AUSTRIAN JOURNAL OF STATISTICS
Volume 36 (2007), Number 3, 179–188

How to keep the Type I Error Rate in ANOVA if Variances are Heteroscedastic

Karl Moder
Institute of Applied Statistics and Computing,
University of Natural Resources and Applied Life Sciences, Vienna

Abstract: One essential prerequisite to ANOVA is homogeneity of variances in underlying populations. Violating this assumption may lead to an increased type I error rate. The reason for this undesirable effect is due to the calculation of the corresponding F-value. A slightly different test statistic keeps the level α. The underlying distribution of this alternative method is Hotelling's T^2. As Hotelling's T^2 can be approximated by a Fisher's F-distribution, this alternative test is very similar to an ordinary analysis of variance.

Zusammenfassung: Eine wesentliche Voraussetzung der Varianzanalyse ist Homoskedastizität in den zu Grunde liegenden Populationen. Eine Verletzung dieser Annahme führt zu einer erhöhten Typ 1 Fehlerrate. Der Grund für diesen unerwünschten Effekt liegt in der Berechnung des entsprechenden F-Wertes. Eine leicht veränderte Teststatistik hält das Niveau α. Die zu Grunde liegende Verteilung dieses alternativen Verfahrens ist Hotelling's T^2. Da Hotelling's T^2 durch eine F-Verteilung approximiert werden kann, ist der alternative Test sehr ähnlich einer normalen Varianzanalyse.

Keywords: ANOVA, Heteroscedasticity, Hotelling's T-squared, Levene-Test.

1 Introduction

ANOVA is one of the most frequently used methods in statistics. A correct application of this method depends on three preconditions: (i) independence of samples; (ii) normal distributed populations, and (iii) homoscedasticity. Dependence can be eliminated by an appropriate model. The effects of non-normal distributed data on significance level are low (see Box and Andersen, 1955) and can be ignored in most cases (see Lindman, 1992). Inhomogeneity of variances however infects α as well as test efficiency. Although Box (1954a) reported only little influence on this error rate with small differences in variances, Box and Andersen (1955) found the effect of unequal variances to be appreciable even when the ratio of block variances is moderate. In a second study Box (1954b) investigated effects of inequality of variance in the two-way classification. For an assumed variance ratio of main effects $1 : \cdots : 1 : 3$ a type I error rate of about 7% was found. In many practical trials variance ratio is much broader and exceeds this values. As for example in Figure 1.

A method proposed by Nelson and Dudewicz (2002) is applicable to such situations, but hypothesis differs from that of analysis of variance and a new test statistic has to be used. Transformation of data (e.g. $\log$-, $\arcsin$-, ..., transformation) is another often used practice in situations where variances are inhomogeneous. In a one factorial experiment,

