Psychology Science, Volume 46, 2004 (4), p. 549 - 554

The Robustness of two sample tests for Means
A Reply on von Eye's Comment

V. GUIARD[1], D. RASCH[2]

Abstract

The question, which test; the t-test or the Wilcoxon test for comparing means of two distributions has to be preferred must be answered by "the t-test". Both tests are robust against non-normality with a little advantage for the Wilcoxon test what the power concerns. But while the t-test is robust against variance heterogeneity or against discrete underlying distributions the Wilcoxon test is not. Both tests fail in robustness if observations within the samples are dependent. Such a situation should be avoided by carefully designing surveys and experiments.

Key words: robustness, non-normality, variance heterogeneity, discrete distributions, autocorrelation, t-test, Wilcoxon test

[1] Research Institute for the Biology of Farm Animals, Dummerstorf, Research unit Genetics & Biometry
[2] Institut für Mathematik, Alpen-Adria Universität Klagenfurt, Austria

1. Introductory Remarks on Robustness

In a paper (Rasch & Guiard, 2004) which was commented by von Eye (2005) we collected results of a systematic robustness research which was done by a statistical research group between 1978 and 1989.

Robustness of statistical procedures is mainly of interest for the *application* of such procedures. Robustness research should therefore be directed on often occurring or expected violations of assumptions needed for the derivation of statistical procedures.

As it was shown by analysing data from different areas of application, non-normality occurs relatively often. Another often occurring violation of an assumption is variance heterogeneity in two populations. In von Eye's comment the problem of dependency within "samples" was additionally discussed.

In the comment on our paper, von Eye considered only a small part of our results, namely the *t*-test for two independent samples. We therefore concentrate here on just this test and its non-parametric counterpart, the Wilcoxon Two-Sample test.

2. Robustness of the two sample t-test and the Wilcoxon test

Because the two sample t-test is one of the most applied statistical procedures, robustness results have been published long before we started our systematic research. Posten's (1978) results mentioned in our paper was obtained by the first *systematic* investigation over the Pearson system of probability distributions. These results where supplemented by those of Tuchscherer & Pierer (1985) considering not only non-normality but also variance heterogeneity both in the Fleishman system of probability distributions. In further investigations we also used the system of truncated normal distributions with different sets of first four moments each. Truncation of normal distributions often occurs after selection (for instance of pupils in the school system and of course in artificial selection in agriculture).

Exact and simulation results on robustness are valid only for the distributions used in the corresponding investigation. In earlier theoretical papers on which the paper of Rasch & Guiard (2004) is based we noticed this several times. If we fix the first four moments, there exist infinitely many distributions with just these four moments. In so far it is correct that robustness is parameter-specific – but this is well known.

Let us first repeat the assumptions for the two-sample t-test and the two-sample Wilcoxon test for testing the hypothesis $H_0 : \mu_1 = \mu_2$ of the equality of two means against a one- or two-sided alternative.

We assume two continuous distributions with existing first four moments and expectations $\mu_1; \mu_2$ and both variances equal to σ^2 (the third and fourth moment is needed for the simulation only). For the t-test we assume additionally normality of the two distributions and for the Wilcoxon tests that all existing moments higher than the second one are equal in both distributions; otherwise the test will not only compare the means.

We further assume that we draw two independent samples $(x_{11}, x_{12}, ..., x_{1n})$ and $(x_{21}, x_{22}, ..., x_{2n})$ from distribution 1 and 2 respectively.

The sizes of the two samples may be unequal; this is a problem only in combination with variance heterogeneity.

A sample in mathematical statistics is defined as a random vector of identically and independently distributed (i.i.d.) random variables. We assumed just such samples in our robustness research. But besides violations of the assumptions concerning the underlying distribution considered in our paper, von Eye draws our attention on the violation of the independence assumption within a random vector (which than is no longer a sample).

The main message of von Eye's remark is that our conclusion that there is no need for a non-parametric counterpart of the two sample *t*-test like the Wilcoxon test is wrong. We will discuss this point in the sequal.

Violations of assumptions can be:

- Non-normality
- Variance heterogeneity
- Correlated observations within samples
- Discrete (non continuous) underlying distributions
- Dependent samples

or any combination of them.

There is another branch of robustness research dealing with the case that not all elements of the random vector modelling the observations are identically distributed (see for instance Huber 1964, 1972). Huber considered the case that there are some outliers in the sample and just this means that not all elements of the vector are identically distributed. This research is really also of practical interest, because outliers can be expected in practical work. But we will drop this point here.

Let us consider the first three entries above only because tests for dependent samples exist and should in the fifth case be used.

2.1 Non-normality

That the *t*-test is extremely robust against non-normality is shown by Posten (1978) and our own investigations, it seems reasonable to recommend this test to users of statistical methods even if they feel that their distribution seems to be far from normal. Posten (1982) could show that in the case of non-normality, where both procedures (t-test and Wilcoxon test) are quite robust within the Pearson system there is in some situations a slightly higher power of the Wilcoxon test if the variances are equal. Nevertheless we see no need to replace the t-test by the Wilcoxon test, because nobody can be sure that variance homogeneity is always present.

