Stat Papers (2011) 52:219–231
DOI 10.1007/s00362-009-0224-x

REGULAR ARTICLE

The two-sample *t* test: pre-testing its assumptions does not pay off

Dieter Rasch · Klaus D. Kubinger · Karl Moder

Received: 13 October 2008 / Revised: 30 March 2009 / Published online: 28 April 2009

Abstract Traditionally, when applying the two-sample *t* test, some pre-testing occurs. That is, the theory-based assumptions of normal distributions as well as of homogeneity of the variances are often tested in applied sciences in advance of the tried-for *t* test. But this paper shows that such pre-testing leads to unknown final type-I- and type-II-risks if the respective statistical tests are performed using the same set of observations. In order to get an impression of the extension of the resulting misinterpreted risks, some theoretical deductions are given and, in particular, a systematic simulation study is done. As a result, we propose that it is preferable to apply no pre-tests for the *t* test and no *t* test at all, but instead to use the Welch-test as a standard test: its power comes close to that of the *t* test when the variances are homogeneous, and for unequal variances and skewness values $|\gamma_1| < 3$, it keeps the so called 20% robustness whereas the *t* test as well as Wilcoxon's *U* test cannot be recommended for most cases.

Keywords Pre-tests · Two-sample *t* test · Welch-test · Wilcoxon-*U* test

1 Introduction

All statistical tests are derived under specific assumptions as concerns the theoretical distribution of a thought-of sample drawn from a certain population. That is, in

D. Rasch · K. Moder
Department of Landscape, Spatial and Infrastructure Sciences, Institute of Applied Statistics and Computing, University of Natural Resources and Applied Life Sciences, Vienna, Austria

K. D. Kubinger (✉)
Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Vienna, Austria
e-mail: klaus.kubinger@univie.ac.at

mathematical statistics, a theoretical (or abstract) sample of size n is given and this sample is defined as a vector of n independent random variables $\mathbf{y}_i$[1], $i = 1, 2, \ldots, n$, with the same distribution—that is to say, they are identically and independently distributed (i.i.d.). Although, however, some assumptions are needed for the derivation of a certain test statistic's distribution, they are often negligible in practical applications.

There usually exists a statistical test (α test) for some null-hypothesis H_0 which is opposed to an alternative hypothesis H_A, given some assumptions about the respective underlying distribution.

Pre-testing means that before the decision for or against H_0 and H_A can be made using such a test, a researcher tests the assumptions about the distribution using the observations of the random sample(s). Doing so, the overall risk of erroneous decisions is difficult to specify that concerns the tested assumptions and the tested null-hypothesis in question. Only if consecutive, independent sampling were applied for both kinds of statistical tests (the pre-test(s) on the one side and the test of H_0 on the other side), could this overall risk of erroneous decisions be calculated using the multiplication rule of probability theory. However, for reasons of feasibility, just a single sampling of data occurs, meaning that the pre-test(s) and the main test are applied at the same observations. As a consequence, the over-all risk can—due to the dependency of the different test statistics—difficult be calculated in closed form.

Hence, many theoretical statisticians nowadays do not recommend pre-testing (see Moser and Stevens 1992, as concerns testing variance homogeneity, Easterling and Anderson 1978; Schucany and Ng 2006, for testing normal distribution[2]).

In this paper, we now will estimate the overall risk of erroneous decisions in the two-sample t test problem using simulation experiments. As a pre-test of normality, we use the Kolmogorov–Smirnov-test (Kolmogorov 1933; Smirnov 1939) and as a pre-test of variance homogeneity the Levene-test (Levene 1960; this because according to Rasch and Guiard 2004, the F test is very sensitive against non-normality and has already been replaced in SPSS).

[1] Random variables are printed in bold.

[2] Unfortunately, statistical program packages, lecture notes and applied statistical text books still recommend a pre-test at least on variance homogeneity in the two-sample location problem. If we *google* for "variance homogeneity" (24 September 2008) the first two notes are as follows:
Homogeneity of variance The assumption of **homogeneity** of **variance** is that the **variance** within each of the populations is equal. This is an assumption of analysis of **variance …** davidmlane.com/hyperstat/A45619.html
Variance homogeneity test Here is a simple test for the **homogeneity** of **variances**, as required in several statistical tests. changingminds.org/explanations/research/analysis/**variance_homogeneity**.htm
At the latter link, note that the F test (later in the text alternatively the Levene-test and Mauchly's test) is recommended as a pre-test in the package XLSTAT. If the F-value is small enough (a table of critical values is given), then it is considered safe to use the t test. Lecture notes and text books are recommended for this topic. Nothing is said about what to do if variances are not equal. But this is done under http://www.sam.sdu.dk/~nks/St2006uk/Variansanalyse-UK.pdf. There Sørensen writes: "If these assumptions not are fulfilled we can conduct a Kruskal–Wallis test". The equivalent of the latter test in our two-sample problem is the Wilcoxon-(Mann–Whitney-)U test (Wilcoxon 1945; Mann and Whitney 1947) which, as a matter of fact, assumes equal moments higher than the first one if the location parameters are to be compared. We found in our *Google*-search more than 500 notes, and most of them recommend pre-tests as concerns the assumption of variance homogeneity.

2 The two-sample t test: its assumptions and testing the null-hypothesis

As indicated, we restrict ourselves to the location parameters of two independent samples; more specifically, in the following we exemplify everything for two expectations (population means).