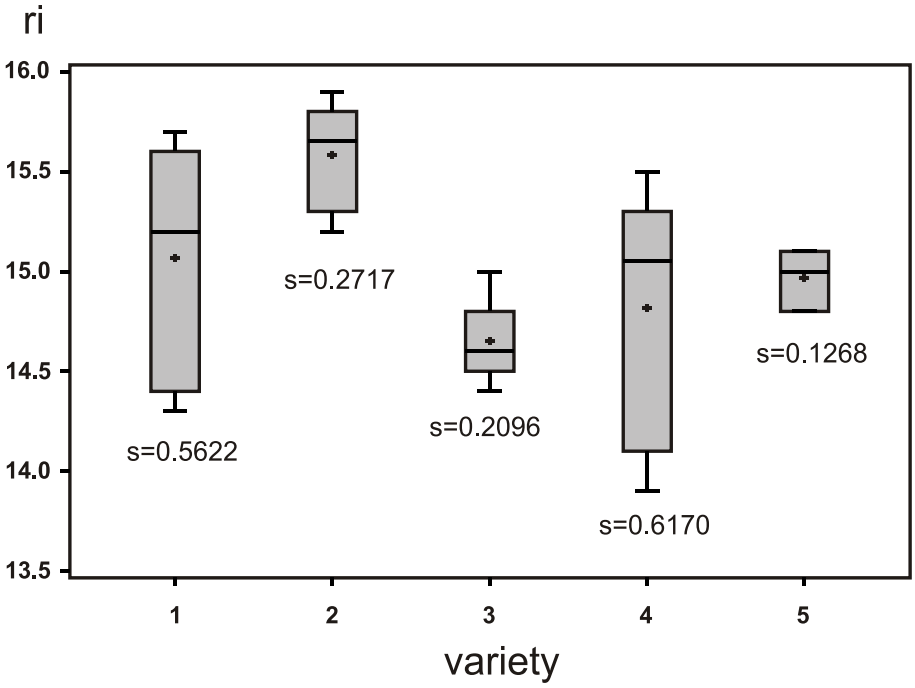

Figure 1: Boxplots for refraction index of apple juice (6 apples per variety), gathered at Landesversuchszentrum Haidegg ($s_1 : \cdots : s_5 = 4.4 : 2.1 : 1.7 : 4.9 : 1$).

this may be useful, if standard deviations are bound to the height of the means. In multi factorial analysis of variance transformations of that kind are not appropriate because of problems with interpretation of parameters and probabilities. In this article a method similar to the analysis of variance with identical hypothesis is introduced and the impacts of inhomogeneous variances on the test of main effects are examined. Type I error rate as well as test efficiency is checked by means of a simulation study.

2 Another View on the F-Ratio

As a very simple case of ANOVA a block analysis is used (although the method is applicable to more complicated situations). An appropriate model looks like

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}\,, \qquad e_{ij} \sim N(0, \sigma_i^2)\,, \quad i = 1, \ldots, I\,; \quad j = 1, \ldots, J\,,$$

where

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0\,.$$

Here x_{ij} is the observation on factor A at level i and factor B at level j, α_i denotes the effect of level i of factor A (treatment effect), β_j the effect of level j of factor B (block effect), e_{ij} is a random effect associated with x_{ij}, I stands for the number of levels of A (number of treatments) and J for the number of levels of B (number of blocks).

An appropriate test statistic for the hypothesis of interest $H_0 : \alpha_1 = \cdots = \alpha_I = 0$ can be calculated as $F = MS_A/MS_E$ with $df_A = I - 1$ and $df_E = (I-1)(J-1)$

degrees of freedom. $MS_A = SS_A/df_A$ is the mean square value for the interesting factor A, SS_A its sum of squares value and df_A its degrees of freedom, $MS_E = SS_E/df_E$ is the mean square value for the error term, SS_E its sum of squares value and df_E its degrees of freedom.

Let $\overline{x}_{..} = \sum_{i=1}^{I} \overline{x}_{i.}/I$, then SS_A can be calculated as

$$\begin{aligned}
SS_A &= J\sum_{i=1}^{I}(\overline{x}_{i.} - \overline{x}_{..})^2 = J\sum_{i=1}^{I}\left(\overline{x}_{i.} - \frac{1}{I}\sum_{i=1}^{I}\overline{x}_{i.}\right)^2 \\
&= J\left(\overline{x}_{1.}^2 + \frac{1}{I^2}\sum_{i=1}^{I}\overline{x}_{i.}^2 - \frac{2}{I}\overline{x}_{1.}\sum_{i=1}^{I}\overline{x}_{i.} + \frac{2}{I^2}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\overline{x}_{i.}\overline{x}_{i^*.}\right. \\
&\qquad\qquad \vdots \\
&\qquad \left. +\overline{x}_{I..}^2 + \frac{1}{I^2}\sum_{i=1}^{I}\overline{x}_{i..}^2 - \frac{2}{I}\overline{x}_{I.}\sum_{i=1}^{I}\overline{x}_{i.} + \frac{2}{I^2}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\overline{x}_{i.}\overline{x}_{i^*.}\right) \\
&= J\left(\sum_{i=1}^{I}\overline{x}_{i.}^2 + \frac{I}{I^2}\sum_{i=1}^{I}\overline{x}_{i.}^2 - \frac{2}{I}\sum_{i=1}^{I}\overline{x}_{i.}^2 - \frac{2}{I}\sum_{i=1}^{I}\sum_{i\neq i^*}\overline{x}_{i.}\overline{x}_{i^*.} + \frac{2I}{I^2}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\overline{x}_{i.}\overline{x}_{i^*.}\right) \\
&= J\left(\frac{I-1}{I}\sum_{i=1}^{I}\overline{x}_{i.}^2 - \frac{2}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\overline{x}_{i.}\overline{x}_{i^*.}\right) = \frac{J}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}(\overline{x}_{i.} - \overline{x}_{i^*.})^2 \\
&= \frac{J}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\bar{d}_{ii^*.}^2,
\end{aligned}$$

