Psychology Science, Volume 46, 2004 (2), p.

The Robustness of Parametric Statistical Methods

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

1. Abstract

In psychological research sometimes non-parametric procedures are in use in cases, where the corresponding parametric procedure is preferable. This is mainly due to the fact that we pay too much attention to the possible violation of the normality assumption which is needed to derive the exact distribution of the statistic used in the parametric approach.

An example is the t-test and its non-parametric counterpart, the Wilcoxon (Mann-Whitney) test. The Wilcoxon test compares the two distributions and may lead to significance even if the means are equal due to the fact that higher moments in the two populations differ. On the other hand the t-test is so robust against non-normality that there is nearly no need to use the Wilcoxon test. In this paper results of a systematic research of the robustness of statistical procedures against non-normality are presented. These results have been obtained in a research group in Dummerstorf (near Rostock) some years ago and have not been published systematically until now. Most of the results are based on extensive simulation experiments with 10 000 runs each. But there are also some exact mathematically derived results for very extreme deviations from normality (two- and three-point distributions). Generally the results are such that in most practical cases the parametric approach for inferences about means is so robust that it can be recommended in nearly all applications.

Key words: - missing -

[1] D. Rasch, BioMath – Institute of Applied Mathematical Statistics in Biology and Medicine Ltd., Rostock

[2] V. Guiard, Research Institute for the Biology, of Farm Animals Dummerstorf, Research Unit Genetics & Biometry

2. Definition of Robustness against Non-Normality

In Herrendörfer, G. (ed.) (1980) and Bock (1982) several robustness concepts like those of Box and Tiao (1964), Huber (1964) and Zielinsky (1977) have been discussed. The concept of ε–robustness was introduced which was used in the robustness research presented in this paper. This concept is here introduced for the construction of confidence intervals in a simplified way.

Definition 1

Let $\underline{d}_a^{\,1}$ [3] be a confidence estimation based on an experimental design V_n of size n concerning a parameter θ of a class G of distributions with (nominal) confidence coefficient 1 - α (0<α<1) in G. For an element $h \in H \supset G$ of a class H of distributions which contains G we denote by $1-\alpha(V_n, h)$ the actual confidence coefficient of $\underline{d}_\alpha$. Then we call $\underline{d}_\alpha$ ε–robust in H if α and $\alpha(V_n, h)$ deviate from each other by less than ε for any element in H (sometimes we also say by less than 100ε% and than speak about 100ε% - robustness).

If we are only interested in not too large $\alpha(V_n, h)$-values (conservative procedures are not considered as non-robust) only, we call $\underline{d}_\alpha$ ε*–robust in H if $\alpha(V_n, h)$ does not exceed α by more than ε*.

In an analogue way the ε– and ε*–robustness of tests or selection procedures can be defined. Due to the fact that a test for testing a null hypothesis $H_0 : \theta = \theta_0$ can be performed by accepting H_0 if θ_0 is inside the confidence interval and reject it otherwise, definition 1 includes the robustness of a test concerning the significance level.

In this paper we use for G the family of univariate normal $(N(\mu,\sigma^2)–)$ distributions and for H the Fleishman system of distributions and/or truncated normal distributions.

Definition 2

A distribution belongs to the Fleishman system (Fleishman, 1978) if its first four moments exist and if it is the distribution of the transform

$$\underline{y} = a + b\underline{x} + c\underline{x}^2 + d\underline{x}^3 \tag{1.1}$$

where $\underline{x}$ is a standard normal random variable (with mean 0 and variance 1).

By a proper choice of the coefficients a, b, c and d the random variable $\underline{y}$ will have any quadruple of first four moments $(\mu,\sigma^2,\gamma_1,\gamma_2)$. By γ_1 and γ_2 we denote the skewness (standardized third moment) and the kurtosis (standardized fourth moment) of any distribution, respectively. For instance any normal distribution (ie. any element of G) with mean μ and variance σ^2 can be represented as a member of the Fleishman system by choosing $a = \mu$, $b = \sigma$ and $c = d = 0$. This shows that we really have $H \supset G$ as demanded in definition 1.

[3] Random variables are underlined

Experimenters know that they nearly never really meet observation from normal distributions. Sometimes they know, that the underlying distribution is skew, truncated or differs in another way from being normal. In Herrendörfer, G. (ed.) (1980) 144 characters from experiments in medicine, plant and animal breeding have been investigated. For each of the characters from a sample of size r the empirical skewness g_1 and kurtosis g_2 have been calculated. The values of r have been 64 for 2 characters, 128 for 11 characters, between 192 and 438 for 38 characters and for all the other characters between 656 and 5466. This means that the point estimates for the skewness and the kurtosis in the most cases are not far from the unknown parameters γ_1 and γ_2 respectively.

It is known (see Herrendörfer, G. (ed.) (1980), page 26) that all probability distributions (with existing fourth order moment) fulfil the inequality

$$\gamma_2 \geq \gamma_1^2 - 2 \tag{1.2}$$

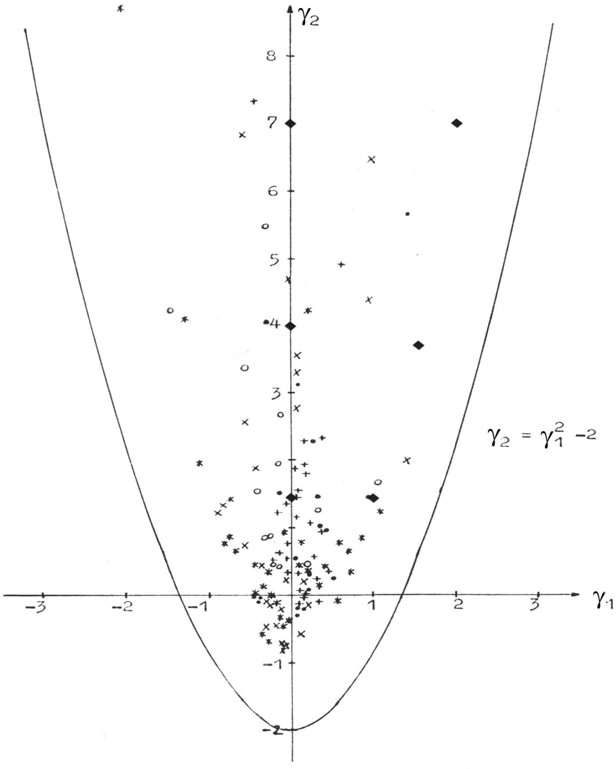

Figure 1:
Values of empirical skewness g_1 and kurtosis g_2 of 144 characters in a (γ_1,γ_2)-plane, by ♦ the parameters (γ_1,γ_2) of 6 distributions of the Fleishman system are denoted

Empirical distribution fulfil an analogue equation in the estimated skewness g_1 and kurtosis g_2 respectively

$$g_2 \geq g_1^2 - 2 \tag{1.3}$$

The equality sign in (2) or (3) defines a parabola in the $(\gamma_1, \gamma_2)-$ plane $\{(g_1, g_2)-$ plane$\}$.

In Figure 1 the position of the (g_1, g_2) –values calculated for the 144 characters in that parabola are shown.

Therefore we selected seven (γ_1, γ_2) –values in that parabola for the robustness investigations reported in this paper. The values together with the coefficient *a*, *b, c* and *d* of the elements in the Fleishman system for the case $\mu = 0$ (what means $a = - c$) and $\sigma = 1$ (what means $b^2 + 6bd + 2c^2 + 15d^2 = 1$) are given in Table 1. they are marked in Figure 1 by a♦.

Table 1:
(γ_1, γ_2) –values and the corresponding coefficients in (2)

No of distribution	γ_1	γ_2	*c* = -*a*	*b*	*d*
1	0	0	0	1	0
2	0	1.5	0	0.866993269415	0.042522484238
3	0	3.75	0	0.748020807992	0.077872716101
4	0	7	0	0.630446727840	0.110696742040
5	1	1.5	0.163194276264	0.953076897706	0.006597369744
6	1.5	3.75	0.221027621012	0.865882603523	0.027220699158
7	2	7	0.260022598940	0.761585274860	0.053072273491

The values of *b, c* and *d* in Table 1 are taken from Table A3 in Nürnberg (1982), some of them can be found already in Fleishman (1972).

1. Simulation Experiments in Robustness Research

In the simulation experiments (as also in the statistical analysis of experimental results) we mainly use linear statistical models with an error term. We investigate the behaviour of statistical procedures like confidence estimation, hypotheses testing and selection procedures which have been derived (and thus are exact) under the assumption of normal error terms. The question is: „What happens with the properties of these procedures if we replace the normal error by an error term following a Fleishman (or a truncated normal) distribution with non-zero γ_1 and γ_2 ?"

There are only very few analytical results to answer this question for sample sizes 2 and 3 which are useless for practical purposes. For very extreme (two- and three-point distributions) exact answers have been given by Herrendörfer (1980) and Herrendörfer and Rasch (1981). The results in our paper are based on simulation experiments.

By simulation we mean that we use random numbers to simulate the observations of an artificial experiment for a given underlying model. In the one-sample case we denote the size of the artificial experiment by *n*.

We use the following example to demonstrate the basic ideas.

Example 1

A sample (that is a vector of identically and independently distributed (abbr. i.i.d.) random variables) of size *n* is drawn from an – distribution. A (realized) confidence interval is calculated from the realization of the random sample (the observations) if σ - as usually- is unknown by:

$$[\overline{y} - t(n-1;1-\frac{\alpha}{2})\frac{s}{\sqrt{n}};\overline{y} + t(n-1;1-\frac{\alpha}{2})\frac{s}{\sqrt{n}}] \qquad (1.4)$$

where the sample mean $\overline{y}$ is the point estimate of μ , the sample standard deviation s is the point estimate of σ and $t(n-1;1-\frac{\alpha}{2})$ is the $(1-\frac{\alpha}{2})$ quantile of the central *t*-distribution with *n* – 1 *d.f.*. The corresponding random interval (with random bounds as functions of the sample) covers the unknown μ with probability $(1-\frac{a}{2})$ as long as the sample really stems from a normal distribution. Amongst all intervals with such a property the interval (4) is optimal what means it has the smallest expected width. Non-parametric intervals have a larger expected width if the normal assumption is true. In this case the linear model for the *n* elements $\underline{y}_i$ of the random sample has the simple form

$$y_{\underline{i}} = \mu + e_{\underline{i}}, i = 1,...,n \qquad (1.5)$$

The experimental design V_n in (1) is simply given by the sample size *n* (there is no structure in the experiment). The robustness against truncation (occurring in any selecting process like in schools or in agricultural artificial selection) of normal distributions of this confidence estimation (and the corresponding test) is investigated in simulation experiment 1 below.