We independently sample n_1 and n_2 observations, respectively, from each of two continuous distributions (population 1 and 2) with the existing fourth moment having potentially different expectations μ_1 and μ_2. That is, the null-hypothesis is $H_0 : \mu_1 = \mu_2$. The alternative hypothesis (two-sided) is then $H_A : \mu_1 \neq \mu_2$, or that is to say $H_A : \delta = \mu_1 - \mu_2 \neq 0$.

Given that the null-hypothesis is true, for any α test, the value of the power function is equal to α for all sample sizes n_1 and n_2. Given that the alternative hypothesis is true, the value of the power function depends on the actual value of δ and the sample sizes n_1 and n_2. Thus when we try for a certain type-II-risk of, for instance $\beta = 0.10$, we must fix the δ as well.

Thus, assuming that

- both distributions are normal and
- they have common variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$

the two-sample t test based on the test statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{1}$$

with

$$s^2 = \frac{\sum_{i=1}^{n_1} \left(y_{1i} - \bar{y}\right)^2 + \sum_{i=1}^{n_2} \left(y_{2i} - \bar{y}_2\right)}{n_1 + n_2 - 2} \tag{2}$$

is the most powerful unbiased test for all α. The test statistic (1) is non-centrally t-distributed with non-centrality parameter

$$\lambda = \frac{\mu_1 - \mu_2}{\sigma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{\delta}{\sigma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{3}$$

and $n_1 + n_2 - 2$ degrees of freedom. Under the null-hypothesis, (1) is (centrally) t-distributed.

Now we first have to determine the sample sizes in dependence of given risks and the non-centrality parameter. They should be equal when the variances are equal, because then λ in (3) is a maximum for fixed $n_1 + n_2 = N$ and of course the power of the test is monotonously increasing with λ. For instance, if we choose $\alpha = 0.05$, $\beta = 0.10$, and $\delta/\sigma = (\mu_1 - \mu_2)/\sigma = 2/3$, we come to 49 individuals from each population by solving $n_1 = n_2 = n = \text{CEIL}\left[\frac{2\sigma^2}{\delta^2} \cdot [t\,(2n-2;\,P) + t\,(2n-2;\,1-\beta_0)]^2\right]$, where CEIL$(x)$ is the smallest integer larger or equal to x (see Rasch et al. 2007a)—or instead by using CADEMO (http://www.biomath.de).

For pre-testing, these sample sizes are certainly too small to receive any reasonable power for either the Kolmogorov–Smirnov-test for normality (twice and

independently applied to each of the two samples), but also for testing variance homogeneity. While the determination of a proper sample size for the Kolmogorov–Smirnov-test is hardly possible due to the fact that the distance from the null-hypothesis cannot be defined unequivocally, we recognize that for comparable precision requirements we would need $\alpha = 0.05$, $\beta = 0.10$, and $\frac{\sigma_1^2}{\sigma_2^2} = \frac{3}{2} = 1.5$ (or w.l.o.g. $\frac{\sigma_{\min}^2}{\sigma_{\max}^2} = \frac{2}{3} = 0.66667$) $n_1 = n_2 = 258$ (two-sided alternative hypothesis; again using CADEMO).

We now calculate the (overall) type-I- and type-II-risks using simulations for the case of using the same data for pre-tests and test.

3 The design of the simulation study

Trying for Student's t test, we already calculated the sizes of 49 for both samples, given the risks and a fixed (minimal) difference of means mentioned above. We used several sample sizes coming more or less close to this value. That is, $n_1 = n_2 = 10$ (30; 50; 100). However, we also used unequal sample sizes $n_1 = 10$, $n_2 = 30$; $n_1 = 30$, $n_2 = 10$; $n_1 = 30, n_2 = 100$; and $n_1 = 100, n_2 = 30$ (whether, for instance, $n_1 = 10, n_2 = 30$ or $n_1 = 30, n_2 = 10$ is of importance depends on whether the variances in the two populations are assumed to be different or not).

In the case of non-normal distribution, we generated data from Fleishman's system of distributions (cf. Fleishman 1978). That is, every distribution for which the first four moments do exist can be obtained by the transformation $\boldsymbol{z} = a + b\boldsymbol{x} + c\boldsymbol{x}^2 + d\boldsymbol{x}^3$, where $\boldsymbol{x}$ is a standard normal random variable. By a proper choice of the coefficient a, b, c and d, the random variable $\boldsymbol{z}$ will have any quadruple of first four moments $(\mu = 0, \sigma^2 = 1, \gamma_1, \gamma_2)$, where (γ_1, γ_2) fulfil the moment inequality $\gamma_2 \geq \gamma_1^2 - 2$. Figure 1 shows the parabola $\gamma_2 = \gamma_1^2 - 2$, in which all (empirical as well as theoretical) distributions with existing fourth moment are included (cf. Rasch and Guiard 2004). The dots in the parabola represent values empirically based on many large data sets collected during the robustness research program, which is described in Rasch and Guiard (2004).

By γ_1 and γ_2, we denote the skewness (standardized third moment) and the kurtosis (standardized fourth moment -3), respectively. For $b = 1$ and $a = c = d = 0$, we receive the standard normal distribution as a special case. By the final transformation $\boldsymbol{y} = \mu + \sigma\boldsymbol{z}$ we can generate distributions with parameters $(\mu, \sigma^2, \gamma_1, \gamma_2)$. Apart from normally distributed data (distribution 1 in Table 1), we generated data according to the distributions 2–5 in Table 1. Primarily, both samples came from the same type of distribution, but a sixth assembly realizes the first sample stemming from distribution type 1 and the second sample stemming from distribution type 5.