where $\bar{d}_{ii^*.}$ denotes the mean difference between level i and i^* of factor A. This means that SS_A is the sum of squares for each possible difference between means of factor A.

Now let $\overline{x}_{.j} = \sum_{i=1}^{I} x_{ij}/I$, then the sum of squares for factor B (blocks) can be calculated as

$$\begin{aligned}
SS_B &= I\sum_{j=1}^{J}(\overline{x}_{.j} - \overline{x}_{..})^2 = I\sum_{j=1}^{J}\left(\frac{1}{I}(x_{1j} - \overline{x}_{1.}) + \cdots + \frac{1}{I}(x_{Ij} - \overline{x}_{I.})\right)^2 \\
&= \frac{1}{I}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \overline{x}_{i.})^2 + 2\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\sum_{j=1}^{J}(x_{ij} - \overline{x}_{i.})(x_{i^*j} - \overline{x}_{i^*.})\right) \\
&= \frac{1}{I}\left(\sum_{i=1}^{I}SS_i + 2\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}SP_{ii^*}\right),
\end{aligned}$$

where SP_{ii^*} is the sum of crossproducts for level i and i^* of the factor A.

Utilizing

$$\overline{x}_{..}^2 = \frac{1}{I^2}\left(\sum_{i=1}^{I}\overline{x}_{i.}\right)^2 = \frac{1}{I^2}\sum_{i=1}^{I}\overline{x}_{i.}^2 + \frac{2}{I^2}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\overline{x}_{i.}\overline{x}_{i^*.}$$

we further get for the total sum of squares

$$\begin{aligned}
SS_T &= \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-\overline{x}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}^2 - IJ\overline{x}_{..}^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}^2 - \frac{IJ}{I^2}\sum_{i=1}^{I}\overline{x}_{i.}^2 - 2\frac{IJ}{I^2}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\overline{x}_{i.}\overline{x}_{i^*.} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}^2 - J\sum_{i=1}^{I}\overline{x}_{i}^2 + (I-1)\frac{J}{I}\sum_{i=1}^{I}\overline{x}_{i.}^2 - 2\frac{J}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\overline{x}_{i.}\overline{x}_{i^*.} \\
&= \sum_{i=1}^{I}SS_i + \frac{J}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}(\overline{x}_i-\overline{x}_{i^*})^2\,.
\end{aligned}$$

Finally, the sum of squares for the error term (SS_E) can be calculated as

$$\begin{aligned}
SS_E &= SS_T - SS_A - SS_B \\
&= \sum_{i=1}^{I}SS_i + \frac{J}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}(\overline{x}_i-\overline{x}_{i^*})^2 - \frac{J}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}(\overline{x}_{i.}-\overline{x}_{i^*.})^2 \\
&\quad -\frac{1}{I}\left(\sum_{i=1}^{I}SS_i + 2\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}SP_{ii^*}\right) \\
&= \frac{1}{I}\left((I-1)\sum_{i=1}^{I}SS_i - 2\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}SP_{ii^*}\right) \\
&= \frac{1}{I}\left(SS_1 - 2SP_{12} + SS_2 + \cdots + SS_{I-1} - 2SP_{I-1,I} + SS_I\right) \\
&= \frac{1}{I}\left(\sum_{j=1}^{J}\big((x_{1j}-\overline{x}_{1.})-(x_{2j}-\overline{x}_{2.})\big)^2 + \cdots + \sum_{j=1}^{J}\big((x_{I-1,j}-\overline{x}_{I-1.})-(x_{I,j}-\overline{x}_{I.})\big)^2\right) \\
&= \frac{1}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\sum_{j=1}^{J}\big((x_{ij}-x_{i^*j})-(\overline{x}_{i.}-\overline{x}_{i^*.})\big)^2 = \frac{1}{I}\sum_{i=1}^{I-1}\sum_{i^*=i+1}^{I}\sum_{i=1}^{J}(d_{ii^*j}-\bar{d}_{ii^*.})^2\,.
\end{aligned}$$

This means that SS_E is calculated as a squared sum of all individual differences between observations of two samples each, minus the according mean difference for all combinations of samples.

