Do not use the Two sample-t-test any more!

Dieter Rasch[1]; Klaus, D. Kubinger[2] & Karl Moder[1]

28.6.2009 – 4.7.2009

,

[1] University of Natural Resources and Applied Life Sciences, Vienna, Institute of Applied Statistics and Computing, E-mail: karl.moder@boku.ac.at, dieter.rasch@boku.ac.at

[2] University of Vienna, Faculty of Psychology, Division of Psychological Assessment and Applied Psychometrics, E-mail: klaus.kubinger@univie.ac.at

Table of contents

Problem

Problem

Given two random variables with expectations

$$E(\boldsymbol{y}_1) = \mu_1; \qquad E(\boldsymbol{y}_2) = \mu_2$$

and existing fourth moments.

Problem

Given two random variables with expectations

$$E(\boldsymbol{y}_1) = \mu_1; \qquad E(\boldsymbol{y}_2) = \mu_2$$

and existing fourth moments.

We have to test the null hypothesis:

$$H_0: \ \mu_1 = \mu_2$$

Problem

Given two random variables with expectations

$$E(\boldsymbol{y}_1) = \mu_1; \qquad E(\boldsymbol{y}_2) = \mu_2$$

and existing fourth moments.

We have to test the null hypothesis:

$$H_0: \ \mu_1 = \mu_2$$

against a two-sided alternative hypothesis:

$$H_A: \ \mu_1 \neq \mu_2$$

t-Test

Given 2 independent random samples

$$\boldsymbol{y}_i^T = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{in_i}),\ n_i;\ i = 1, 2$$

distributed as $N(\vec{\mu}_i; \sigma^2 I_{n_i}); \quad \sigma^2 > 0$

Then

$$\boldsymbol{s}^2 = \frac{\boldsymbol{SQ}_{y_1} + \boldsymbol{SQ}_{y_2}}{n_1 + n_2 - 2}$$

is the pooled estimator of σ^2.

t-Test

From this it follows that

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

is distributed as

$$t\left(n_1 + n_2 - 2; \lambda = \frac{\mu_1 - \mu_2}{\sigma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\right).$$

t-Test

From this it follows that

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

is distributed as

$$t\left(n_1 + n_2 - 2; \lambda = \frac{\mu_1 - \mu_2}{\sigma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\right).$$

A uniformely best unbiased α-test for all $0 < \alpha < 1$ is

$$k\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{matrix} 1, \text{ for } |t| > t(n_1 + n_2 - 2; 1 - \frac{\alpha}{2}) \\ 0, \text{ otherwise} \end{matrix}$$

Welch-Test

Welch-Test

In the case of unequal variances Welch(1947) and Trickett and Welch(1954) proposed an approximate test based on

$$\boldsymbol{t}^* = \frac{\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2}{\sqrt{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}}$$

Welch-Test

In the case of unequal variances Welch(1947) and Trickett and Welch(1954) proposed an approximate test based on

$$\boldsymbol{t}^* = \frac{\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2}{\sqrt{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}}$$

The test is:

$$k\begin{pmatrix}\boldsymbol{y}_1 \\ \boldsymbol{y}_2\end{pmatrix} = \begin{matrix} 1, \text{ for } |t^*| \; t > (f; 1 - \frac{\alpha}{2}) \\ 0, \text{ otherwise} \end{matrix}$$

Welch-Test

In the case of unequal variances Welch(1947) and Trickett and Welch(1954) proposed an approximate test based on

$$\boldsymbol{t}^* = \frac{\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2}{\sqrt{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}}$$

The test is:

$$k\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix} = \begin{matrix} 1, & \text{for } |t^*| \; t > (f; 1 - \frac{\alpha}{2}) \\ 0, & \text{otherwise} \end{matrix}$$

with:

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

Which Test?

Hence, many theoretical statisticians nowadays do not recommend pre-testing (see Moser & Stevens, 1992), as concerns testing variance homogeneity, Easterling & Anderson, 1978, and Schucany & Ng, 2006, for testing normal distribution. Nevertheless in applied statistics pre-testing is often applied.