Inhomogeneous variances with the square roots of the ratio of the variances $\sigma_{\max}^2/\sigma_{\min}^2 = \sigma_1^2/\sigma_2^2$ from 1 (homogeneous variance) to 10 (heterogeneous variances each) were generated in steps of 1. W.l.o.g. we set $\sigma_2^2 = 1$. Furthermore, not only $\delta = 1$ (and $\delta = 0$, i.e. a true null-hypothesis) was simulated, but $\delta = 2, 3, 4$, and 5, in order to get an impression of the power of the test (procedure). Each combination of all these conditions is based on 100,000 pairs of two samples. The number of sample pairs used after the splitting because of the pre-tests' results is shown in Table 2 (in brackets).

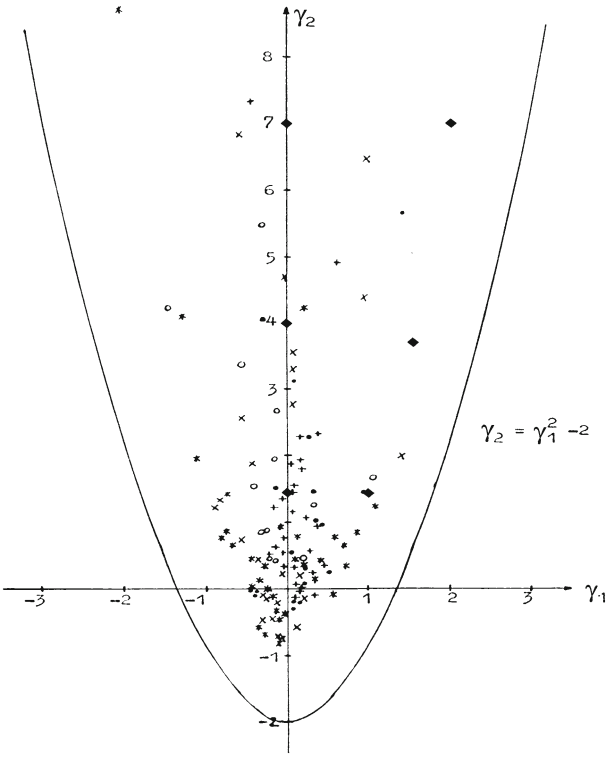

Fig. 1 Values of empirical skewness g_1 and kurtosis g_2 of 144 characters in a (γ_1, γ_2)-plane; by ♦ the parameters (γ_1, γ_2) of some distributions of the Fleishman system are denoted

Table 1 Coeficients of Fleishman's distributions ($a = -c$) with corresponding kurtosis and skewness

Distribution	b	c	d	Skewness	Kurtosis
1	1.000000000	0.000000000	0.000000000	0.0	0
2	2.459073610	0.000000000	−0.755751411	0.0	15
3	0.436347284	0.038013227	0.160514106	0.5	15
4	−1.534711303	0.170094613	0.306848289	1.0	15
5	−0.146400045	0.700711980	0.054356229	3.0	15

Three different statistical tests were applied: Student's t test, the Welch-test (cf. Welch 1947), and the Wilcoxon-U test. They were investigated for two procedures:

(I) pre-tests are used, meaning that the Kolmogoroff–Smirnov-test for testing the normality of each of the two samples and the Levene-test for testing the homogeneity of the variances of the two samples were applied. If normality was rejected in at least one of the two populations, the Wilcoxon-U test was used, even if we had different second, third or fourth moments in the two populations. This was done in order to test to what extent the Wilcoxon-U test is appropriate for

Table 2 A representative summary of the simulation study's results (an extract of different δ, different sample sizes, and different square roots of the ratio of the variances $\sigma^2_{\max}/\sigma^2_{\min} = \sigma^2_1/\sigma^2_2$)