The interesting F-value results as a pooled estimation of all squared mean differences divided by a pooled value of individual differences for observations of two samples each for all combinations of samples. In a heteroscedastic situation this pooling is responsible for an enhanced type I error rate.

For the paired t-test homogeneity of variances is of no interest, as there is only one variable created from two dependent ones. By replacing the pooled sum of differences with a sum of individual paired differences, we find a test statistic which is Hotelling's T^2 distributed (see Hotelling, 1947) as the counterpart of Student's paired t-value.

3 Hotelling's T^2

Hotelling's T^2 for a single group of samples is calculated as

$$T^2 = J(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \,,$$

where J is the number of observations within each sample, $\bar{\boldsymbol{X}}$ denotes the sample mean vector of I elements and $\boldsymbol{\mu}$ the mean parameter vector of I elements, and $\boldsymbol{S}$ is the $(I \times I)$ sample covariance matrix.

The null hypothesis is formulated as $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$. As mentioned above, the analysis of variance is a test based on all possible pairwise mean differences. For this situation Hotelling's T^2 can be easily applied:

1. $\boldsymbol{\mu}$ is replaced by $\mathbf{0}$, a vector of zeros.
2. $\bar{\boldsymbol{X}}$ is replaced by $\bar{\boldsymbol{D}}$, a vector of all $I(I-1)/2$ possible pairwise mean differences of a factor, respectively the vector of all $I-1$ independent pairwise mean differences leading to equivalent results, i.e. $\bar{\boldsymbol{D}}' = (\bar{x}_1 - \bar{x}_2, \bar{x}_2 - \bar{x}_3, \ldots, \bar{x}_{I-1} - \bar{x}_I)$.
3. $\boldsymbol{S}$ is calculated from all samples of individual differences corresponding to the mean difference vector $\bar{\boldsymbol{D}}$. As individual differences include covariances between particular samples, it is not necessary to calculate off-diagonal elements in $\boldsymbol{S}$. Thus, it is sufficient to calculate $\boldsymbol{S}$ as a matrix of individual variances of sample differences. For independent differences, $\boldsymbol{S}$ looks like $\boldsymbol{S} = \text{diag}(s^2_{d_{12}}, \ldots, s^2_{d_{I-1,I}})$, where $s^2_{d_{i-1,i}}$ is the variance of the differences between all mean adjusted observations in samples $i-1$ and i, for $i = 2, \ldots, I$.

As a consequence, Hotelling's $\boldsymbol{T^2}$ simplifies to

$$T^2 = J \bar{\boldsymbol{D}}' \boldsymbol{S}^{-1} \bar{\boldsymbol{D}} \,,$$

which is $T^2_{I-1,J-1}$ distributed. Probability levels for T^2 can be found by approximating

$$\frac{J-I+1}{(J-1)(I-1)} T^2 \sim F_{I-1,J-I+1} \,.$$

4 Simulation Results

By means of a simulation study the impacts of inhomogeneous variances on the empirical type I error rate with a given $\alpha = 0.05$ were investigated. Figures 2 to 8 are based on $8 \times 8 = 64$ simulation configurations with 10000 runs each (treatment factor $i = 1, \ldots, I$ with number of levels $I = 3, \ldots, 8$, block factor $j = 1, \ldots, J$ with number of replications $J = 3, \ldots, 8$). Software packages `R` and `SAS` were used for these purposes. The errors e_{ij} were generated from $N(0, \sigma_i^2)$.

With homoscedastic variances both procedures meet the α-level. This is not true for the analysis of variance as soon as there are differences in the σ_i^2-levels. From Figure 2 we find that the empirical significance level when the ratio of the true standard deviations is $\sigma_1 : \sigma_2 : \cdots : \sigma_v = 3 : 1 : \cdots : 1$ rises up to 12% (depending on the number of factor levels), whereas the alternative test keeps the predefined value of α.