We use besides the t-quantiles other abbreviations (for quantiles) as given in the following Table.

Table 2:
Abbreviations used in this paper

Term	Abbreviation
Degrees of freedom	d.f.
Identically and independently distributed	i.i.d.
P-quantile (percentile) of the *N*(0;1)distribution	*u*(*P*)
P-quantile (percentile) of the *t*-distribution with f d.f.	*t*(*f; P*)
P-quantile (percentile) of the *F*-distribution with d.f.	$F(f_1, f_2; P)$
P-quantile (percentile) of the χ^2 -distribution with f d.f.	*CQ*(*f; P*)

3.1 Pseudo-Random Number Generators

Because we cannot generate random numbers by means of a computer program, in simulation experiments so-called pseudo-random numbers (PRN) are used. In Herrendörfer, G. (ed.) (1980) and by FEIGE et. al. (1985) 18 generators of PRN uniformely distributed in (0, 1) have been tested against several properties which may negatively influence the simulation results (like periodicity or correlation).

Finally we used a generator which was developed by Teuscher (1979) according to an idea of Mac Laren and Marsaglia (1965). This generator is a combination of the following two generators:

$$x_k = 2^{10} x_{k-1} \bmod 1049339 \tag{1.6}$$

and

$$y_i = 8323 y_{i-1} \bmod 2^{28}, y_0 = 1$$

Within the initialisation phase the values $x_1, ..., x_{128}$ will be generated. In order to calculate the random values $Z_i, i = 1...,$ we calculate the index I with $1 \le I \le 128$, being the integer part of $1 + y_i / 2^{21}$. Then we select x_I from the x-values above and calculate $Z_i = x_I / 1049339$. Within the vector $(x_1, ..., x_{128})$ the value x_I will be substituted by the next value of the x_k -values, i.e. x_{128+i}.

The *Z*-values generated from (6) are PRN uniformly distributed in the interval (0,1). They have to be transformed into PRN with a standard normal distribution and then by (1) in PRN of the Fleishman system. FEIGE et. al. (1985) tested 5 different transformations of the *Z*- values into standard normal PRN *u* and proposed the following for further use. At first we define for $Z < 0.5$

$$z = \sqrt{-2 \ln Z} \text{ and } c = 1$$

and for $Z \geq 0.5$

$$z = \sqrt{-2\ln(1-Z)} \text{ and } c = -1$$

Then with coefficients given by Odeh and Evans (1974) (see also Feige et. al., 1985, page 31) calculate

$$u \ = c \cdot [\ z + \frac{\left\{\left[\left(p_4 \cdot z + p_3\right)z + p_2\right]z + p_1\right\}z + p_0}{\left\{\left[\left(q_4 \cdot z + q_3\right)z + q_2\right]z + q_1\right\}z + q_0} \ ].$$

These N(0;1)-PRN have been used directly (after truncation) in simulation experiment 1 and have been transformed by (1) into Fleishman variates for the other experiments.

3.2 Planning the Size and the Scope of the Simulation Experiments

A simulation experiment has to be planned in the same way as experiments in psychology (see Rasch, 2003) or other fields. The aim of the simulation experiments discussed in this papers is to evaluate probabilities as the significance level of a test or the confidence coefficient of an interval estimation or the probability of a correct selection in selection procedures. More specific are we interested to find out, whether under non-normality the nominal probability differs by more than an ε from the actual value. We restrict ourselves in the present paper on a significance level of $\alpha = 0.05$, a confidence coefficient 1-α =0.95 and a probability of a correct selection $\beta = 0.05$. For ε^* in definition 1 we choose 20% of α and this is 0.01. That means on the basis of the simulation of N samples of size n each we will test the pair of hypotheses (for selection problems α must be replaced by β) :

$$H_0 : \alpha = 0.05$$
$$H_A : \alpha > 0.05$$

(if α is smaller than 0.05 we do not exceed the nominal significance level).

To determine the number of simulated samples we proceed like an experimenter in life science planning a real life experiment. We follow the principle explained in paragraph 4.4 in Rasch (2003) and fix the significance level of this test at α*=0.01 and the value of the power function at α + d = 0.05 + 0.006 = 0.056 at 0.85. To determine the number of simulated samples we edit the constants above into the CADEMO module MEANS, branch „Comparing a probability with a constant" and receive the result 9894. Therefore 10 000 runs were used in each of the simulation experiments below.

3.3 The Layout of the Simulation Experiments for several problems

We report simulation experiments for the investigation of the robustness of tests, confidence estimation and selection procedures. Each of the statistical inference methods was perfomed 10 000 times for each case included in the research program. Then the number R of negative results has been observed and used to calculate the observed error rate by $\hat{\alpha}$ as an estimate of the actual risk α . The elements h of the Fleishman system are given in Table 1. In paragraph 4 we report only the results for the nominal risk $\alpha = 0.05$ but in Herrendörfer, G. (ed.) (1980), Guiard, V. (ed.) (1981), Rasch, D. und G. Herrendörfer (ed.) (1982), Rasch, D. (ed.) (1985), Rudolph, P.E. (ed.) (1985 a), and Guiard, V. und Rasch, D. (ed.) (1987) results for $\alpha = 0.01$ and $\alpha = 0.1$ are also given. The normal distribution was included on one hand to check the correctness of the programs. On the other hand – like in the investigation of asymptotic tests – the results for the normal have been of interest too.

We demonstrate this in the following example which is also used to show how more complicated cases than that in example 1 can be handled.

Example 2

Let us consider the non-linear regression model with unknown parameters α, β and γ.

$$\underline{y}_i = \alpha + \beta e^{\gamma x_i} + \underline{e}_i \qquad i = 1,...,n) \tag{1.7}$$

with i.i.d. random errors $\underline{e}_i$. Here V_n is given not only by the sample size n but by the allocation of the n x-values. We define a design with k different x-values $x_1, x_2,..., x_k$ which occur with the frequencies $n_1, n_2,..., n_k$, respectively, by

$$V_n = \begin{pmatrix} x_1 & x_2 & ... & x_k & ,k \geq 3, \sum_{j=1}^{k} n_j = n \\ n_1 & n_2 & ... & n_k & \end{pmatrix} \tag{1.8}$$

The distribution of the least squares estimators of the three parameters is unknown for finite samples even if the error terms are normally distributed. What is known is the asymptotic distribution on which statistical tests can be based (see RASCH, 1995).

Let us now describe the simulation experiments whose results will be reported in paragraph 4. The first 6 experiments deal with tests for means and variances. But the results are also useful for the construction of confidence intervals. Between a confidence interval with coefficient 1-α and a test with a significance level α a one-to-one correspondence exist. As long as the hypothesis value μ_0 lies inside the interval, H_0 is accepted and rejected otherwise. On the other hand, all accepted μ_0– values are inside the interval and all rejected outside. We report the tests here because they give additional information about the power (or the risk of the second kind).

Simulation Experiment 1 – One-sample tests for the mean

This experiment has been planned and performed by TEUSCHER (1985).
In this experiment three test have been compared to test the hypothesis $H_0 : \mu = \mu_0$ against:

a) $H^{+}{}_{A} : \mu > \mu_0$,

b) $H^{-}{}_{A} : \mu < \mu_0$ and

c) $H_A : \mu \neq \mu_0$.

with a sample of size *n.*

If the variance σ^2 of the assumed normal distribution is known, the u-test was used as follows:

Test 1

Calculate $u = z = \frac{\overline{y} - \mu_0}{\sigma}\sqrt{n}$ and reject H_0 :

a) if $u > u(1\text{-}\alpha),$

b) if $u < - u(1\text{-}\alpha),$

c) if $u > u\left(1-\frac{\alpha}{2}\right)$.

For unknown σ^2 two tests have been investigated:

Test 2

The one-sample t-test:

Calculate $\mathrm{t} = \frac{\overline{\mathrm{y}} - \mu_0}{\mathrm{s}}\sqrt{n}$ and reject H_0 :

a) if $t > t(n\text{-}1;1\text{-}\alpha),$

b) if $t <\text{-}\ t(n\text{-}1;1\text{-}\alpha),$

c) if $t > t\left(n-1;1-\frac{\alpha}{2}\right)$.

Test 3

The modified Johnson-test based on JOHNSON N, J. (1978)
Calculate with an estimate g_1 of the skewness γ_1:

$$t_J = \left(\bar{y} - \mu_0 + \frac{g_1}{6s^2 n} + \frac{g_1}{3s^4 \sqrt{n}} \left(\bar{y} - \mu_0 \right)^2 \right) \frac{\sqrt{n}}{s}$$ and reject H_0:

a) if $t_J > t(n\text{-}1;1\text{-}\alpha)$,
b) if $t_J < - t(n\text{-}1;\ 1\text{-}\alpha)$,
c) if $t_J > t\left(n-1;1-\frac{\alpha}{2}\right)$.

This test corrects the test statistic by the empirical skewness g_1. There is no correction for kurtosis.

In this experiment the class *H* in definition 1 was not the Fleishman system but the system of truncated normal distributions. If a standard normal distribution is truncated at *v* (lower truncation point) and *w* (upper truncation point) then the truncated distribution is skew if $w \neq -v$ and it has a non-zero kurtosis for most (*v, w*).

Teuscher used the distributions 1-7 of Table 3 and sample sizes of *n* = 5, 10, 30 and 50.