2.2 Variance heterogeneity

Our simulations have further shown that the *t*-test in the case of equal sample sizes is robust against variance heterogeneity but the Wilcoxon test is not. This is also a theoretical result of Posten, Yeh & Owen (1982).

2.3 Correlated observations

Von Eye in his remark is now considering the situation that the elements of the random vector are not independent. We see in the case of carefully planning and performing a survey no danger that such a situation will occur. But let us assume, autocorrelation in samples is a sometimes occurring phenomenon in psychological research.

Von Eye has correctly shown that the *t*-test in such cases is not robust. But now von Eye is using just this fact to conclude that the Wilcoxon test is important and should in such cases be used in place of the *t*-test. This conclusion could be (and, as it can be seen below) wrong because he did not show that the Wilcoxon test is working well if autocorrelations are present.

Therefore we did some simulations with 10000 runs – this means that we simulated 10 000 times two samples and tested their equality by the Wilcoxon test and also by the *t*-test.

Let $(e_{11}, e_{12}, ..., e_{1n})$ and $(e_{21}, e_{22}, ..., e_{2n})$ be two i.i.d. independent random vectors of size n = 100 from an N(0;1)- distribution. For $i = 1,2$ calculate $x_{i1} = e_{i1}$ and $x_{ij} = \rho x_{ij-1} + \sqrt{1-\rho^2}\, e_{ij};\ j = 1,...,100$ respectively. Then the components of both vectors $(x_{11}, x_{12}, ..., x_{1n})$ and $(x_{21}, x_{22}, ..., x_{2n})$ are N(0;1)- distributed having an autocorrelation ρ.

These vectors will now be used in a usual Wilcoxon test and in a two-sample *t*-test for the null hypothesis of the equality of the two expectations (variance assumed to be unknown but equal) against a two-sided alternative neglecting the autocorrelation.

In the simulations we always used a nominal α = 0,05 and equal means (w.l.o.g. = 0) and also w.l.o.g. variances equal to 1 in both samples. That means, the null hypothesis is assumed to be true. The results for the Wilcoxon test and the t-test presented in Table 1 give the relative frequency of rejecting the null hypothesis (called actual α) for the 10 000 pairs of samples.

The t-test as well as the Wilcoxon test are both theoretically derived for samples and not for random vectors with autocorrelation. Because contrary to normality this assumption is intrinsic (both tests are highly sensitive), this means that correlated observations must be avoided by a careful data recording in a carefully planned survey or experiment.

Table 1:
Relative Frequency (actual α) of rejecting the null hypothesis by the Wilcoxon test and the *t*- test for different autocorrelation coefficients ρ in both samples

ρ	actual α Wilcoxon test	actual α t-test
-0.8	0.0000	0.0000
-0.4	0.0044	0.0036
0	0.0478	0.0479
0.4	0.1950	0.2020
0.8	0.5053	0.5193

2.4 Discrete underlying distributions

We like to make a further remark in this connection. To denote a test as *distribution free* as done in many papers or books does not make any sense for us.

Neither does the test statistic of a test from this group follow no probability distribution nor is the application of such a test free of assumptions on the underlying distribution but just this is suggested by the term "distribution-free". In the case of the Wilcoxon test the distributional assumption is: We need two independent samples (i.i.d. vectors) from continuous distributions.

Summarising we still think there are more disadvantages then advantages in using the Wilcoxon test in place of the t-test.

References

1. Huber, P.J. (1964): Robust estimation of a location parameter. Ann. Math. Statist. 35, 73 – 101.
2. Huber, P.J. (1972): Robust Statistics: A review. Ann. Math. Statist. 43, 1041 - 1067.
3. Posten, H. O. (1982): Two-sample Wilcoxon-Power over the Pearson system and comparison with the t-test. J. of Statistical Computation and Simulation 16, 1-18.
4. Posten, H.O (1978): The Robustness of the Two-Sample T-test over the Pearson System. J. of Statistical Computation and Simulation, 6, 295-311.
5. Posten, H.O., Yeh, H.C. and Owen, D.B. (1982): Robustness of the two-sample t-test under violations of the homogeneity of variance assumptions. Communications in Statistics: Theory and Methods 11, 109-126.
6. Rasch, D. and Guiard, V. (2004): The Robustness of parametric statistical methods. Psychology Science, vol. 46 (2), p. 175-208.
7. Rasch, D., Häusler, J., Kubinger, K.-D., Herrendörfer, G. and Guiard, V. (2005): How Robust are Non-parametric significance tests in cases of non-continuous distributions? Paper to be presented at the 5th St. Petersburg Conference on Simulation, St. Petersburg, June 2005, Proceedings will be published in a special issue of the Journal of Statistical Planning and Inference.
8. Tuchscherer, A. and Pierer, H. (1985): Simulationsuntersuchungen zur Robustheit verschiedener Verfahren zum Mittelwertvergleich im Zweistichprobenproblem (Simulationsergebnisse). In: Rudolph, P.E. (ed.) (1985): Robustheit V - Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik, Heft 15, Dummerstorf-Rostock, S. 1-42.
9. von Eye, A. (2005): Robustness is parameter-specific. A comment on Rasch and Guiard's robustness study. Psychology Science, vol. 46 (2004).