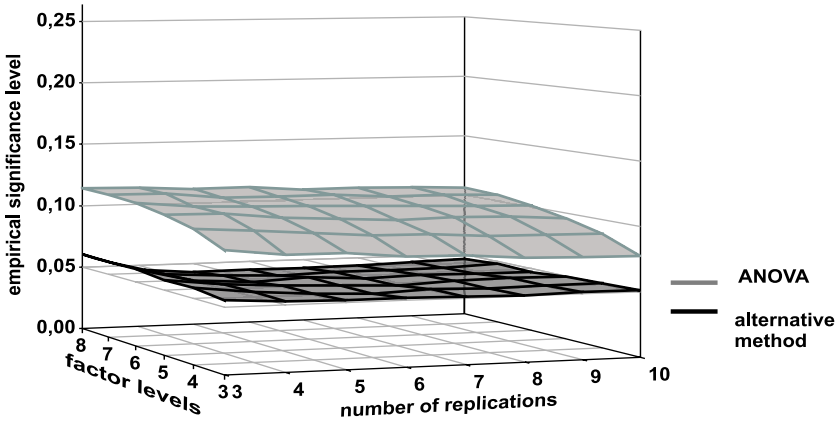

Figure 2: Empirical significance levels when the ratio of the true standard deviations is $\sigma_1 : \sigma_2 : \cdots : \sigma_v = 3 : 1 : \cdots : 1$ ($\alpha = 0.05$)

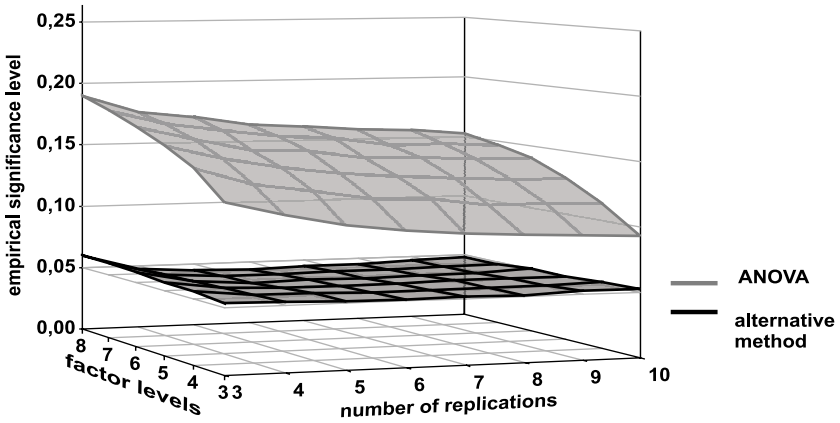

Figure 3: Empirical significance levels when the ratio of the true standard deviations is $\sigma_1 : \sigma_2 : \cdots : \sigma_v = 6 : 1 : \cdots : 1$ ($\alpha = 0.05$)

For Figure 3 the ratio of standard deviations is wider ($6 : 1 : \cdots : 1$) than for Figure 2. As a consequence the type I error rate rises up to 18% for ANOVA. Maybe the results in Figure 1 reflect this situation, as for ANOVA the null hypothesis is rejected ($p = 0.0136$), whereas the alternative method does not reject the null ($p = 0.1727$).

5 Power Comparison

An important question which arises with all kinds of tests concerns test efficiency. In the following figures several situations for a true alternative hypothesis with different variance ratios were investigated.

Figure 4 shows a higher power for analysis of variance if all variances are homogeneous, especially with a low number of replications. In Figure 5 the power of ANOVA seems to be superior to that of the alternative method. The apparent advantages are partly due to an enhanced type I error rate. This means, that a lot of significant results are not caused by differences in factor levels, but on random influences.