Table 3:
Values of *v* and *w* and of the skewness and kurtosis of the simulated distributions

No of distribution	v	w	γ_1	γ_2
1	-6	6	0	-0.06
2	-2.2551	2.2551	0	-0.5
3	-1.183	1.183	0	-1.0013
4	-1	6	0.5918	0.0014
5	-0.29	2.743	0.75	0
6	-0.7041	2.2778	0.5	-0.5
7	1	1.52513	1.3159	1.998
8	0.5	3	1.0057	0.5915
9	-1.5	3	0.348	-0.3481
10	2.85	4.71	1.505	3.75

Simulation Experiment 2 – one-sample sequential tests for the mean

RASCH (1985a, b) investigated the behaviour of three sequential one – sample tests for non-normal situations. These tests are sequential tests analogue to those of experiment 1. The null hypothesis:

$$H_0 : \mu = \mu_0$$

was tested against:

$$H_A : \left(\mu - \mu_0 \right)^2 = \sigma^2 d^2$$

The test statistics t_1, t_2, and t_3 of the three tests can be found in the reference, test 1 was based on the original WALD-test.

For each of the tests with test statistic t (t_1, t_2, and t_3) and fixed risk α of the first and β of the second kind, respectively, the decision rule was as follows:

Accept H_0 if

$$t_1, t_2 \le b = \ln B \text{ or if } t_3 \le B$$

Reject H_0 if

$$t_1, t_2 \ge a = \ln A \text{ or if } t_3 \ge A$$

and take a further observation otherwise. Here we put

$$A = \frac{1-\beta}{\alpha}; B = \frac{\beta}{1-\alpha} \tag{1.9}$$

with (α, β) = (0.05; 0.1); (0.05, 0.2); (0.1; 0.2) and (0.1; 0.5); d =0.6; 1 and 1.6.

In the simulation a normal distribution (no 1 in Table 4), the distributions 3-6 of Table 1 (number 5-8 in Table 4) and 8-10 of Table 3 (2-4 in Table4) have been used. Further the average sample size of the 10 000 replication has been calculated for each situation.

Table 4:
Simulated distributions and their skewness and kurtosis

No of distribution	Type of distribution	γ_1	γ_2
1	Normal	0	0
2	Truncated normal	1.0057	0.5915
3	Truncated normal	0.348	-0.3481
4	Truncated normal	1.505	3.75
5	Fleishman system	0	3.75
6	Fleishman system	0	7
7	Fleishman system	1	1.5
8	Fleishman system	1.5	3.75

Simulation Experiment 3 and Posten's experiment – Tests for comparing two means

Comparing the means of two distributions may be one of the most frequent applications of statistics in many research fields. We discuss here the independent samples case, this means, the samples of size n_1 and n_2 have been drawn independently from two populations with parameters $\mu_1, \sigma_1^2, \gamma_{11}, \gamma_{12}$ and $\mu_2, \sigma_2^2, \gamma_{21}, \gamma_{22}$, respectively. The dependent samples case as it occurs if the same persons are investigated before and after a treatment can be reduced to the one-sample case discussed in experiment 1.

Our purpose therefore is to take two independent random samples ($y_{11}, ..., y_{1n_1}$) and ($y_{21}, ..., y_{2n_2}$) of sizes n_1 and n_2 from the two populations in order to test the null hypothesis

$H_o : \mu_1 = \mu_2$

against one of the following one- or two-sided alternative hypotheses

a) a)$H_A : \mu_1 > \mu_2$
b) b)$H_A : \mu_1 < \mu_2$
c) c)$H_A : \mu_1 \neq \mu_2$

The following tests were investigated:

Test 1 – two sample t-test:

Calculate

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_{1.})^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_{2.})^2}{n_1 + n_2 - 2} \qquad (1.10)$$

and the test statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} .$$

and reject H_0

in case a) if $t > t(n_1+n_2–2; 1–\alpha)$,
in case b) if $t<–t(n_1+n_2–2; 1–\alpha)$ and
in case c) if $|t| > t(n_1+n_2–2; 1–\frac{\alpha}{2})$,

and otherwise accept H_0.

Test 2 – Welch test

WELCH (1947) proposed the following test statistic:

$$t_W = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1.11}$$

where s_1^2 and s_2^2 are the two sample variances. Taking f^* as

$$f^* = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{(n_1 - 1)n_1^2} + \frac{s_2^4}{(n_2 - 1)n_2^2}} \tag{1.12}$$

reject H_0:

in case a), if $t > t(f^*; 1\text{-}\alpha)$,
in case b) if $t < - t(f^*; 1\text{-}\alpha)$ and
in case c), if $|t| > t(f^*; 1\text{-} \alpha/2)$

and accept it otherwise (the three cases correspond to those in test 1).

Test 3 Wilcoxon (Mann-Whitney) test

WILCOXON (1945) and later MANN and WHITNEY (1947) proposed a two-sample test based on the ranks of the observations. This test is not based on the normal assumption, in its exact form only two continuous distributions with all moments existing are assumed, we call it the Wilcoxon test. As it can be seen from the title of the second paper, this test is testing whether one of the underlying random variables is stochastically larger than the other. It can be used fo testing the equality of means under additional assumptions: The null hypothesis tested by the Wilcoxon test corresponds with our null hypothesis $H_o : \mu_1 = \mu_2$ if and only if all higher moments of the two populations are equal. Otherwise a rejection of the Wilcoxon hypothesis says few about the rejection of $H_o : \mu_1 = \mu_2$. The test runs as follows:

Calculate:

$$d_{ij} = \begin{cases} 1 & \text{, if} \quad y_{1i} > y_{2j} \\ 0 & \text{, if} \quad y_{1i} < y_{2j} \end{cases} \tag{1.13}$$

and then

$$W = W_{12} = \frac{n_1\left(n_1+1\right)}{2} + \sum_{i=1}^{n_1}\sum_{j=1}^{n_j} d_{ij} \ . \tag{1.14}$$

Reject H_o

in case a), if $W > W\left(n_1, n_2; 1-\alpha\right)$

in case b) if $W < W\left(n_1, n_2; \alpha\right)$ and

in case c), if either $W < W\left(n_1, n_2; \frac{\alpha}{2}\right)$ or $W > W\left(n_1, n_2; 1-\frac{\alpha}{2}\right)$

and accept it otherwise (the three cases correspond to those in test 1). Extensive tables of the quantiles $W\left(n_1, n_2; \alpha\right)$ are given by VERDOOREN (1963).

Test 4 – Range test of Lord

LORD (1947) proposed the following test for small samples:

Calculate the ranges (maximum minus minimum sample value) w_1 and w_2 of the two samples and then

$$T_{Lord} = \frac{\bar{y}_1 - \bar{y}_2}{w_1 + w_2}$$

and reject H_o

in case a), if $T_{Lord} > \tau\left(n_1, n_2; 1-\alpha\right)$

in case b) if $T_{Lord} < \tau\left(n_1, n_2; \alpha\right)$, and

in case c), if either $T_{Lord} < \tau\left(n_1, n_2; \frac{\alpha}{2}\right)$ or $T_{Lord} > \tau\left(n_1, n_2; 1-\frac{\alpha}{2}\right)$

and accept it otherwise (the three cases correspond to those in test 1). Tables of the quantiles $\tau\left(n_1, n_2; P\right)$ are given in Lord's paper.

The following tests are due to TIKU (1980).

Test 5 Tiku's T-test

A modified maximum likelihood test for censored samples described in TUCHSCHERER (1985, page 161 as test T_{12}).

Test 6 Tiku's T_c-test

A Welch type modification of test 5 described in TUCHSCHERER (1985, page 161 as test T_{13}).

The situation of experiment 3 concerning test 1 and test 3 was one of the few studies already investigated with a sufficient large number of runs before our robustness research started. These simulations were done by Posten (1985) using 40 000 runs for test 3 and100 000 runs for test 1.

Posten used the union of Pearson system distributions with the normal distributions as the class H of distributions in definition 1. He simulated distributions in the Pearson system in a grid of (γ_1 , γ_2) – values with γ_1 = 0(0.4)2 and γ_2 = - 1.6(0.4)4.8 and sample sizes of 5(5)30 equal in both samples.

Posten stated in his 1985 paper:

„*It would seem , therefore, that further studies of the effects of variance heterogeneity on the two tests would be needed over an extensive practical family of non-normal distributions before a single procedure might be specified as a somewhat general choice for the two-sample location problem"*

This was the reason that Tuchscherer started a further experiment (Tuchscherer 1985, Tuchscherer and PiererI 1985).

The authors of the second paper used the 7 Fleishman distributions of Table 1, first kind risks of 0.01; 0.05 and 0.1; the test 1-6 and 7 further tests and a ratio of the two variances of 0.5; 1; and 1.5 and a sample size of 5 in population 1 and of 5 and 10 in population 2.

In Guiard and Rasch (ed.) (1987) for the tests 1 – 4 further results can be found for truncated normal distributions with (γ_1 , γ_2) – values: (1.0057; 0.5915); (0.348; -0.3488) and (1.505; 3.75).

Simulation Experiment 4 – Sequential tests for comparing two means

FRICK (1985) investigated the robustness of two tests against non-normality. Besides the normal distribution the Fleishman distributions in Table 4 and two truncated distribution have been used. In some of the simulations different distributions have been generated in the two samples. The null hypothesis.

$$H_0 : \mu_1 = \mu_2$$

was tested against:

$$H_A : \frac{(\mu_1 - \mu_{20})^2}{\sigma^2} = d^2$$

Test 1 – Hajnal's test

HAJNAL (1961) proposed the following procedure:

Calculate

$$Q = \exp\left(\frac{-d^2}{2K}\right) H\left\{\frac{f+1}{2}, \frac{1}{2}, \frac{d^2t^2}{2K\left(f+t^2\right)}\right\}$$

with

$$K^2 = \frac{1}{n_1} + \frac{1}{n_2}, f = n_1 + n_2 - 2, s_p^2 \text{ ,}$$

from (10),

$$t = \frac{\overline{y}_1 - \overline{y}_2}{s_P \cdot K}$$

and the confluent hyper-geometric function *H.*

With *A* and *B* as in experiment 2, we accept H_0 if $Q<B$ and reject it, if $Q>A$. Otherwise experimentation continues.