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 test" (24^{th} Sept, 2008) a note is as follows:

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 test" (24^{th} Sept, 2008) a note is as follows:

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

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 test" (24^{th} Sept, 2008) a note is as follows:

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 *N.K.* $S\phi rensen$ writes:

"If these assumptions not are fulfilled we can conduct a Kruskal-Wallis 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 *N.K. 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 & Whitney, 1947) which, as a matter of fact, assumes equal moments higher than the first one if the location parameters are to be compared.

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 *N.K.* $S\phi rensen$ 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 & 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.

Pre-testing

Pre-testing means that before the decision between the two hypotheses is made by the 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.

Pre-testing

Pre-testing means that before the decision between the two hypotheses is made by the 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.

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.

In this paper, we now will estimate the overall risk of erroneous decisions in the two-sample t-test problem using simulation experiments.

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.

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,1930; Smirnov, 1939) and as a pre-test of variance homogeneity the Levene-test (Levene, 1960; this because according to Rasch & Guiard, 2004, the F-test is very sensitive against non-normality and has already been replaced in SPSS).

Considerations about problems with Pre-test

Considerations about problems with Pre-test

If the same sample (as usual) is used as well for the pre-test as also for the final test we have a sample size problem.
We go back to our two-sample problem.

Considerations about problems with Pre-test

If the same sample (as usual) is used as well for the pre-test as also for the final test we have a sample size problem.
We go back to our two-sample problem.

We have to test the null hypothesis:

$$H_0: \ \mu_1 = \mu_2$$

Considerations about problems with Pre-test

If the same sample (as usual) is used as well for the pre-test as also for the final test we have a sample size problem.
We go back to our two-sample problem.

We have to test the null hypothesis:

$$H_0: \ \mu_1 = \mu_2$$

against a two-sided alternative hypothesis:

$$H_A: \ \mu_1 \neq \mu_2$$

Which sample size for pre-tests?

Which sample size for pre-tests?

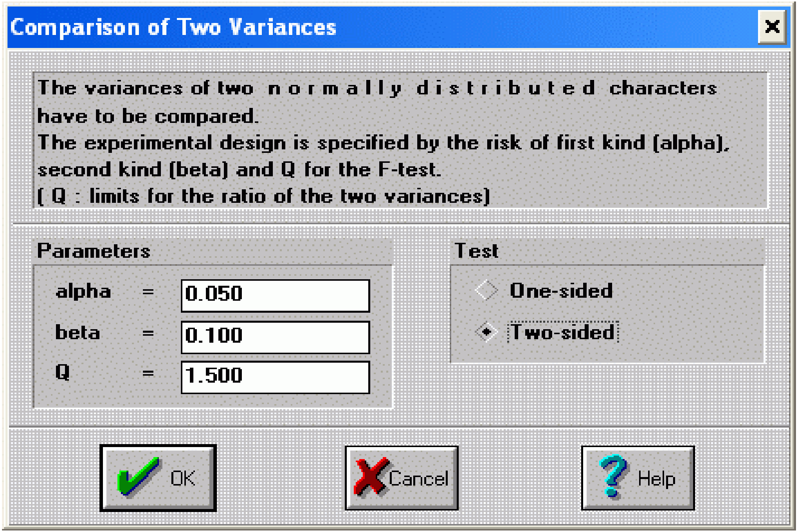
Let us assume we like to test all hypotheses with a first kind risk of 0.05 and a second kind risk of 0.1.

Which sample size for pre-tests?

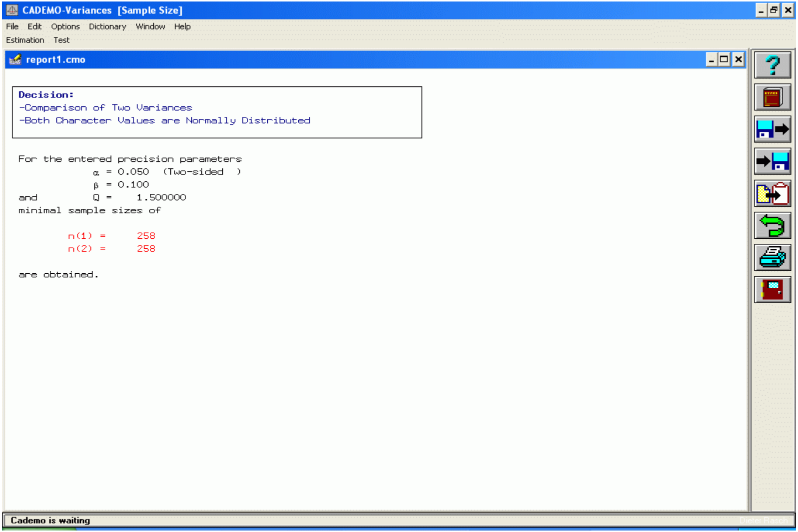
Let us assume we like to test all hypotheses with a first kind risk of 0.05 and a second kind risk of 0.1.