Distribution type	$\sqrt{(\sigma^2_1/\sigma^2_2)}$	n_1	n_2	$\delta = 0$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
1	1	10	10	4.95	4.55 (4.84)	0.00 (0.02)	4.96	4.82	5.21
		30	30	4.97	4.82 (4.98)	11.11 (0.09)	4.96	4.95	4.92
		10	30	5.00	4.87 (4.31)	0.00 (0.06)	5.01	5.05	4.87
		30	10	4.97	5.09 (4.32)	0.00 (0.06)	4.96	5.15	5.00
		30	100	4.86	5.00 (4.80)	9.09 (0.11)	4.84	4.91	4.82
	2	10	10	6.08	3.33 (32.99)	0.00 (0.03)	5.38	4.93	6.06
		30	30	7.37	4.78 (91.89)	10.00 (0.10)	5.15	4.98	5.88
		10	30	1.02	5.90 (30.18)	0.00 (0.07)	0.90	5.00	2.19
		30	10	19.38	2.98 (73.73)	16.67 (0.06)	15.51	5.19	10.09
		30	100	1.33	4.99 (98.38)	0.00 (0.11)	0.63	4.96	1.89
2	1	10	10	4.21	9.14 (3.61)	8.82 (3.29)	4.36	4.17	5.23
		30	30	4.09	6.47 (1.61)	6.54 (28.00)	4.19	4.15	4.92
		10	30	4.64	9.43 (5.55)	6.33 (16.42)	4.92	4.58	5.06
		30	10	4.51	9.53 (5.50)	6.20 (16.46)	4.91	4.45	4.96
		30	100	4.12	0.80 (1.75)	5.43 (70.58)	4.98	4.33	4.99
	2	10	10	7.78	1.27 (44.10)	12.42 (3.30)	4.87	4.42	6.96
		30	30	3.60	4.31 (48.93)	8.57 (28.01)	4.34	4.17	6.64
		10	30	1.47	4.47 (39.22)	3.15 (16.52)	1.13	4.17	1.88
		30	10	18.56	1.96 (59.48)	12.74 (16.49)	13.84	4.76	12.43
		30	100	0.27	2.85 (18.64)	1.94 (70.22)	0.86	4.36	1.57
3	1	10	10	3.87	7.62 (3.28)	5.35 (4.11)	3.90	3.60	5.21
		30	30	4.72	5.32 (2.56)	5.17 (36.40)	4.70	4.63	5.07
		10	30	4.61	1.90 (4.96)	5.47 (21.77)	4.92	3.90	4.97
		30	10	4.63	1.88 (4.95)	5.23 (21.59)	4.92	4.02	4.94
		30	100	5.11	1.09 (1.26)	5.03 (86.41)	4.96	4.51	5.02
	2	10	10	3.80	6.00 (9.63)	6.45 (4.03)	4.07	3.57	5.81
		30	30	4.43	3.95 (22.55)	6.10 (35.91)	4.56	4.40	5.75
		10	30	1.20	10.24 (1.24)	3.38 (21.88)	1.08	4.29	3.03
		30	10	10.95	1.99 (33.52)	9.07 (21.82)	13.86	3.84	8.65
		30	100	0.82	1.95 (3.95)	3.12 (86.33)	0.71	4.69	3.02
4	1	10	10	4.65	11.01 (5.63)	14.63 (0.82)	5.07	4.92	5.32
		30	30	4.47	9.23 (3.76)	9.97 (7.22)	4.92	4.89	4.99
		10	30	4.36	19.84 (5.95)	8.06 (4.22)	5.08	5.45	5.05
		30	10	4.36	19.22 (5.96)	7.84 (4.21)	5.04	5.37	5.02
		30	100	4.23	4.84 (2.66)	5.64 (44.67)	5.05	4.98	5.02
	2	10	10	11.13	1.37 (57.53)	21.52 (0.79)	5.65	5.25	7.42
		30	30	5.48	4.53 (84.08)	16.48 (7.16)	5.19	4.99	7.98
		10	30	1.91	4.06 (71.26)	7.23 (4.15)	0.93	4.80	2.07

Table 2 continued

Distribution type	$\sqrt{(\sigma_1^2/\sigma_2^2)}$	n_1	n_2	$\delta = 0$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
		30	10	34.42	2.30 (81.09)	15.55 (4.18)	15.39	5.67	13.01
		30	100	0.26	2.69 (47.90)	4.58 (44.39)	0.70	4.86	2.64
5	1	10	10	2.08	21.75 (6.98)	7.42 (13.75)	3.80	2.99	5.27
		30	30	3.93	12.12 (0.99)	5.07 (87.04)	4.55	4.39	4.99
		10	30	2.75	5.93 (3.39)	5.02 (66.79)	4.45	7.32	5.01
		30	10	2.57	5.53 (3.44)	5.18 (66.60)	4.43	7.15	5.03
		30	100		0.00 (0.00)	4.92 (100.00)	4.85	6.48	4.92
	2	10	10	9.62	5.84 (13.41)	34.61 (13.84)	8.56	8.07	20.84
		30	30	14.23	0.80 (5.11)	49.95 (87.02)	6.60	6.47	47.48
		10	30	3.62	12.01 (1.65)	27.16 (66.54)	2.76	4.05	21.59
		30	10	27.81	0.34 (15.38)	32.89 (66.71)	16.75	12.26	32.33
		30	100		0.00 (0.00)	69.30 (100.00)	1.36	4.63	69.30
The first sample type 1 the second type 5	1	10	10	4.88	10.65 (13.62)	8.89 (7.20)	5.75	5.33	7.44
		30	30	4.74	7.13 (9.18)	11.09 (64.23)	5.42	5.35	10.07
		10	30	7.77	2.55 (6.38)	10.68 (64.02)	7.04	5.05	10.23
		30	10	2.54	26.40 (14.28)	7.84 (7.40)	3.84	8.27	5.61
		30	100		0.00 (0.00)	15.02 (100.00)	5.87	5.01	15.02
	2	10	10	8.73	12.38 (9.91)	27.96 (7.26)	9.01	8.67	14.35
		30	30	9.74	2.43 (8.42)	37.21 (64.17)	7.01	6.92	31.16
		10	30	4.03	7.50 (0.68)	18.89 (64.06)	3.15	5.51	15.87
		30	10	16.66	4.60 (33.56)	42.08 (7.20)	15.83	12.31	20.60
		30	100		0.00 (0.01)	43.91 (100.00)	1.45	5.17	43.91
Distribution type	$\sqrt{(\sigma_1^2/\sigma_2^2)}$	n_1	n_2	$\delta = \sigma_2 = 1$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
1	1	10	10	56.22	54.60 (4.78)	66.67 (0.03)	56.26	55.68	54.90
		30	30	96.67	96.99 (4.98)	100.00 (0.09)	96.68	96.66	95.90
		10	30	76.04	60.81 (4.21)	66.67 (0.06)	76.07	72.52	73.30
		30	10	76.33	59.17 (4.36)	75.00 (0.08)	76.31	72.86	73.61
		30	100	99.75	99.57 (4.70)	100.00 (0.09)	99.74	99.69	99.62
	2	10	10	30.03	20.62 (33.12)	40.00 (0.05)	27.65	26.12	27.34
		30	30	72.42	66.02 (91.67)	60.00 (0.10)	67.15	66.61	64.21
		10	30	27.15	54.61 (30.34)	42.86 (0.07)	25.80	51.56	34.62
		30	10	58.53	22.54 (73.37)	37.50 (0.08)	52.89	28.55	40.01
		30	100	83.56	95.38 (98.42)	91.67 (0.12)	81.73	95.47	87.56
2	1	10	10	32.09	40.49 (3.67)	33.85 (3.25)	31.67	31.08	35.91
		30	30	76.58	78.05 (1.55)	76.40 (28.01)	66.32	66.08	80.45