Test 2 – a Welch modification of test 1

Reed and Frantz (1979) proposed to proceed as follows:
In both populations observations could be taken with different probabilities

$$\pi_1, \pi_2;\ 0 < \pi_i < 1; \left(i = 1, 2\right); \pi_1 + \pi_2 = 1 .$$

Replace *Q* in test 1 by:

$$Q_W = \exp\left(\frac{-d_W{}^2}{2K}\right) H\left\{\frac{f_W+1}{2}, \frac{1}{2}, \frac{d_W{}^2 t_W{}^2}{2K\left(f_W + t_W{}^2\right)}\right\}$$

with $f_w = f^*$ in (12), t_W from (11), $d_W = \dfrac{2\left(\mu_1 - \mu_2\right)^2}{\pi_1\sigma_1^2 + \pi_2\sigma_2^2}$ and the other symbols as in test 1. The decision rules are as for test 1 with Q_w in place of *Q*.

We announce here that Häusler (2003) investigated the robustness of the *triangular sequential test* described by Schneider (1992) for some Fleishman distributions.

Simulation Experiment 5 - Comparing more than two means

If we take independent samples from k normal distributions and in each distribution a model analogue to (5) can be assumed in the form

$$\underline{y}_i = \mu_i + \underline{e}_i, i = 1,...,k; E(e_i) = 0; \mathrm{var}(e_i) = \sigma^2$$

for all *i*.

Several hypothesis can be tested. If both risk of a pair-wise comparison of all the μ_i are defined pair-wise, the pair-wise *t*-test is the appropriate procedure. This means, for each comparison a two-sample *t*-test with a pooled variance estimate is applied. The tests of STEEL (1959) and NEMENY (1963) mentioned below can be applied in this case, too. Their properties are investigated for comparing means with a control.

To test

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

against the alternative that at least two of the means are different, the *F*-test (of a one-way ANOVA) usually is used.

There are many multiple comparison procedures parametric and non-parametric ones for several situations and comparison-wise or experiment-wise significance levels (see the post hoc tests in SPSS ANOVA branch).

If the risk of the first or the second kind must be understood for one comparison only, this risk is called comparison-wise. If the risk must be understood for any comparison in the experiment, it is called experiment-wise.

This is also true **for comparing *k* –1 means with a standard** or control where comparing the *k*-1 non-standard procedures with each other is not the aim of the experiment. To test (if the standard mean is μ_k) the null hypotheses

$$H_{0ik} : \mu_i = \mu_k \text{ against } H_{Aik} : \mu_i \neq \mu_k$$

we use for a **comparison-wise significance level** the *t*-test for each comparison with a pooled variance estimate (equal variances assumed):

$$s^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.} \right)^2}{\sum_{i=1}^{k} n_i - k} \tag{1.15}$$

A Wilcoxon type rank test of Steel (1959) rejects (based on samples of size *n*) for the significance level α

$$\mathrm{H}_{0_{ik}}: \mu_i = \mu_k \ (i = 1, ..., \ k - 1), \text{ if}$$

$$T_S = \max\left(W_{ik};W_{ik}^*\right) > r\left(k-1,n,1-\alpha\right),\ \ i=1,...,k-1,.$$

For the bounds $r\left(k-1,n,1-\alpha\right),\ \ i=1,...,k-1,$ see Steel (1959). The W_{ik} are defined analogue to (14) and the $W*_{ik}$ are the corresponding values for the inverse rank order.

The non-parametric Kruskal – Wallis –test in the version of NEMENY (1963) is based on

$$T_{KW} = \left|\frac{1}{n}\sum_{j=1}^{n} R_{kj} - \frac{1}{n}\sum_{j=1}^{n} R_{ij}\right|, \left(i=1,...,k-1\right)$$

with the ranks R_{lj} of the j-th element of the l-th sample in the overall ranking of all observations.

$H_{0_{ik}}$: $\mu_i = \mu_k$ (i = 1, ..., k - 1) is rejected, if $T_{KW} > d\left(k-1;\infty;1-\alpha\right)\sqrt{\frac{kn(kn+1)}{12}}\sqrt{\frac{2}{n}}$ with the $d\left(k-1;\infty;1-\alpha\right)$ of Dunnett’s test below.

For an **experiment-wise significance level** we use the Dunnett-test (see RASCH et.al., 1999, page 133):

Reject $H_{0_{ik}}$: $\mu_i = \mu_k$ (i = 1, ..., k - 1) with experiment-wise significance level α_e, whenever

$$\left|\bar{y}_{i.} - \bar{y}_{k.}\right| > \sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_k}\right)}\,.\,d^*$$

holds. The value of d^* is for α = 0.05, $n_1 = n_2 = ... = n_p = n$ and $n_k = n\sqrt{k-1}$ the value d(k-1;f;0.95) from Table A5 and for $n_l = n$ (l = 1,..., k) it is d(k-1;f;0.95) from Table A6 in RASCH et.a.l., 1999.

Again, s^2 is a pooled estimator of σ^2 with f $d.f.$.

RUDOLPH (1985 a,b) compared in a simulation experiment the properties of the Dunnett test with those of the test of Steel and Kruskal and Wallis for k = 3, n = 6 and n = 21; α = 0.01 and α = 0.05 and the following variance structures:

Table 5:
Variance structures of the simulation experiment

Case	$\sigma_1{}^2$	$\sigma_2{}^2$	$\sigma_3{}^2$
1	0.5	1	1
2	0.5	0.5	1
3	0.5	2	1
4	2	2	1
5	1	2	1
6	1	1	1

The power was estimated as the relative frequency P_i of rejecting $H_{0ik} : \mu_i = \mu_3$ (i =1,2) if

$$\frac{|\mu_i - \mu_3| \cdot \sqrt{n}}{\sqrt{\sigma_1{}^2 + \sigma_2{}^2}} = \Delta_i \tag{1.16}$$

Simulation Experiment 6 – Tests for comparing variances

For comparing two variances we assume samples (y_{11}, ..., y_{1n_1}) and (y_{21}, ..., y_{2n_2}) of size n_1 and n_2 drawn independently from two populations with parameters $\mu_1, \sigma_1^2, \gamma_{11}, \gamma_{12}$ and $\mu_2, \sigma_2{}^2, \gamma_{21}, \gamma_{22}$, respectively.

Our purpose is to test the null hypothesis

$$H_0 : \sigma_1{}^2 = \sigma_2{}^2$$

against

$$H_A : \sigma_1{}^2 \neq \sigma_2{}^2$$

The usual test for comparing two variances is the F-test based on the two sample variances on the statistic

$$F = \frac{\max(s_1^2; s_2^2)}{\min\left(s_1^2; s_2^2\right)}$$

and to reject H_0 *if* $F > F\left(n_1 - 1; n_2 - 1; 1 - \alpha\right)$, if $s^2{}_1 > s^2{}_2$; otherwise the *d.f.* in the F-quantile have to be interchanged. NÜRNBERG (1982) compared the power of this F-test with that of 10 further tests (Bartlett, modified Bartlett, modified χ^2, Cochran, Range, Box-Scheffé, Box-Andersen, Jackknife, Levene-z, Levene-s). We report here only the most important members of this set only.

Levene Test: For this test we use the observations y_{1i} and y_{2i} to calculate the quantities

$$u_{1i} = \left|y_{1i} - \bar{y}_1\right|; u_{2i} = \left|y_{2i} - \bar{y}_2\right| \text{ for } \textbf{Levene's } z\textbf{-test}$$

and

$$v_{1i} = \left(y_{1i} - \bar{y}_1\right)^2; v_{2i} = \left(y_{2i} - \bar{y}_2\right)^2 \text{ for } \textbf{Levene's } s\textbf{-test}.$$

Then we carry out an independent samples *t*-test with the *u*-values or the *v*-values for the *z*- and *s*-version of the test, respectively.

For the **Box-Scheffé-test** (also known as Box-test or Box-Kendall-test) (Box, 1953, Scheffé, 1963) the samples are randomly divided into c groups, c given. These groups contain m_{ij}, $i = 1,2; j = 1,...,c$, observations. Let s_{ij}^2 denote an estimator of σ_i^2 and define $z_{ij} = \ln(s_{ij}^2)$ and

$$F^+ = \frac{2(c-1)c\sum_{i=1}^{2}(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2}{\sum_{i=1}^{2}\sum_{j=1}^{c}(z_{ij} - \bar{z}_{i\cdot})^2}$$

The null hypothesis will be rejected if $F^+ > F(1, 2(c-1), 1-\alpha)$. In his simulation study, Nürnberg (1985) considered the cases $c = 2$ and $c = 3$.

Simulation Experiment 7 – Tests for the parameters in regression analysis

Tests and confidence estimation in linear and quasi-linear (for instance polynomial) regression can be expected to behave like the corresponding inferences for means. Because in psychological research intrinsically non-linear regression – like exponential regression – play no important role, we only summarize the results briefly. The different non-linear regression functions investigated in the robustness research are described in Rasch (1995, chapter 16).

In the intrinsically non-linear regression problems arise which are not known in the procedures discussed so far. Parameters can be estimated only iteratively and even for normal distributions the distribution of the estimators (inclusive their expectation and variance) are only known asymptotically (i.e. for a sample size tending to infinity). Therefore we investigated together with the robustness also the behaviour of the estimators and the tests for the regression parameters in the normal case for small samples.

Simulation Experiment 8 - Selection procedures

As already mentioned in Rasch et al (1999) and Rasch (2003) selection procedures should be preferred to multiple comparison because most practical problems are selection problems.

Let us take independent samples from a normal distributions. If the model of the simulation experiment 5 is applied, than we are interested to find the greatest value of the expectations μ_i $(i = 1,...,a)$. Without loss of generality, we assume, that

$$\mu_1 \le \mu_2 \le \cdots \le \mu_a .$$

But this order is not known for the user. There are two formulations of this problem:

The Indifference-zone formulation (Bechhofer, 1954)

We apply the following selection rule.
Choose that value μ_i which belongs to the sample with the greatest mean $\bar{y}_i$.

In order to describe the error probability we define the event of *d-correct selection (dCS):*

A *dCS* occurs, if the selected μ_i is greater as $\mu_a - d$ (Guiard, 1996).