For the Kolmogorov-Smirnov-tests (normality of each distribution) the sample size is difficult to calculate. But we know that it is relatively large $\rightarrow$ 500.

Sample size for comparing 2 variances



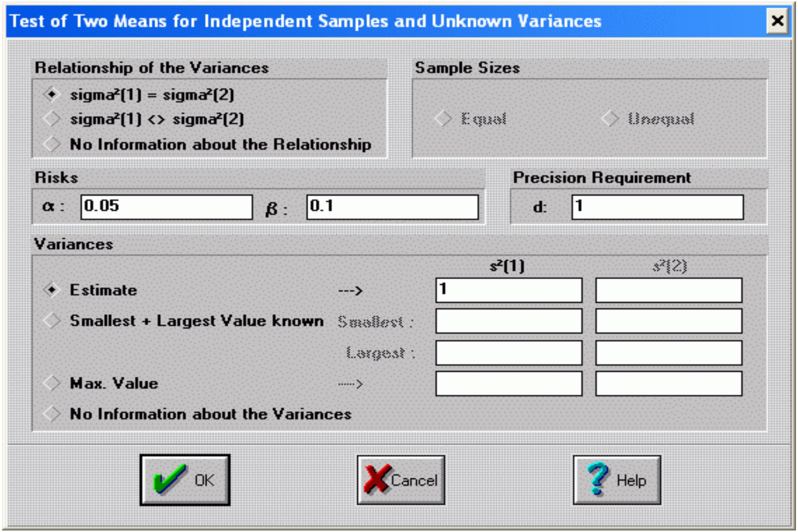
Results (sample size for comparing 2 variances)



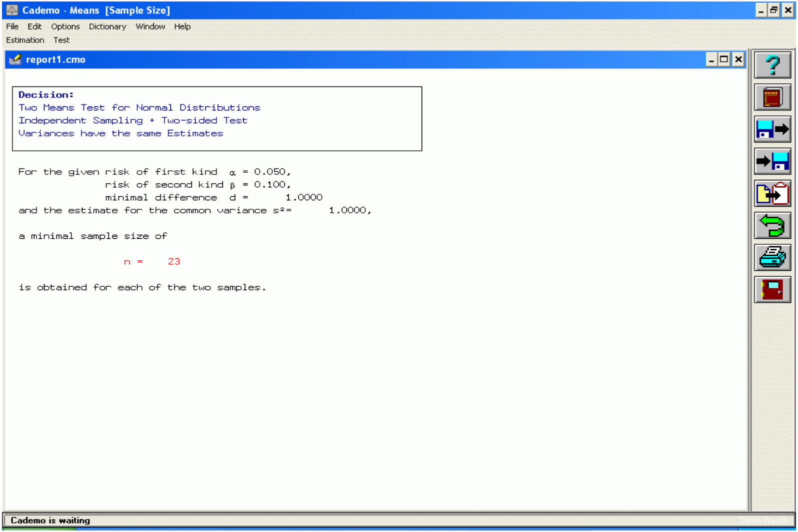
Which sample size for pre-tests?

For the F-test (equality of variances) the sample size is calculated by CADEMO for a variance ratio of at least 1.5. We need 258 observations in each sample.

Sample size for comparing 2 means



Results (sample size for comparing 2 means)



Which sample size?

Which sample size?

Summarizing:
per sample we need:

Which sample size?

Summarizing:
per sample we need:

- 500 units for testing normality

Which sample size?

Summarizing:
per sample we need:

- 500 units for testing normality
- 258 units for testing homogeneity of variances

Which sample size?

