multiple contrast tests. *Biometrical Journal 39*, 467-479.

Ludbrook, J. & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *American Statistician 52*, 127-132.

Marozzi, M. (2004a). A bi-aspect nonparametric test for the two-sample location problem. *Computational Statistics and Data Analysis 44*, 639-648.

Marozzi, M. (2004b) A bi-aspect nonparametric test for the multi-sample location problem. *Computational Statistics and Data Analysis 46*, 81-92.

Neuhäuser, M., Büning, H. & Hothorn, L. A. (2004). Maximum test versus adaptive tests for the two-sample location problem. *Journal of Applied Statistics 31*, 215-227.

Neuhäuser, M. & Hothorn, L. A. (1997). Adaptive tests for trend. In: Kitsos, C. P., Edler, L. (eds.): *Industrial Statistics: Aims and Computational Aspects*, Physica-Verlag, Heidelberg, pp. 269-273.

Neuhäuser, M. & Hothorn, L. A. (1999). An exact Cochran-Armitage test for trend when dose-response shapes are a priori unknown. *Computational Statistics and Data Analysis 30*, 403-412.

Neuhäuser, M., Seidel, D., Hothorn, L. A. & Urfer, W. (2000). Robust trend tests with application to toxicology. *Environmental and Ecological Statistics 7*, 43-56.

O'Gorman, T.W. (2001). An adaptive permutation test procedure for several common tests of significance. *Computational Statistics and Data Analysis 35*, 335-350.

Pesarin, F. (2001): *Multivariate permutation tests*. Chichester: Wiley,

Randles, R. H. & Hogg, R. V. (1973). Adaptive distribution-free tests. *Communications in Statistics 2*, 337-356.

Randles, R. H. & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York, N.Y.: Wiley.

Weerahandi, S. (1995). *Exact statistical methods for data analysis*. New York, N.Y.: Springer.

Weichert, M. & Hothorn, L. A. (2002). Robust hybrid tests for the two-sample location problem. *Communications in Statistics – Simulation and Computation 31*, 175-187.

Zheng, G., Freidlin, B. & Gastwirth, J. L. (2002). Robust TDT-type candidate-gene association tests. *Annals of Human Genetics 66*, 145-155.