If *P(dCS)* is the probability of *d*-correct selection, than $\beta = 1 - P(dCS)$ denotes the probability of an error. For given values of *d* and β an appropriate value of *n* can be calculated (e.g. Rasch, Herrendörfer, Bock, Victor, Guiard, 1996).

We denote a selection rule to be robust if it's actual error probability β_A is smaller than $1.2 \cdot \beta$, where β denotes the required error probability. Using other estimations $\hat{\mu}_i$ of the μ_i the selection rule can be modified. Following Domröse (1987), we get further selection rules $\mathbf{R}$ using the following estimations $\hat{\mu}_i$.

RB: The classical selection rule, RB, proposed by Bechhofer (1954), uses

$$\hat{\mu}_i = \underline{\bar{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \underline{y}_{ij} .$$

For normal distributions Rasch (1995) presents tables for planning the sample size n.

RTr$_{0.1}$ and RTr$_{0.2}$: The trimmed means $\underline{\bar{y}}_{i\alpha} = \frac{1}{n(1-2\alpha)}\sum_{j=n\alpha+1}^{n(1-\alpha)} \underline{\bar{y}}_{i(j)}$ are less sensitive to outliers. Here $\underline{\bar{y}}_{i(j)}$ denotes the order statistics $\underline{\bar{y}}_{i(1)} \le \underline{\bar{y}}_{i(2)} \le \cdots \le \underline{\bar{y}}_{i(n)}$ of the variables $\underline{\bar{y}}_{ij}$, $j = 1,...,n$. If $n \cdot \alpha$ is not integer, $\underline{y}_{i([n\alpha]+1)}$ and $\underline{y}_{i([n(1-\alpha)]+1)}$ have to be counted by weights $1-(n\alpha - [n\alpha])$, where $[x]$ denotes the integer part of x. We consider the estimators $\hat{\mu}_i = \underline{\bar{y}}_{i\,0,1}$ for rule RTr$_{0.1}$, and $\hat{\mu}_i = \underline{\bar{y}}_{i\,0,2}$ for rule RTr$_{0.2}$, respectively.

RTi$_{0.1}$ and **RTi$_{0.3}$:** Tiku (1981) proposed the estimators

$$\hat{\mu}_i = \underline{\bar{y}}^T_{ir'} = \frac{1}{n-2r+2r\beta'}\left[\sum_{j=r+1}^{n-r} \underline{y}_{i(j)} + r\beta'(\underline{y}_{i(r+1)} + \underline{y}_{i(n-r)})\right],$$

where $r = [0.5 + r'\cdot n]$ and $\beta' = -\varphi(t)(t-\varphi(t))\cdot n/r)n/r$ $(0 < \beta' < 1)$ with t from $1-\Phi(t) = r/n$. Following Tiku (1981) we use $r' = 0.1$, for RTr$_{0.1}$ and $r' = 0.3$ for RTr$_{0.3}$, respectively.

RA: Randlers, Ramberg, and Hogg (1973) described an adaptive selection rule. This rule works in two steps. In the first step by means of a particular estimator of the kurtosis, the distribution of the data is grouped into one of the three groups „light-tailed", „medium-tailed" and „heavy-tailed". In the second step a selection rule is applied depending on the group, found within the first step.

RRS: The selection rule RRS uses the rank sums $R_i = \sum_{j=1}^{n} r_{ij}$ instead of $\hat{\mu}_i$, where r_{ij} denote the rank of y_{ij} among all $n\cdot a$ observations.

For the simulation we apply the so called least favourable configuration:

$$\mu_1 = \mu_2 = \cdots = \mu_{a-1} = \mu_a - d$$

with different values for d . For all other configurations the $P(dCS)$ would be greater. The distributions used in the simulation study of Domröse and Rasch (1987) are shown in Table 6.

Table 6:
(γ_1, γ_2) -values used in the simulation experiment

type of distribution	*N*	*U*	*F*	*F*	*F*	*F*
γ_1 =	0	0	0	1	2	- 2
γ_2 =	0	- 1.2	6	1.5	6	6

N: normal, U: uniform, F: Fleishman

Moreover, the following combinations of n and d/σ are used:

n	d/σ
11	1; 0.75
21	1; 0.75
47	0.75; 0.5

Subset selection formulation (Gupta, 1956))

The goal of a subset selection procedure consists in selecting a subset, s , of the a distributions such, that the probability $P(CS)$, that the „best“ distribution belongs to this subset, is not smaller than a prespecified value P^* . The first of the following developed the following selection rules was developed by Gupta for the case of normal distributions $N(\mu_i, \sigma^2)$:

RG: (Gupta, 1956, 1965) Calculate the sample means $\overline{y}_i$, $i = 1,...,a$, and the pooled variance estimate (15) . Select the distribution i , if

$$\overline{y}_i \geq \max_{1 \leq l \leq a} \overline{y}_l - t_{a-1}(d.f.,P^*)\sqrt{2s^2 / n} .^4$$

Here $t_{a-1}(d.f.;P^*)$ denotes the P^*-quantile of the $(a-1)$-dimensional t-distribution with $d.f.$ degrees of freedom and equal correlations $\rho = 0.5$.

RH: (selection rule of Hsu, 1980): Let $D_{[1]}^{(ij)} \leq \ldots \leq D_{[\mathrm{n}^2]}^{(ij)}$ denote the ordered differences $y_{ig} - y_{jh}$ $(g,h = 1,...,n)$ between the samples of the distributions i and j . Then select distribution i if $\min_{j \neq i} D_{[\mathrm{c}(P^*)\text{-n(n+1)/2}]}^{(\mathrm{ij})} > 0$ and/or $\min_{j \neq i} D_{[\mathrm{medium}]}^{(ij)} > 0$, using

$$D_{[\mathrm{medium}]}^{(ij)} = \begin{cases} D_{[k+1]}^{(ij)} & , \text{ if } n^2 = 2k+1 \\ \frac{1}{2}(D_{[k]}^{(ij)} + D_{[k+1]}^{(ij)}) & , \text{ if } n^2 = 2\mathrm{k} \end{cases} .$$

The values $c(P^*)$ are given as $c(P^*) = r^\alpha$ (one-tailed, $\alpha = 1 - P^*$) in Table VIII of Miller (1966)

RA: The rule RA (Hogg, 1974) is an adaptive selection rule which switches between the rules R_G an R_H in dependence of special estimators of skewness and kurtosis and of the Levene-s-test for comparing the variances (Levene, 1960; Nürnberg, 1985).

Within their simulation experiment Listing and Rasch (1996) used the following distributions:

Table 7:
Simulated distributions

type of distribution	U	N	F	F	F	Chi	F	F
$\gamma_1 =$	0	0	0	0	1.5	1.5	2	-1.5
$\gamma_2 =$	-1.2	0	3.75	7	3.75	3.75	7	3.75

U: uniform; N: normal; F: Fleishman: Chi: non central Chi-square distribution with $d.f. = 1$ and non-centrality parameter $\lambda = 2.426$

[4] Listing and Rasch (1996) erroneously wrote $\sqrt{s^2 / 2}$ instead of $\sqrt{2s^2 / n}$

In order to estimate $P(CS)$, Listing and Rasch (1996) simulated the least favourable configuration

$$\mu_1 = \mu_1 = \cdots = \mu_a .$$

For estimating the expected number $f(d)$ of selected false populations, the configurations

$$\mu_1 = \mu_2 = \cdots = \mu_{a-1} = \mu_a - d \cdot \sigma$$

with different values of d were simulated.

The number of simulation runs was at least $N_s = 9000/a$. Listing and Rasch (1996) simulated some further selection rules, but in the present paper only the best ones were described.

4. Summary of Results and Recommendations

We present here the results of the simulation experiments described in paragraph 3. By Result *x* we denote the result of simulation experiment *x*.

Result 1 - One-sample tests and confidence intervals for the mean

If we use in definition 1 an ε of 20% of the significance level $\alpha = 0.05$, a test is robust as long the actual α_{act} lies between 0.04 and 0.06. In Table 8 the sample sizes are given for which a 20% robustness was found.

Table 8:
Sample sizes, which give a 20% robustness

Test	No. of distribution in Table 1 and its $\gamma_1 -, \gamma_2 - values$	Case a)	Case b)	Case c)
Test 1 (*u*)	1 [0; 0],2 [0; 1.5],3[0; 3.75]	5	5	5
Test 1 (*u*)	4 [0 ; 7],5 [1 ; 1.5],6 [1.5: 3.75]	5	10	10
Test 1 (*u*)	7 [2; 7]	5	50	50
Test 2 (*t*)	1 [0; 0],2 [0; 1.5]	5	5	5
Test 2 (*t*)	3 [0; 3.75]	10	5	5
Test 2 (*t*)	4 [0 ; 7],6 [1.5: 3.75]	10	50	50
Test 2 (*t*)	5 [1 ; 1.5]	30	50	50
Test 2 (*t*)	7 [2; 7]	50	50	50
Test 3 (t_J)	1 [0; 0],2 [0; 1.5]	10	5	5
Test 3 (t_J)	3 [0; 3.75]	50	50	30
Test 3 (t_J)	4 [0 ; 7]	30	50	30
Test 3 (t_J)	5 [1 ; 1.5],6 [1.5: 3.75]	50	30	50
Test 3 (t_J)	7 [2; 7]	50	50	50

It is not surprising that test 3 is quite good if the kurtosis is small. Then it behaves better than the *t*-test (test 2). But for sample sizes larger or equal to 50 all the tests are 20%-robust.

Result 2 – one-sample sequential tests for the mean

The average sample sizes for the 10000 replications did not differ too much for the three tests if d =1, otherwise they were larger for test 3 than for the other tests. It is known that sequential tests guarantee the risks only approximately even in the normal case. This can be seen from Table 9, but the test is conservative.