Summarizing:
per sample we need:

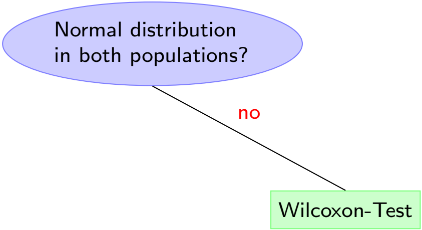
- 500 units for testing normality
- 258 units for testing homogeneity of variances
- 23 units for testing equality of means

What a nonsense!!

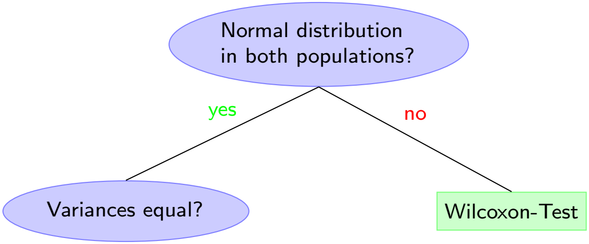
Test-algorithm

Normal distribution
in both populations?

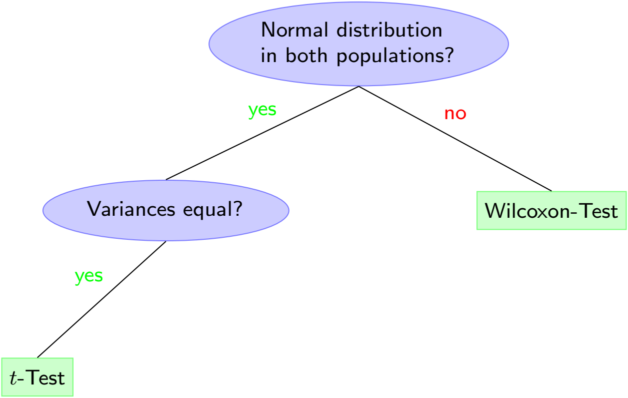
Test-algorithm



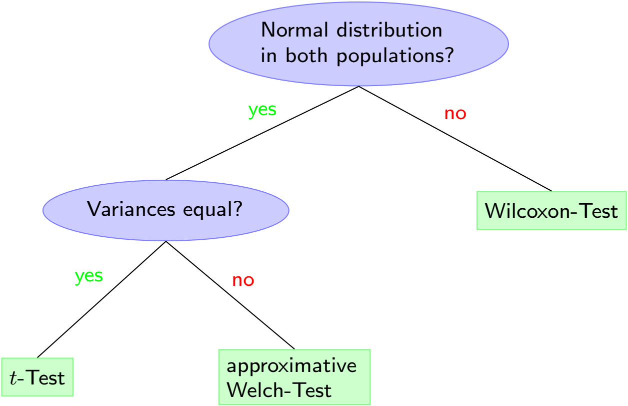
Test-algorithm



Test-algorithm



Test-algorithm



The risk of the second kind

The risk of the second kind is the most important one in pre-testing. Accepting a wrong null hypothesis leads to the wrong final test.

The risk of the second kind

The risk of the second kind is the most important one in pre-testing. Accepting a wrong null hypothesis leads to the wrong final test.

Simulation experiment:

In the case "always t – Test" is the actual first kind risk α_{act}

The risk of the second kind

The risk of the second kind is the most important one in pre-testing. Accepting a wrong null hypothesis leads to the wrong final test.

Simulation experiment:

In the case "always t – Test" is the actual first kind risk α_{act}

- "near to" the nominal one α_{nom}?

The risk of the second kind

The risk of the second kind is the most important one in pre-testing. Accepting a wrong null hypothesis leads to the wrong final test.

Simulation experiment:

In the case "always t – Test" is the actual first kind risk α_{act}

- "near to" the nominal one α_{nom}?
- "near to" means

$$|\alpha_{act} - \alpha_{nom}| \leq 0{,}2\alpha_{nom}$$

Skewness - Kurtosis

For the standardised 3. (Skewness) and 4. moment -3 (Kurtosis) of any distribution we have:

Skewness - Kurtosis

For the standardised 3. (Skewness) and 4. moment -3 (Kurtosis) of any distribution we have:

$$\gamma_2 \geq \gamma_1^2 - 2$$

Skewness - Kurtosis

For the standardised 3. (Skewness) and 4. moment -3 (Kurtosis) of any distribution we have:

$$\gamma_2 \geq \gamma_1^2 - 2$$

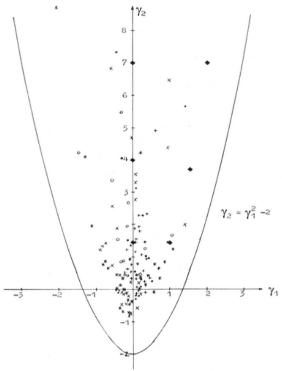
In the by $\gamma_2 \geq \gamma_1^2 - 2$ defined parabola we find all theoretical (and empirical) distributions with a fourth moment.