Table 2 continued

Distribution type	$\sqrt{(\sigma_1^2/\sigma_2^2)}$	n_1	n_2	$\delta = \sigma_2 = 1$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
		10	30	48.567	29.97 (5.51)	46.28 (16.68)	43.06	42.68	50.17
		30	10	48.37	29.39 (5.51)	48.58 (16.55)	43.28	42.68	50.43
		30	100	94.38	27.29 (1.73)	93.35 (70.67)	77.43	80.12	94.08
	2	10	10	18.86	10.66 (43.75)	18.87 (3.18)	15.44	14.34	14.11
		30	30	31.03	43.94 (48.87)	29.54 (27.96)	36.18	35.62	29.07
		10	30	11.22	36.12 (39.15)	12.62 (16.48)	10.85	27.25	11.03
		30	10	35.66	12.37 (59.44)	22.83 (16.47)	33.70	15.64	22.33
		30	100	41.14	56.25 (18.24)	37.61 (70.70)	32.37	59.63	37.99
3	1	10	10	66.43	68.83 (3.23)	82.58 (4.19)	65.48	64.36	80.15
		30	30	97.84	83.77 (2.65)	99.89 (36.24)	94.93	94.86	99.89
		10	30	83.09	38.57 (4.84)	93.82 (21.85)	78.83	77.57	93.12
		30	10	83.59	50.00 (5.01)	93.36 (21.54)	78.20	78.23	93.76
		30	100	100.00	26.23 (1.26)	100.00 (86.63)	99.16	98.15	100.00
	2	10	10	37.76	43.50 (9.74)	55.91 (4.06)	37.83	35.87	53.68
		30	30	75.51	64.01 (22.59)	94.95 (36.05)	71.64	71.16	94.63
		10	30	36.70	75.00 (1.25)	77.55 (21.91)	34.06	58.80	73.52
		30	10	65.47	25.86 (33.63)	68.96 (22.00)	59.69	41.06	69.78
		30	100	92.06	35.47 (3.85)	99.79 (86.58)	82.00	94.03	99.77
4	1	10	10	59.42	58.97 (5.62)	54.76 (0.84)	59.20	58.82	56.45
		30	30	97.69	92.15 (3.72)	92.54 (7.24)	96.28	96.24	96.24
		10	30	80.06	65.60 (6.06)	63.59 (4.12)	78.48	75.47	73.45
		30	10	79.81	65.78 (5.84)	79.16 (4.27)	78.11	76.71	75.51
		30	100	99.92	50.95 (2.76)	99.28 (44.49)	99.18	98.51	99.54
	2	10	10	24.80	24.53 (57.56)	24.05 (0.79)	26.13	24.21	18.34
		30	30	53.54	69.91 (84.12)	38.10 (7.27)	69.92	69.34	42.57
		10	30	17.39	59.60 (71.12)	12.44 (4.10)	26.42	54.32	18.95
		30	10	34.84	24.81 (81.19)	32.86 (4.20)	53.29	25.46	28.14
		30	100	86.49	59.40 (47.97)	46.43 (44.33)	83.52	95.04	61.73
5	1	10	10	68.88	59.27 (6.91)	85.91 (13.91)	65.93	65.00	82.31
		30	30	98.24	18.79 (1.01)	99.87 (87.07)	94.81	94.72	99.87
		10	30	89.70	8.76 (3.31)	90.99 (66.50)	79.24	73.64	91.17
		30	10	82.59	50.40 (3.31)	98.69 (66.59)	77.16	83.77	98.07
		30	100		0.00 (0.00)	99.98(100.00)	98.88	96.81	99.98
	2	10	10	21.34	69.67 (13.51)	53.94 (13.85)	27.81	24.35	49.43
		30	30	76.11	19.64 (5.20)	92.94 (87.14)	75.09	74.63	92.98
		10	30	41.04	54.17 (1.69)	62.95 (66.39)	27.31	60.14	65.40
		30	10	32.73	16.31 (15.53)	72.96 (66.75)	55.52	21.66	68.62
		30	100		0.00 (0.00)	98.26 (100.00)	86.69	93.42	98.26