Table 9:
Percentage of false rejections of H_0 for 8 distributions from Table 4 (columns 3 –10) for $\alpha = 0.05$, $\beta = 0.2$ and $d = 0.6$

d	Test t_i	1	2	3	4	5	6	7	8
		[0; 0]	[1.006;0.6]	[0.35; -0.35]	[1.5; 3.75]	[0; 3.75]	[0; 7]	[1; 1.5]	[1.5; 3.75]
0.6	1	4.34	7.44	4.93	10.41	3.23	2.95	6.17	7.57
0.6	2	4.64	1.69	3.59	11.11	3.33	2.63	1.78	1.36
0.6	3	4.9	9.18	5.41	12.54	3.33	3.53	7.15	7.02

Table 10:
Percentage of false acceptions of H_0 for 8 distributions from Table 4 (columns 3 –10 with [$\gamma_1;\gamma_2$]) for $\alpha = 0.05$, $\beta = 0.2$ and $d = 0.6$

d	Test t_i	1	2	3	4	5	6	7	8
		[0; 0]	[1.006;0.6]	[0.35; -0.35]	[1.5; 3.75]	[0; 3.75]	[0; 7]	[1; 1.5]	[1.5; 3.75]
0.6	1	15.54	8.95	13.78	5.05	14.64	14.76	10.46	7.24
0.6	2	15.46	27.8	19.45	31.3	13.15	11.53	24.01	27.01
0.6	3	18.9	29.87	21.68	38.84	17.74	16.8	27.21	22.75

The Wald test (test 1) has for the normal distribution (distribution 1) the best approximation for both risks and is always robust (more than that, conservative) for the risk of the second kind. With exception of the extremely non-normal distribution 2, 4, 7 and 8 it is also robust for the significance level. From the 3 tests examined test 2 must be preferred even if its power is lower than for test 3 (this is a consequence of the higher first kind risk, the power function of test 3 dominates that of test 2 and 1 with exception of distribution 8) in both positions and for all distributions. But for practical distributions in which either the skewness or the kurtosis is low, the test is robust and the average sample size moderate.

Result 3 – Tests for comparing two means

The results of the experiments described above let to a huge data set and it seems therefore better to repeat here a summary which was published in Guiard, V. und Rasch, D. (ed.) (1987) pages 37-73 unifying both the results of Posten (1978) and Tuchscherer and Pierer (1985). This summary is given in figure 2.

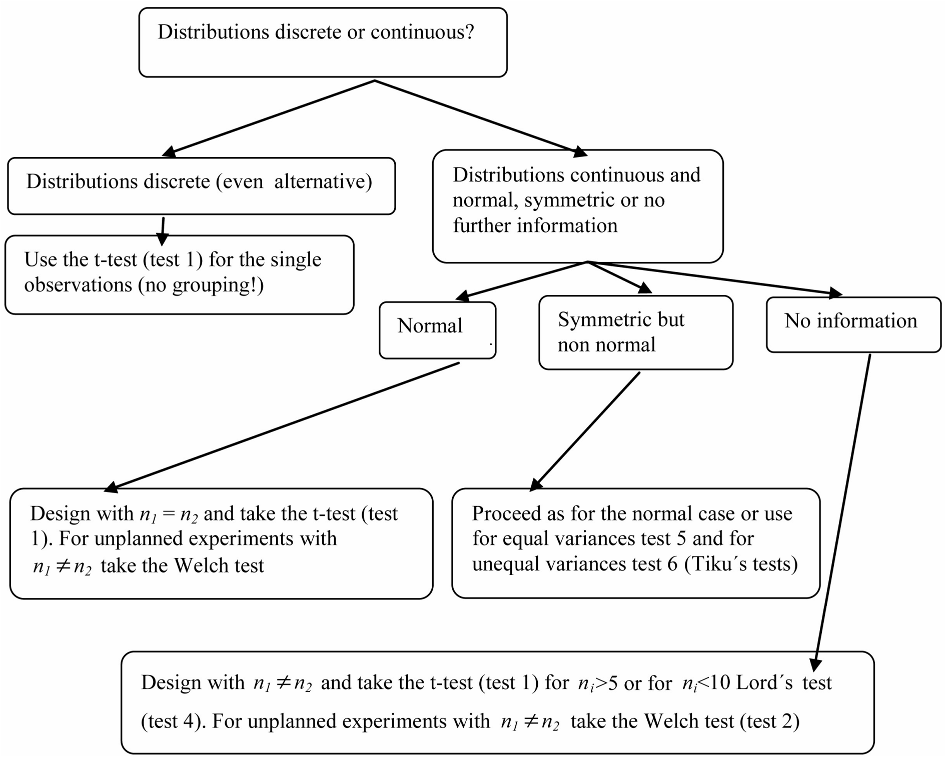
Distributions discrete or continuous?
Distributions discrete (even alternative)
Distributions continuous and normal, symmetric or no further information
Use the t-test (test 1) for the single observations (no grouping!)
Normal
Symmetric but non normal
No information
Design with $n_1 = n_2$ and take the t-test (test 1). For unplanned experiments with $n_1 \neq n_2$ take the Welch test
Proceed as for the normal case or use for equal variances test 5 and for unequal variances test 6 (Tiku´s tests)
Design with $n_1 \neq n_2$ and take the t-test (test 1) for n_i>5 or for n_i<10 Lord´s test (test 4). For unplanned experiments with $n_1 \neq n_2$ take the Welch test (test 2)

Figure 2:
Proposals for the comparison of two means

The Wilcoxon test is in the case of equal variances and equal sample sizes also quite good but there is no real need to use it even with small samples. The tests proposed are in general more robust in cases where no prior knowledge is available and this is in most practical situations the case. Before an experiment is performed, it has to be planned amongst others its size has to be determined (see RASCH, 2003). We recommend to design the experiment with equal sizes in both samples. This not only because the total size $N = n_1 + n_2$ is for a given precision a minimum, if the two sample sizes are equal but also the robustness is larger in this case and the *t*-test can always be applied.

Result 4 – Sequential tests for comparing two means

For pairwise sampling at each step and if the non-normality of the two populations is of the same type both tests are conservative for both risks and 20%-robust. The skewness than has nearly no influence on the risks. Pairwise sampling is recommended.

If the non-normality in both population is different, the skewness influences the risks and the tests are not always robust. Than the kurtosis has no big influence. We think that in practical situations differences in the higher order moments in both populations can not be expected.

For pair-wise sampling in both populations with different probabilities $\pi_1 = 0.75; \pi_2 = 0.25$ the result are given in Table 11.

Table 11:
Estimated actual risks $\hat{\alpha}$ and $\hat{\beta}$ for Hajnal's (suffix H) test and the Welch modification (suffix W) for pairwise sampling and sampling with probabilities $\pi_1 = 0.75; \pi_2 = 0.25$;
$d = d_w = 1.5$ are given for $\alpha = 0.05$ and $\beta = 0.1$

Sample 1		Sample 2		pair-wise sampling				$\pi_1 = 0.75; \pi_2 = 0.25$			
γ_1	γ_2	γ_1	γ_2	$\hat{\alpha}_H$	$\hat{\beta}_H$	$\hat{\alpha}_W$	$\hat{\beta}_W$	$\hat{\alpha}_H$	$\hat{\beta}_H$	$\hat{\alpha}_W$	$\hat{\beta}_W$
0	0	0	0	0.034	0.061	0.031	0.055	0.039	0.067	0.060	0.062
0	1.5	0	1.5	0.036	0.065	0.027	0.064	0.039	0.070	0.061	0.066
0	3.75	0	3.75	0.030	0.067	0.023	0.067	0.037	0.069	0.050	0.065
0	7	0	7	0.022	0.070	0.200	0.068	0.037	0.070	0.044	0.068
0	-1	0	-1	0.038	0.054	0.035	0.055	0.041	0.065	0.091	0.050
0.5	0	0.5	0	0.035	0.058	0.032	0.059	0.041	0.062	0.075	0.043
1	1.5	1	1.5	0.036	0.063	0.026	0.066	0.040	0.058	0.070	0.036
1.5	3.75	1.5	3.75	0.032	0.062	0.024	0.071	0.041	0.059	0.080	0.035
2	7	2	7	0.025	0.070	0.019	0.070	0.039	0.060	0.077	0.034
2	7	0	0	0.046	0.031	0.039	0.093	0.068	0.091	0.071	0.077
0	0	2	7	0.043	0.031	0.038	0.091	0.033	0.025	0.096	0.016
2	7	-2	7	0.095	0.115	0.085	0.119	0.086	0.109	0.122	0.110

Summarizing the results of Table 11 and of further simulation results by Frick (1985) for unequal variances we propose pair-wise sampling and the Welch modification (test 2) which is also robust in the case of equal non-normality but unequal variances in both populations. Recently it was shown by Häusler (2003) that the triangular sequential test (Schneider, 1992) behaves much better than the tests above. More information will be found in Rasch, Kubinger, Schmidtke and Häusler, 2004.

Result 5 – Comparing more than two means

For the multiple *t*-test the robustness is at least as good as for the two-sample *t*-test, because robustness increases with the *d.f.* of the *t*-test. If the size of each sample is *n*, the two-sample *t*-test has 2(*n*-1) *d.f.* but in the *k*-sample case we have *k*(*n* - 1) *d.f.*. Therefore no new simulation is needed. This is analogue in the case of comparing with a standard.

That the *F*-test is very robust against non-normality (but be careful: the *F*-test for comparing variances is extremely non-robust, see result 6) was already shown by ITO (1964) and no further simulation was needed. For the many existing multiple comparisons as for instance listed under „post hoc tests" in the SPSS ANOVA branch no general recommendation can be given but we think that the Tukey and the Student-Newman Keuls – test (based on range) is as robust as the *t*-test due to the two-sample results for the *t*- and the Lord (range) – test. This will be supported by the poor power results of the rank tests in Rudolph's experiment, where these test have been compared with the Dunnett-test. The result of Rudolph (1985b) shows that the Dunnett test must be preferred to the rank tests for *n*< 15, for larger *n* it is still good but the rank tests behave better than for small samples. For non-normal distributions **and** variance heterogeneity no test is really robust. Here a multiple Welch test may be helpful as in the two-sample problem.