Skewness - Kurtosis

For the standardised 3. (Skewness) and 4. moment -3 (Kurtosis) of any distribution we have:

$$\gamma_2 \geq \gamma_1^2 - 2$$

In the by $\gamma_2 \geq \gamma_1^2 - 2$ defined parabola we find all theoretical (and empirical) distributions with a fourth moment.



Simulation

Simulated distributions

Simulation

Simulated distributions

- Let $\boldsymbol{u}$ $N(0;1)$.

Simulation

Simulated distributions

- Let $\boldsymbol{u}$ $N(0;1)$.
- By the transformation

$$\boldsymbol{y} = a + b\boldsymbol{u} + c\boldsymbol{u}^2 + d\boldsymbol{u}^3$$

we obtain a distribution at each point within the parabola.
Fleishman-System, Fleishman, 1978

Simulation

Simulated distributions

- Let $\boldsymbol{u}$ $N(0;1)$.
- By the transformation

$$\boldsymbol{y} = a + b\boldsymbol{u} + c\boldsymbol{u}^2 + d\boldsymbol{u}^3$$

we obtain a distribution at each point within the parabola.
Fleishman-System, Fleishman, 1978

Simulation experiment

- Each test was done 100 000 times with simulated data. The relative frequency of rejecting H_0 is an estimate of the power function.

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

and further:

$\alpha_{nom} = 0{,}01;\ 0{,}05$ and $0,10$

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

and further:

$\alpha_{nom} = 0{,}01;\ 0{,}05$ and $0,10$

$\delta/\sigma = (\mu_1 - \mu_2)/\sigma =$ 0; 1; 2; 3; 4 and 5;
$\sigma_1/\sigma_2 =$ 1, 2, . . . , 10.

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

and further:

$\alpha_{nom} = 0{,}01;\ 0{,}05$ and $0,10$

$\delta/\sigma = (\mu_1 - \mu_2)/\sigma =$ 0; 1; 2; 3; 4 and 5;
$\sigma_1/\sigma_2 =$ 1, 2, . . . , 10.

$n_1 = n_2 =$ 10; 30; 50; 100;

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

and further:

$\alpha_{nom} = 0{,}01;\ 0{,}05$ and $0,10$

$\delta/\sigma = (\mu_1 - \mu_2)/\sigma =$ 0; 1; 2; 3; 4 and 5;
$\sigma_1/\sigma_2 =$ 1, 2, . . . , 10.

$n_1 = n_2 =$ 10; 30; 50; 100;
$n_1 = 10;\ n_2 = 30;$

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

and further:

$\alpha_{nom} = 0{,}01;\ 0{,}05$ and $0,10$

$\delta/\sigma = (\mu_1 - \mu_2)/\sigma =$ 0; 1; 2; 3; 4 and 5;
$\sigma_1/\sigma_2 =$ 1, 2, . . . , 10.

$n_1 = n_2 =$ 10; 30; 50; 100;
$n_1 = 10;\ n_2 = 30;\qquad n_1 = 30;\ n_2 = 10;$

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

and further:

$\alpha_{nom} = 0{,}01;\ 0{,}05$ and $0,10$

$\delta/\sigma = (\mu_1 - \mu_2)/\sigma =$ 0; 1; 2; 3; 4 and 5;
$\sigma_1/\sigma_2 =$ 1, 2, . . . , 10.