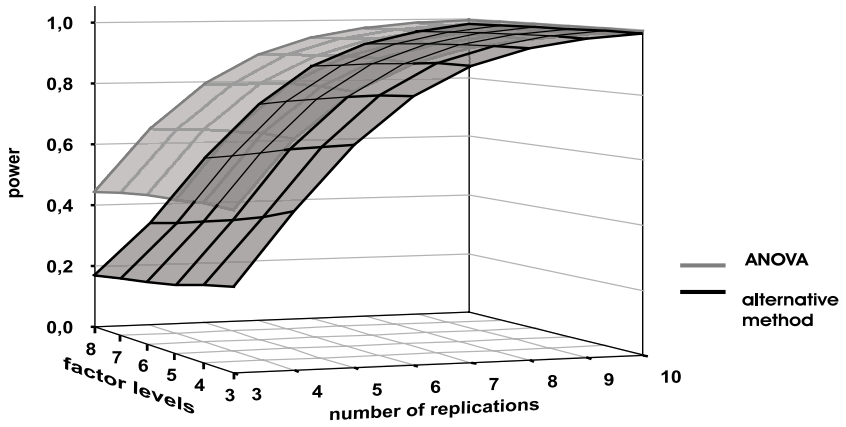

Figure 4: Power functions of ANOVA and Hotelling's T^2 test ($\alpha = 0.05$). Data of the first group are from $N(2, 1)$, for all other groups from $N(0, 1)$.

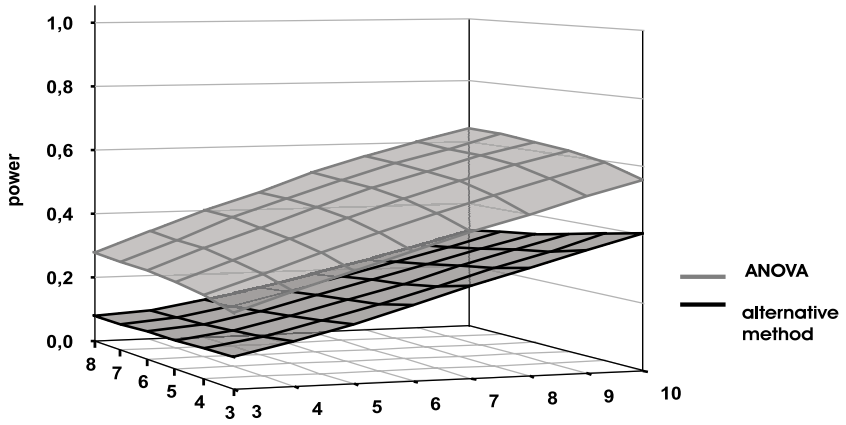

Figure 5: Power functions of ANOVA and Hotelling's T^2 test ($\alpha = 0.05$). Data of the first group are from $N(2, 3^2)$, for all other groups from $N(0, 1)$.

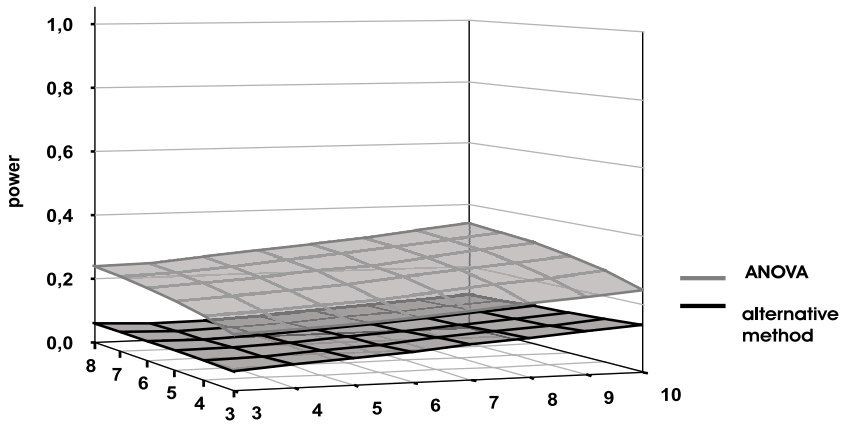

Figure 6: Power functions of ANOVA and Hotelling's T^2 test ($\alpha = 0.05$). Data of the first group are from $N(2, 6^2)$, for all other groups from $N(0, 1)$.