Table 12:
Estimated actual risks $100\hat{\alpha}$ for $\alpha = 0.05$, $k = 3$, $n = 6$ and $n = 21$ and variance homogeneity (structur 6) and $100P_1$ and $100P_2$ for $\Delta_1 = \Delta_2 = 3.46$ and $n = 6$ and $n = 21$

γ_1	γ_2	Test	n=6	n = 21	$\frac{P_1+P_2}{2}, n=6$	$\frac{P_1+P_2}{2}, n=21$
0	0	Dunnett	4.9	4.8	83.48	63.46
		Steel	4.85	4.82	75.93	55.58
		Kruskal - Wallis	4.72	4.64	71.48	56.52
0	1.5	Dunnett	4.16	5.13	83.34	63.67
		Steel	4.27	5.34	76.33	62.42
		Kruskal - Wallis	4.31	5.32	73.05	63.29
0	3.75	Dunnett	4.26	4.74	83.87	64.66
		Steel	4.89	4.82	77.29	70.50
		Kruskal - Wallis	4.61	4.64	74.98	71.11
0	7	Dunnett	3.98	4.82	84.05	66.22
		Steel	4.74	5.34	77.06	77.33
		Kruskal - Wallis	4.55	5.32	76.16	78.94
1	1.5	Dunnett	4.75	4.90	83.53	64.04
		Steel	5.01	4.95	76.19	64.28
		Kruskal - Wallis	4.75	4.72	72.63	64.49
1.5	3.75	Dunnett	4.23	4.97	83.51	64.59
		Steel	4.71	5.09	76.01	72.98
		Kruskal - Wallis	4.3	4.98	73.26	72.59
2	7	Dunnett	4.17	4.3	83.62	65.49
		Steel	5.08	5.1	76.79	80.81
		Kruskal - Wallis	4.73	5.01	76.06	80.19

We now report some of Rudolph's results in Table 12. Because P_1 and P_2 differed only by not more than 0.4 we report the arithmetic mean.

The results can not be interpreted as if the power does not increase with increasing *n.* From (16) we see that the same Δ for larger *n* means a smaller difference in the means. The following conclusion can be drawn. For small samples the Dunnett test is uniformely more powerful than the non-parametric competitors. If the sample size becomes larger, the non-parametric tests are slightly better than the Dunnett test for kurtosis of 3.75 and larger. In Figure 1 those values rather than the rule are the exception and due to the fact that a sample size determination is more easy for the Dunnett test, we see no need to replace this test by a non-parametric one. Furthermore we recommend the selection rules in place of multiple comparisons, not only that they need less number of observations but they are also more robust as shown in Result 7.

Result 6 – Tests for comparing variances

Nürnberg (1985) gave the following recommendations:

For small sample size *n* (*n*=6) only the Box-Scheffé-test (*c*=2 or 3) is 20%-robust for all investigated distributions. The *F*-test for comparing variances should never be used and be deleted from statistical program packages.

For *n*=18 and 42 the following four tests are 20%-robust: modified Bartlett, Box-Scheffé, Box-Andersen, Levene-s. The other tests are not 20%-robust for some distributions.

SPSS deleted the *F*-test and replaced it by Leven's -*z*-test as part of the procedure in an independent two-samples *t*-test.

Result 7 - Selection procedures

Indifference-zone formulation

Some of the estimations of *P*(*dCS*), calculated by means of 6000 simulation runs, are shown in Table 13.

From these results we derive the following conclusions and recommendations:

1. The easely handled Bechhofers selection rule **RB** with extensive tables for planning sample size *n* in Rasch et al. (1996) can be recommended.
2. If the underlying distribution is symmetric with unknown kurtosis, we recommend adaptive selection rule **RA**.
3. If the underlying distribution possibly has a negative kurtosis or is skew with low kurtosis and must be trimmed on account of outliers, α should be kept as small as possible and both $\underline{y}_{(n(\alpha+1))}$ and $\underline{y}_{(n(1-\alpha))}$ should be given larger weight, say by the use of $\mathbf{RTr}_{\mathbf{0.1}}$.
4. If the underlying distribution is distinctly skew, we recommend the application of the rank sum selection rule **RRS**.

Table 13:
Estimates $\hat{P}(dCS)$ for different selection rules given $1-\beta = 0.95$

Selection rule	n	d/σ	$\gamma_1 = 0$	0	0	1	2	-2
			$\gamma_2 = -1,2$	0	6	1.5	6	6
RB	21	0.75	0.9553	0.9496	0.9515	0.9454	0.9320	0.9627
	47	0.5	0.9533	0.9495	0.9512	0.9392	0.9483	0.9627
$\mathrm{RTr}_{0.1}$	21	0.75	0,8970	0.9425	0.9847	0.9449	0.9605	0.9787
	47	0.5	0.8940	0.9405	0.9878	0.9456	0.9635	0.9763
$\mathrm{RTr}_{0.2}$	21	0.75	0.8367	0.9313	0.9903	0.9359	0.9585	0.9810
	47	0.5	0.8393	0.9256	0.9907	0.9346	0.9650	0.9790
RTi_{01}	21	0.75	0.9280	0.9423	0.9760	0.9427	0.9485	0.9742
	47	0.5	0.9270	0.9442	0.9783	0.9483	0.9557	0.9730
$\mathrm{RTi}_{0.2}$	21	0.75	0.8150	0.9253	0.9893	0.9290	0.9523	
	47	0.5	0.8168	0.9156	0.9888	0.9244	0.9557	
RA	21	0.75	0.9862	0.9483	0.9897	0.9434	0.9335	0.9677
	47	0.5	0.9995	0.9495	0.9913	0.9485	0.9412	0.9637
RRS	21	0.75	0.9244	0.9434	0.9867	0.9755	0.9982	0.9844
	47	0.5	0.9369	0.9403	0.9884	0.9746	1.	0.9920

The subset selection formulation:

The estimated values of *P*(*CS*) for *P**=0.95 are shown in Table 14.

In case of $\hat{P}(CS) < P^* - \varepsilon$, this entry was set in boldface. As robustness condition we use $\varepsilon = \beta/5 = 0.01$ where $\beta = 1 - P^*$ denotes the error probability. The estimations of the expected probabilities, $f(d)$, of selecting non-best populations are shown in Table 15. In the cases where a subset selection procedure fails in robustness we have marked the corresponding entry with a „*“ sign.

For the case of equal variances, Listing and Rasch (1996) recommend to use the Gupta rule **RG** if it is known, that the distributions are approximately normal. If the distributions are completely unknown, the adaptive rule **RA** should be used.

For the case of unequal variances it was not possible to give a clear recommendation.

5. Final Conclusions

We now summarize the recommendations above in some short conclusions. By *inferences* we mean as well *selection procedures*, *confidence estimations* as also *hypothesis testing*. We restrict ourselves on a significance level α = 0.05 or what means the same on a confidence coefficient of 1-α = 0.95 and on a probability of incorrect selection (which plays the role as the α in testing and confidence estimation) of β = 0.05 . Results for $\alpha(\beta)$ = 0.01 and $\alpha(\beta)$ = 0.1 can be found in the corresponding references.

Table 14:
$\hat{P}(CS)$ for $P^* = 0.95$, $n = 35$, and $k = 10$

Distribution	type	U	N	F	F	F	Chi	F	F
Rule	$\gamma_1 =$	0	0	0	0	1.5	1.5	2	-1.5
	$\gamma_2 =$	-1.2	0	3.75	7	3.75	3.75	7	3.75
RG		0.950	0.950	0.950	**0.949**	**0.942**	**0.939**	**0.940**	0.957
RA		0.950	0.950	0.950	0.952	0.953	0.952	0.952	0.952

Table 15:
$\hat{P}(CS)$ for $d = 0.5$, $P^* = 0.95$, $n = 35$, and $k = 10$

Distribution	type	U	N	F	F	F	Chi	F	F
Rule	$\gamma_1 =$	0	0	0	0	1.5	1.5	2	-1.5
	$\gamma_2 =$	-1.2	0	3.75	7	3.75	3.75	7	3.75
RG		0.614	0.626	0.618	0.612	0.616	0.62*	0.61*	0.618
RH		0.621	0.631	0.532	0.431	0.494	0.399	0.373	0.485
RA		0.072	0.666	0.530	0.431	0.494	0.399	0.373	0.485

In general we can formulate that non-parametric tests are not really needed for the inferences discussed in this article. Only the Wilcoxon test may be applied in some cases but only in few cases (see result 4) it is really better than the *t*-test. But when we consider the problems arising in non-parametric procedures for determining optimal sample sizes as discussed in Rasch (2003) we recommend the use of parametric inferences in general. In special cases as well in Rasch et al. (1996, 1998) as also in the design software CADEMO a warning is given, if the sample size is too small to apply parametric inferences. In designing an experiment, the sample size must be increased to reach robustness and in the analysis of undesigned experiments (we hope that such an approach will disappear in psychological research in the near future) the parametric procedure should be supported by the corresponding non-parametric one.

Selection procedures, the classical and the sequential (test 2) *t*-test (as well as the *F*-test for the effects in ANOVA models) for means and for regression coefficients in linear or quasi-linear regression models are robust. For the *t*-test in the two- or *k*- sample problem this is true, if the variances are equal in all populations. For unequal variances in the two-sample problem the Welch-test is recommended.

For comparing variances the *F*-test is extremely non-robust or sensitive, it should never be used (don't forget that the *F*-test for comparing means is pretty robust). To compare variances, the Box-Scheffé-test (*c*=2 or 3) is 20%-robust for $n \geq 6$ and therefore recommended.

For $n \geq 18$ Levene-s test can also be applied.