$n_1 = n_2 =$ 10; 30; 50; 100;
$n_1 = 10;\ n_2 = 30;\qquad n_1 = 30;\ n_2 = 10;$
$n_1 = 30;\ n_2 = 100;$

Simulation - parameter

We selected:

Type	Skewness	Kurtosis
0	0	0
1	0	15
2	0,5	15
3	1	15
4	3	15

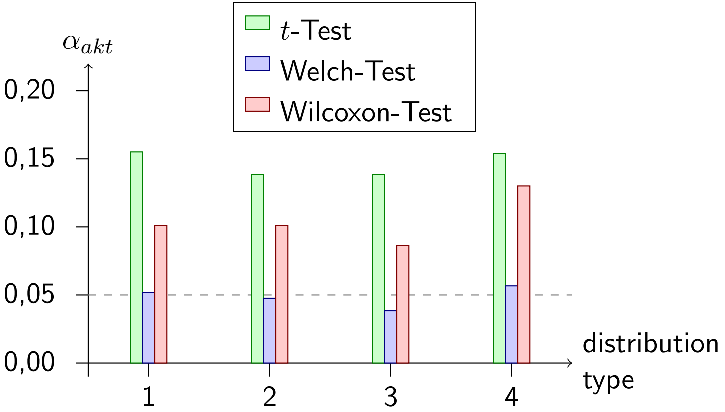
and further:

$\alpha_{nom} = 0{,}01;\ 0{,}05$ and $0,10$

$\delta/\sigma = (\mu_1 - \mu_2)/\sigma =$ 0; 1; 2; 3; 4 and 5;
$\sigma_1/\sigma_2 =$ 1, 2, . . . , 10.

$n_1 = n_2 =$ 10; 30; 50; 100;
$n_1 = 10;\ n_2 = 30;\qquad n_1 = 30;\ n_2 = 10;$
$n_1 = 30;\ n_2 = 100;\qquad n_1 = 100;\ n_2 = 30;$

Graphic of specific results (without pre-testing)



Empirical risk of the 1. kind for $t-$, Wilcoxon-, and Welch-test if H_0 is true and α_{nom} $=$ 0.05. The ratio of standard deviations for the first and second sample (σ_1, σ_2) are in ratio 1:2 (n_1=30, n_2=10).

Graphic of specific results (pre-testing)

Empirical risk of the 1. kind for $t-$ and Wilcoxon-test with pre-testing if H_0 is true and $\alpha_{nom} = 0.05$. The ratio of standard deviations for the first and second sample (σ_1, σ_2) are in ratio 1:2 (n_1=30, n_2=10).

Results - conclusions

Results - conclusions

- Never do a pre-test.

Results - conclusions

- Never do a pre-test.
- Choose always the approximate Welch-Test.

Results - conclusions

- Never do a pre-test.
- Choose always the approximate Welch-Test.
- The Wilcoxon – Test is useless, if higher moments differ in both popolations,

Results - conclusions

- Never do a pre-test.
- Choose always the approximate Welch-Test.
- The Wilcoxon – Test is useless, if higher moments differ in both popolations,
- The t-test can also not be recommanded.

Literatur

- Aspin, A.,A. (1949), *Tables for use in comparisons whose accuracy involves two variances, seperately estimated*. Biometrika 36, 290-296
- Fleishman, A. J. (1978) *A method for simulating non-normal distributions*. Psychometrika, 43, 521-532
- Rasch, D. (1995); *Mathematische Statistik*, J. Ambrosius Barth, Leipzig; Wiley, Berlin.
- Rasch, D. and Guiard, V. (2004) *The Robustness of Parametric Statistical Methods*. Psychology Science, 46; 175-208.
- Rasch, D.; Teuscher, F. and Guiard, V. (2007). *How Robust are tests for two independent samples in case of ordered categorical data*? Journal of Statistical Planning and Inference 133, 2706-2720
- Rasch, D., Herrendörfer, G., Bock, J., Victor, N., and Guiard, V. (2008) *Verfahrensbibliothek Versuchsplanung und -auswertung*. 2. Auflage mit CD, R. Oldenbourg Verlag München-Wien
- Rasch, D.; Kubinger, K.D. and Moder, K. (2009) *The two-sample t-test: pre-testing its assumptions does not pay*. Journal of Statistical Theory and Practice, accepted
- Satterthwaite, F.E. (1946) *An approximate distribution of estimates of variance components*. Biometrics Bulletin 2, 110-114
- Trickett, W.H. and Welch, B.L. (1954), *On the comparisons of two means: Further discussion of iterative methods for calculating tables*. Biometrika 41, 361-374
- Welch, B.L. (1947), *The generalization of Students problem when several different population variances are involved*. Biometrika 34, 28-35