Table 2 continued

Distribution type	$\sqrt{(\sigma_1^2/\sigma_2^2)}$	n_1	n_2	$\delta = \sigma_2 = 1$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
The first sample type 1 the seond type 5	1	10	10	64.84	39.47 (13.76)	47.13 (7.15)	60.07	58.56	58.39
		30	30	99.74	56.80 (8.99)	95.71 (64.27)	98.65	98.62	96.40
		10	30	87.57	29.37 (6.34)	70.77 (64.21)	80.00	73.37	72.93
		30	10	84.92	82.22 (14.48)	66.62 (7.19)	80.33	86.91	79.87
		30	100		0.00 (0.00)	99.18 (100.00)	99.89	99.79	99.18
	2	10	10	19.66	43.05 (10.07)	9.27 (7.44)	21.22	19.18	20.30
		30	30	80.06	37.42 (8.47)	43.12 (64.20)	75.61	75.05	51.05
		10	30	36.39	43.75 (0.66)	24.41 (64.03)	23.80	52.60	28.76
		30	10	34.61	29.43 (33.99)	8.941 (7.27)	52.02	20.68	27.20
		30	100		0.00 (0.00)	68.92 (100.00)	87.92	97.32	68.92
Distribution type	$\sqrt{(\sigma_1^2/\sigma_2^2)}$	n_1	n_2	$\delta = \sigma_2 = 5$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
1	1	10	10	100.00	100.00 (4.76)	100.00 (0.03)	100.00	100.00	100.00
		30	30	100.00	100.00 (4.99)	100.00 (0.10)	100.00	100.00	100.00
		10	30	100.00	100.00 (4.29)	100.00 (0.06)	100.00	100.00	100.00
		30	10	100.00	100.00 (4.33)	100.00 (0.05)	100.00	100.00	100.00
		30	100	99.99	99.79 (4.69)	100.00 (0.13)	100.00	100.00	100.00
	2	10	10	99.99	99.97 (33.21)	100.00 (0.03)	100.00	100.00	100.00
		30	30	100.00	99.91 (91.94)	100.00 (0.09)	100.00	100.00	100.00
		10	30	100.00	99.97 (29.91)	100.00 (0.06)	100.00	100.00	100.00
		30	10	100.00	99.95 (73.42)	100.00 (0.06)	100.00	100.00	100.00
		30	100	100.00	99.89 (98.41)	100.00 (0.11)	100.00	100.00	100.00
2	1	10	10	98.89	99.73 (3.69)	100.00 (3.24)	98.06	97.85	99.98
		30	30	100.00	93.18 (1.64)	100.00 (27.84)	99.93	99.93	100.00
		10	30	99.99	93.57 (5.51)	100.00 (16.50)	99.35	98.81	100.00
		30	10	99.81	92.12 (5.45)	100.00 (16.56)	99.39	98.93	99.99
		30	100	100.04	34.55 (1.72)	100.00 (70.74)	100.00	99.98	100.00
	2	10	10	91.04	99.61 (43.72)	99.69 (3.24)	93.72	93.13	99.87
		30	30	100.00	97.31 (48.78)	100.00 (27.82)	99.09	99.06	100.00
		10	30	99.73	98.43 (39.00)	100.00 (16.60)	95.59	97.74	100.00
		30	10	99.09	91.08 (59.40)	99.94 (16.52)	97.38	94.58	99.98
		30	100	100.00	70.65 (18.33)	100.00 (70.54)	99.90	99.96	100.00
3	1	10	10	100.00	100.00 (3.25)	100.00 (4.15)	99.98	99.98	100.00
		30	30	100.00	86.38 (2.60)	100.00 (36.11)	100.00	100.00	100.00
		10	30	100.00	93.84 (5.04)	100.00 (21.92)	100.00	99.99	100.00

Table 2 continued

Distribution type	$\sqrt{(\sigma_1^2/\sigma_2^2)}$	n_1	n_2	$\delta = \sigma_2 = 5$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
		30	10	100.00	93.38 (4.95)	100.00 (21.73)	100.00	99.99	100.00
		30	100	99.92	26.04 (1.25)	100.00 (86.51)	100.00	100.00	100.00
	2	10	10	99.56	99.49 (9.73)	99.75 (4.06)	99.56	99.47	99.95
		30	30	99.98	84.04 (22.33)	100.00 (36.01)	100.00	100.00	100.00
		10	30	99.99	97.69 (1.27)	100.00 (21.59)	99.98	99.98	100.00
		30	10	99.96	89.07 (33.67)	100.00 (21.56)	99.94	99.57	99.99
		30	100	100.00	37.13 (3.94)	100.00 (86.33)	100.00	100.00	100.00
4	1	10	10	99.99	99.82 (5.52)	100.00 (0.82)	99.96	99.95	100.00
		30	30	100.00	92.95 (3.56)	100.00 (7.47)	100.00	100.00	100.00
		10	30	100.00	97.02 (5.87)	100.00 (4.25)	100.00	99.96	100.00
		30	10	100.00	97.06 (5.94)	100.00 (4.20)	100.00	99.99	100.00
		30	100	100.00	53.35 (2.71)	100.00 (44.46)	100.00	100.00	100.00
	2	10	10	99.71	99.83 (57.81)	100.00 (0.82)	99.81	99.77	100.00
		30	30	100.00	95.78 (83.85)	100.00 (7.40)	100.00	100.00	100.00
		10	30	100.00	97.67 (70.91)	100.00 (4.19)	99.96	99.96	100.00
		30	10	99.93	97.35 (80.92)	100.00 (4.15)	99.97	99.83	100.00
		30	100	100.00	60.42 (48.00)	100.00 (44.29)	100.00	100.00	100.00
5	1	10	10	100.00	96.14 (6.97)	100.00 (13.85)	99.99	99.98	100.00
		30	30	100.00	20.65 (1.01)	100.00 (87.27)	100.00	100.00	100.00
		10	30	100.00	53.91 (3.38)	100.00 (66.68)	100.00	99.99	100.00
		30	10	100.00	52.78 (3.32)	100.00 (66.62)	100.00	100.00	100.00
		30	100		0.00 (0.00)	100.00 (100.00)	100.00	100.00	100.00
	2	10	10	99.99	95.44 (13.40)	100.00 (13.88)	99.98	99.97	100.00
		30	30	100.00	19.27 (5.09)	100.00 (87.18)	100.00	100.00	100.00
		10	30	100.00	57.34 (1.68)	100.00 (66.71)	100.00	99.97	100.00
		30	10	100.00	43.80 (15.67)	100.00 (66.71)	100.00	100.00	100.00
		30	100		0.00 (0.00)	100.00 (100.00)	100.00	100.00	100.00
The first sample type 1 the second type 5	1	10	10	100.00	98.06 (13.63)	100.00 (7.24)	100.00	100.00	100.00
		30	30	100.00	58.70 (9.24)	100.00 (64.13)	100.00	100.00	100.00
		10	30	100.00	61.13 (6.40)	100.00 (63.79)	100.00	100.00	100.00
		30	10	100.00	96.05 (14.34)	100.00 (7.30)	100.00	100.00	100.00
		30	100		0.00 (0.00)	100.00 (100.00)	100.00	100.00	100.00
	2	10	10	100.00	98.23 (9.99)	100.00 (7.19)	100.00	99.99	100.00
		30	30	100.00	39.03 (8.45)	100.00 (64.03)	100.00	100.00	100.00
		10	30	100.00	58.26 (0.67)	100.00 (64.31)	100.00	100.00	100.00