6. References

1. Bechhofer, R.E. (1954): A single-sample multiple decision procedure for ranking means of normal populations with known variances. Am. Math. Statist., 16-39.
2. Bock, J. (1982): Definition der Robustheit. In: Guiard, V. (ed.) 1981, 1-9.
3. Box, G.E.P. & Tiao, G.C. (1964): A note on criterion robustness and inference robustness. Biometrika 51, 1-34.
4. Domröse, H., Rasch, D. (1987): Robustness of Selection Procedures. Biom. J. 29, 5, 541-553.
5. Feige, K.-D., Guiard, V., Herrendörfer, G., Hoffmann, J., Neumann, P., Peters, H., Rasch, D. & Vettermann, Th. (1985): Results of Comparisons Between Different Random Number Generators. In: Rasch, D. & Tiku, M.L. (editors): (1985), 30-34.
6. Fleishman, A. J. (1978): A method for simulating non-normal distributions. Psychometrika 43, 521-532.
7. Frick, D. (1985): Robustness of the two-sample sequential t-test. In: Rasch, D. & Tiku, M.L. (editors): (1985), 35-36
8. Guiard, V. (ed.) (1981): Robustheit II – Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik, Heft 5, Dummerstorf-Rostock.
9. Guiard, V. & Rasch, D. (ed.) (1987): Robustheit Statistischer Verfahren. Probleme der angewandten Statistik, Heft 20, Dummerstorf-Rostock.
10. Guiard, V. (1996): Different definitions of Δ-correct selection for the indifference zone formulation. J. of Statistical Planing and Inference 54, 175-199.
11. Gupta, S.S. (1956): On a decision rule for a problem in ranking means. Mimeogr. Ser. No. 150, Univ. of North Carolina, Chapel Hill.
12. Gupta, S.S. (1965): On some multiple decision (selection and ranking rules). Technometrics 7, 225-245.
13. Gupta, S.S. & Hsu, J.C. (1980): Subset selection procedures with application to motor-vehicle fatality data in a two-way layout. Technometrics 22, 543-546.
14. Häusler, (2003): Personal communication.
15. Hajnal, J. (1961): A two-sample sequential t-test. Biometrika 48, 65-75.
16. Herrendörfer, G. (ed.) (1980): Robustheit I - Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik, Heft 4, Dummerstorf-Rostock.
17. Herrendörfer, G. & Rasch, D. (1981): Definition of Robustness and First Results of an Exact Method. Biometrics 37, 605.
18. Herrendörfer, G., Rasch, D. & Feige, K.-D. (1983): Robustness of Statistical Methods. II Methods for the one-sample problem. Biom. Jour. 25, 327-343.
19. Hogg, R.V. (1974): Adaptive robust procedures: a partial review and some suggestions for future applications and theory. JASA 69, 909-927.
20. Hsu, J.C. (1980): Robust and nonparametric subset selection procedures. Comm. Statist. Theory Methods A 9, 1439-1459.
21. Huber, P.J. (1964): Robust estimation of a location parameter. Ann. Math. Statist. 35, 73-101.
22. Huber, P.J. (1972): Robust Statistics: A review. Ann. Math. Statist. 43, 1041-1067.
23. Ito, K. (1969): On the effect of heteroscedasticity and non-normality upon some multivariate test procedures. Proc. Internat. Symp. Multiv. Analysis vol. 2, New York Academic Press, 87-120.
24. Johnson, N.J. (1978): Modified t-tests and confidence intervals for asymmetric populations. JASA 73, 536-544.
25. Listing, J., Rasch, D. (1996): Robustnes of subset selection procedures. J. Statist. Planning and Inference 54, 291-305.

26. Nemeny, P. (1963): Distribution-free multiple comparisons. PhD thesis Princeton Univ., Princeton N.J.
27. Nürnberg, G. (1982): Beiträge zur Versuchsplanung für die Schätzung von Varianzkomponenten und Robustheitsuntersuchungen zum Vergleich zweier Varianzen. Probleme der angewandten Statistik, Heft 6, Dummerstorf-Rostock.
28. Nürnberg, G. (1985): Robustness of two-sample tests for variances. In:. Rasch, D. & Tiku, M.L. (editors): (1985), 75-82.
29. Nürnberg, G. & Rasch, D. (1985): The Influence of Different Shapes of Distributions with the same first four Moments on Robustness. In: Rasch, D. & Tiku, M.L. (editors): (1985), 83-84.
30. Mac Laren , M.D. & Marsaglia, G. (1965): Uniform Random Number Generators. J. Assoc. Comput. Mach. 12, 83-89.
31. Mann, H.B. & Whitney, D.R. (1947): On a Test whether One of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Statist. 18, 50-60.
32. Odeh, R.E. & Evans, J.O. (1974): Algorithm AS70: The percentage points of the normal distribution. Appl. Statist. 23, 96-97.
33. Posten, H.O (1985): Robustness of the Two-Sample T-test. In: Rasch, D. & Tiku, M.L. (editors): (1985), 92-99.
34. Posten, H.O (1978): The Robustness of the Two-Sample T-test over the Pearson System. J. of Statist. Comput. and Simulation, 6, 295-311.
35. Posten, H.O. (1982): Two-Sample Wilcoxon Power over the Pearson System. J. of Statist. Comput. and Simulation, 16, 1-18.
36. Randlers, R.H., Ramberg, J.S. & Hogg, R.V. (1973): An adaptive procedure for selecting the population with largest location parameters. Technometrics 15, 769-778.
37. Rasch, D. & Herrendörfer G. (ed.) (1982): Robustheit III - Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik, Heft 7, Dummerstorf-Rostock
38. Rasch, D. (1983): First results on robustness of the one-sample sequential t-test. Transactions of the 9th Prague Conf. Inf. Theory, Statist Dec. Funct., Rand. Processes. Academia Publ. House CAS, 133-140.
39. Rasch, D. (1984): Robust Confidence Estimation and Tests for Parameters of Growth Functions. In: Gyori, I (Ed.): Szamitastechnikai es Kibernetikai Modserek. Alkalmazasa az orvostudomangban es a Biologiaban, Szeged, 306-331.
40. Rasch, D. & Schimke, E. (1985): Die Robustheit von Konfidenzschätzungen und Tests für Parameter der exponentiellen Regression. In: Rasch, D. (ed.) (1985), 40-92
41. Rasch, D. & Tiku, M.L. (editors) (1985): Robustness of Statistical Methods and Nonparametric Statistics. Proceedings of the Conference on Robustness of Statistical Methods and Nonparametric Statistics, held at Schwerin, May 29-June 2, 1983. Reidel Publ. Co. Dortrecht, Boston, Lancaster, Tokyo.
42. Rasch, D. (1985 a): Robustness of Three Sequential One-Sample Tests Against Non- Normality. In: Rasch, D. & Tiku, M.L. (editors): (1985), 100-103.
43. Rasch, D. (1985 b): Robustness of Sequential Tests for Means. Biom. Jour. 27, 139-148.
44. Rasch, D. (ed.) (1985c): Robustheit IV - Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik, Heft 13, Dummerstorf-Rostock.
45. Rasch, D. (1995): Mathematische Statistik. Joh. Ambrosius Barth, Berlin, Heidelberg (851 S.).
46. Rasch, D. (2003): Determining the Optimal Size of Experiments and Surveys in Empirical Research. Psychology Science, vol. 45 ; suppl. IV ; 3-47.
47. Rasch, D. & Herrendörfer, G. (1981 a): Review of Robustness and Planning of Simulation Experiments for its Investigation. Biometrics 37, 607.

48. Rasch, D. & Herrendörfer, G. (1981 b): Robustheit statistischer Methoden. Rostocker Math. Kolloq. 17, 87-104.

49. Rasch, D., Herrendörfer, G., Bock, J., Victor, N. & Guiard, V. (1998): Verfahrensbibliothek Versuchsplanung und -auswertung. R. Oldenboug Verlag MünchenWien, Band I 1996; Band II 1998.

50. Rasch, D., Kubinger, K., Schmidtke, J. & Häusler (2004): Use and misuse of Hypothesis Testing. In preparation

51. Reed A.H. & Frantz, M.E. (1979): A Sequential Two-Sample t Test using Welch Type Modification for unequal variances. Comm. Statist. A –Theory and Methods, 14, 1459-1471.

52. Rudolph, P.E. (ed.) (1985 a): Robustheit V - Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik, Heft 15, Dummerstorf-Rostock.

53. Rudolph, P.E. (1985 b): Robustness of many-one statistics. Rasch, D. & Tiku, M.L. (editors): (1985), 128-133.

54. Schneider, B. (1992): An Interactive Computer Program for Design and Monitoring of Sequential Clinical Trials. Internatiol Biometric Conference 1992, Hamilton, New Zealand, Invited Paper Volume.

55. Steel, R.G.D. (1959): A multiple comparison rank sum test. Biometrics, 15, 560-572.

56. Teuscher, F. (1979): Ein hierarchischer Pseudo-Zufallszahlengenerator. Unpublished paper at the research centre Dummerstorf-Rostock.

57. Teuscher, F. (1985): Simulation Studies on Robustness of the t- and u- Test against Truncation of the Normal distribution. In: Rasch, D. & Tiku, M.L. (editors): (1985), 145-151.

58. Tiku, M.L. (1980): Robustness of MML estimators based on censored samples and robust test statistics. J. Statist. Planning and Inference, 4, 123-143.

59. Tiku, M.L. (1981): Testing linear contrasts of means in experimental design without assuming normality and homogeneity of variances. Invited paper presented at the March 22 to 26, 1981 Biometric Colloquium of the GDR-Region of the Biometric Society.

60. Tiku, M.L. (1982): Robust statistics for testing equality of means or variances. Comm. Statist Theory Methods A 11, 2543-2558.

61. Tuchscherer, A. (1985): The robustness of some procedures for the two-sample location problem – a simulation study (concept). In: Rasch, D. & Tiku, M.L. (editors): (1985), 159-164.

62. Tuchscherer, A. & Pierer, H. (1985): Simulationsuntersuchungen zur Robustheit verschiedener Verfahren zum Mittelwertvergleich im Zweistichprobenproblem (Simulationsergebnisse). In: Rudolph, P.E. (ed.) (1985 a), 1-42.

63. Verdooren, L.R. (1963): Extended tables of critical values for Wilcoxons test. Biometrika, 50, 177-185.

64. Wald, A. (1947): Sequential Analysis. John Wiley, New York.

65. Welch, B.L. (1947): The Generalization of Student’s Problem when Several Different Population Variances are Involved. Biometrika, 34, 28-35.

66. Wilcoxon, F. (1945): Individual Comparisons by Ranking Methods. Biometrics, 1, 80-82.

67. Zielinsky, R. (1977): Robustness: a quantitative approach. Bull. Laced. Polonaise de Science., Ser. Math., Astr. et Phys., Vol. XXV, 12, 1281-1286.