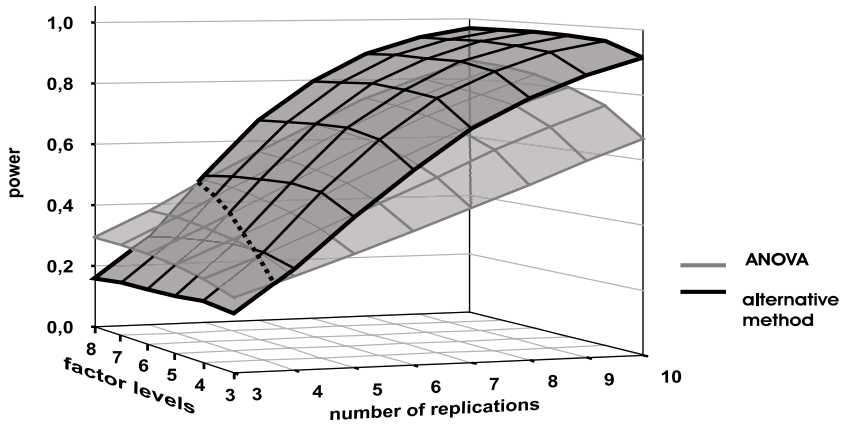

Figure 7: Power functions of ANOVA and Hotelling's T^2 test ($\alpha = 0.05$). Data of the first group are from $N(0, 3^2)$, of the second group from $N(2, 1)$, and of all other groups from $N(0, 1)$.

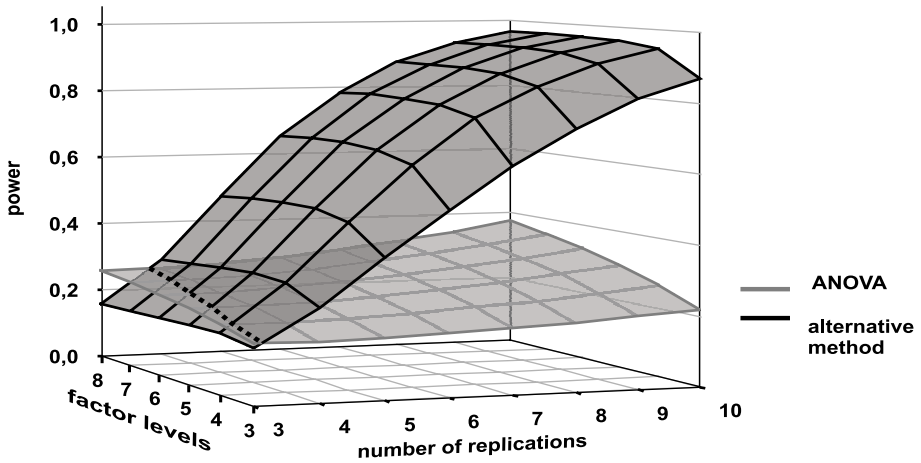

Figure 8: Power functions of ANOVA and Hotelling's T^2 test ($\alpha = 0.05$). Data of the first group are from $N(0, 6^2)$, of the second group from $N(2, 1)$, and of all other groups from $N(0, 1)$.

Figure 6 shows comparable results to those in Figure 5, but it is difficult to find any differences in the factor levels even if the number of observations is high. Whereas in Figures 5 and 6 the factor level with the largest effect was bound to the largest standard deviation this is not true in the following.

If the largest level of the factor does not correspond to the level with the largest standard deviation as in Figure 7, the alternative method is superior to ANOVA in most situations.

If heteroscedasticity is high (as in Figure 8), significant results of ANOVA are similar to that of Figure 3. This means that it is almost impossible to find differences in factor levels even if the sample size is large. However, the alternative method shows a large power.