Table 2 continued

Distribution type	$\sqrt{(\sigma_1^2/\sigma_2^2)}$	n_1	n_2	$\delta = \sigma_2 = 5$					
				With pre-tests			Without pre-tests		
				t	Welch (Levene)	U (KS)	t	Welch	U
		30	10	100.00	94.02 (33.80)	100.00 (7.28)	100.00	100.00	100.00
		30	100		0.00 (0.00)	100.00 (100.00)	100.00	100.00	100.00

The entries give the percentage of rejecting the (final) hypothesis H_0:$\mu_1 = \mu_2$. t is for Student's t test and U for the Wilcoxon-U test. For the procedure (I), where pre-tests were used, the percentage of significant results is additionally given in *parentheses*. (Missing results indicate that because of division by zero no calculation was possible)

comparing first moments (means) if higher moments differ in the two populations. If normality was accepted, the homogeneity of variances was tested using the Levene-test. If homogeneity was accepted, we continued with the t test, otherwise with the Welch-test for testing the equality of means. For a nominal type-I-risk of $\alpha_{\text{nom}} = 0.01, 0.05$, and 0.10, the actual type-I-risk was estimated using the relative frequency $\hat{\alpha}_{\text{act}}$ of (erroneously) rejected cases in the 100,000 generated sample pairs. By using different means in the two populations, we tried for information about the actual power function.

(II) no pre-tests were applied but each of the three mentioned tests were used for the comparison of type-I- and type-II-risk.

Calculation of Student's t test, the Welch-test, the Wilcoxon-U test, the Kolmogorov–Smirnov-test, and the Levene-test was done using Fortran programs—the accuracy of the programs was tested by comparing the results with those of SAS 9.2.

4 Results of the simulation study

Concerning procedure (I), Fig. 2 gives the exemplary case of $\alpha = 0.05$ and $n_1 = n_2 = 30$ if the null-hypothesis ($H_0 : \mu_1 = \mu_2$; i.e. $\delta = 0$) is actually true and normal distribution as well as homogeneity of variances is given. Additionally, Fig. 2 gives the cases $\delta = 1, 2, 3, 4$, and 5, and the cases of the square root of the ratio of the variances $\sigma^2_{\text{max}}/\sigma^2_{\text{min}} (= \sigma_1^2/\sigma_2^2$ without loss of generality; $\sigma_2^2 = 1$) being 2–10 (i.e. there are heterogeneous variances). The ordinates give the estimated probability of a significant test. Bear in mind that Fig. 2 represents procedure (I), but nevertheless all three tests (Student's t test, the Welch-test, and the Wilcoxon-U test) were calculated. Figure 3 discloses similar cases, but now both the distributions were chosen according to type 3 in Table 1, and $n_1 = 100$, $n_2 = 30$ was used. More detailed results are given in Table 2, which gives a representative summary of the all the simulations mentioned above.

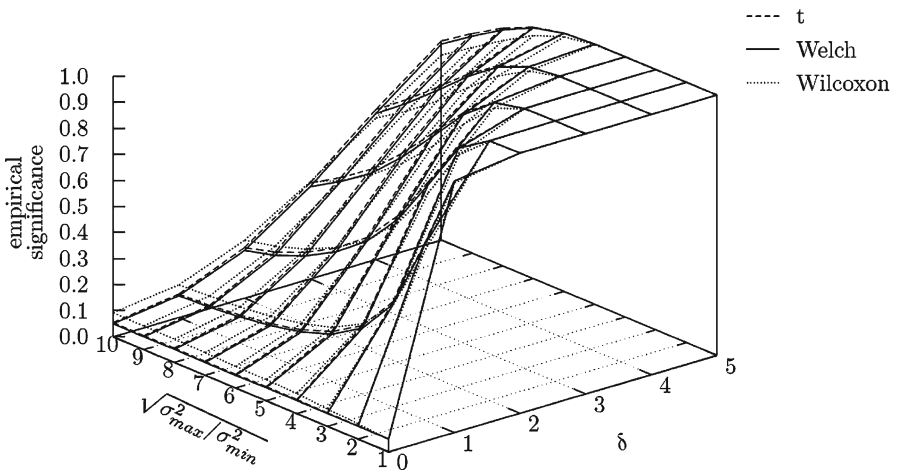

Fig. 2 The estimated actual type-I-risk and the estimated power of the three pertinent tests for $n_1 = n_2 = 30$ at different square roots of the ratio of the variances $\sigma^2_{\max}/\sigma^2_{\min}$—both distributions are normal and using a nominal type-I-risk of $\alpha_{\text{nom}} = 0.05$

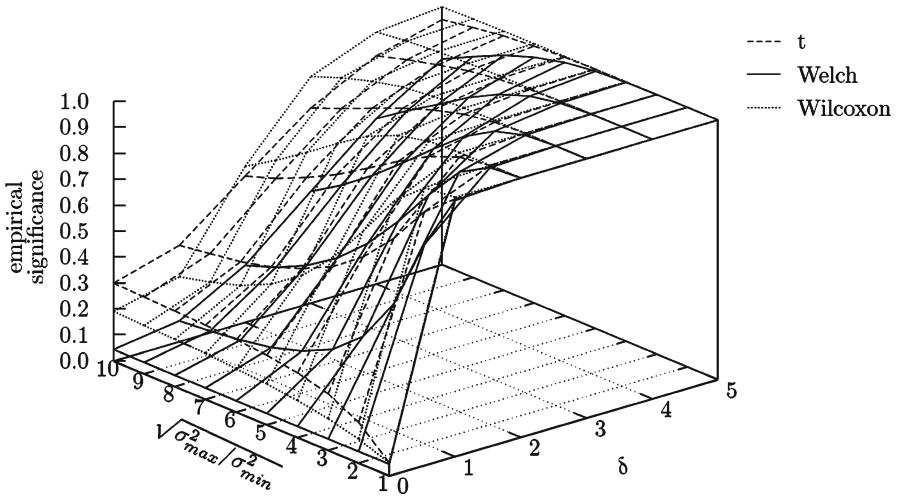

Fig. 3 The estimated actual type-I-risk and the estimated power of the three pertinent tests for $n_1 = 100$ and $n_2 = 30$ at different square roots of the ratio of the variances $\sigma^2_{\max}/\sigma^2_{\min}(= \sigma^2_1/\sigma^2_2)$—both distributions are of type 3 and using a nominal type-I-risk of $\alpha_{\text{nom}} = 0.05$

5 Discussion

As Table 2 illustrates, pre-testing will not lead to any improvement as concerns erroneous decisions in comparison to applying a (proper) test without pre-testing. In most cases the Wilcoxon-*U* test has a lower power than the other tests and if there are unequal variances and/or unequal higher moments, it has an actual type-I-risk which is too large and of course entails a higher power. The *t* test is equivalent to the Welch-test

if variances are equal but does not keep the type-I-risk if the variances are heterogeneous. The Welch-test behaves very well most of the time. Only in the extreme case of distributions of type 5, that is with $\gamma_1 = 3$; $\gamma_2 = 15$ (which, however, occurs very rarely according to to Fig. 1) or in the case of very small samples if one distribution is normal and the other of type 5, does the Welch-test perform more poorly than desired. In this case, however, there is no alternative. Given any other distribution, nearly all estimated actual type-I-risks in Table 2 lie near 5% for the Welch-test. Furthermore, the power of the test is acceptable, though it is no surprise that the power of a test starting under H_0 at a higher α_{act} than 0.05 is higher than the power of an 0.05-test.

The results for $\alpha = 0.01$ and $\alpha = 0.10$ are similar to those for $\alpha = 0.05$, and therefore it is most likely that the nominal α does not play any role for our conclusions.

6 Conclusion

From Fig. 2 and in much more detail from Table 2 (but also for all the results left by the authors), we can conclude the following:

- The assumptions underlying the two-sample t test should not be pre-tested.
- The Welch-test should be introduced in text books and statistical program packages as the standard test for comparing expectations.
- The Wilcoxon-U test should not be used in the given context.

References

Easterling RG, Anderson HE (1978) The effect of preliminary normality goodness of fit tests on subsequent inference. J Stat Comput Simul 8:1–11

Fleishman AJ (1978) A method for simulating non-normal distributions. Psychometrika 43:521–532

Kolmogorov AV (1933) Sulla determinazione empirica di una legge di distribuzione. Inst Ital Attuari Gorn 4:1–11

Levene H (1960) Robust tests for equality of variances. In: Olkin I (ed) Contributions to probability and statistics. Essays in honor of Harold Hotelling. University Press, Stanford, pp 278–292

Mann HB, Whitney DR (1947) On a test whether one of two random variables is stochastically larger than the other. Ann Math Stat 18:50–60

Moser BK, Stevens GR (1992) Homogeneity of variance in the two-sample means test. Am Stat 46:19–21

Rasch D, Guiard V (2004) The robustness of parametric statistical methods. Psychol Sci 46:175–208

Rasch D, Teuscher F, Guiard V (2007a) How robust are tests for two independent samples? J Stat Plan Inference 137:2706–2720

Rasch D, Verdooren LR, Gowers JI (2007b) Design and analysis of experiments and surveys (2nd edn.). Oldenbourg, München

Schucany WR, Ng HKT (2006) Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. Commun Stat Theory Methods 35:2275–2286

Smirnov VI (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull Math Univ Moscou 2:3–14

Welch BL (1947) The generalisation of "Student's" problem when several different population variances are involves. Biometrika 34:28–35

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80–82