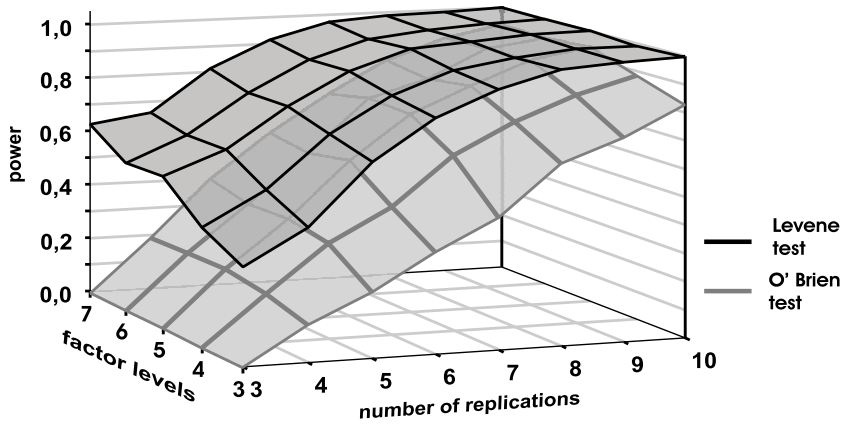

Figure 9: Power functions of Levene and O'Brien tests with a ratio of true standard deviations $\sigma_1 : \sigma_2 : \cdots : \sigma_v = 1 : 7 : 5 : 3 : 2 : 4 : 6$ ($\alpha = 0.05$).

6 Tests on Homogeneity of Variances

There are various different tests on homogeneity of variances available (Conover, Johnson, and Johnson, 1981). Levene's test (Levene, 1960) is one of the most popular ones. O'Brien's test (see O'Brien, 1979) is a modification for Levene's test, which is believed to be one of the most sensitive ones (Abdi, 2007) especially with platycurtic distributions (Algina, Olejnik, and Ocanto, 1989). In a simulation study with 1000 runs each, these tests are investigated.

The simulation is performed in such a way, that with 3 levels of the factor the ratio of standard deviations is $1 : 7 : 5$. When there are 4 levels this ratio is set to $1 : 7 : 5 : 3$. Following this strategy the ratio of standard deviations for 7 factor levels is set to $1 : 7 : 5 : 3 : 2 : 4 : 6$. The power functions of these tests are shown in Figure 9. In case of normal distributed data Levene's test performs better than O'Briens. But no matter which of these tests is used, there is a relatively high risk to oversee inhomogeneous variances even with a wide ratio of standard deviations.

7 Conclusions

Heteroscedasticity can be found in a lot of practical trials. The consequences of such a situation in concern to analysis of variance are subsumed in the following:

- ANOVA leads to an enhanced type I error rate, if variances are non-homogeneous.
- An alternative test based on Hotelling's T^2 keeps the α-level independently from the variance ratio.
- As soon as a factor effect comes with an enhanced standard deviation, the power of each test is very low.
- If the enhanced standard deviation is not bound to an enhanced factor effect, the alternative method shows very large power compared to ANOVA.
- If variances are homogeneous, ANOVA shows larger power than the alternative.
- Tests on homogeneity of variances show only low power. If there are doubts concerning homogeneity of variances, an alternative procedure is preferable.

References

Abdi, H. (2007). O'brien test for homogeneity of variance. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics.* Thousand Oaks, CA: Sage.

Algina, J., Olejnik, S., and Ocanto, R. (1989). Type I error rates and power estimates for selected two-sample tests of scale. *Annals of Educational Statistics*, *14*(4), 373-384.

Box, G. (1954a, June). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*(2), 290-302.

Box, G. (1954b, September). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. effect of inequality of variances and of correlation of errors in the two-way classification. *Annals of Mathematical Statistics*, *25*(3), 484-498.

Box, G., and Andersen, L. (1955). Theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society (Series B)*, *17*, 1-34.

Conover, W., Johnson, M., and Johnson, M. (1981, November). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(4), 351-361.

Hotelling, H. (1947). Multivariate quality control. In C. Eisenhart, M. W. Hastay, and W. A. Wallis (Eds.), *Techniques of statistical analysis.* New York: McGraw-Hill.

Levene, H. (1960). Robust tests for the equality of variance. *Contributions to Probability and Statistics*, 278-292.

Lindman, H. R. (1992). *Analysis of variance in experimental design* (S. Fienberg and I. Olkin, Eds.). Springer-Verlag New York, Inc.

Nelson, P. R., and Dudewicz, E. J. (2002). Exact analysis of means with unequal variances. *Technometrics*, *44*(2), 152-160.

O'Brien, R. G. (1979). A general anova method for robust tests of additive models for variances. *Journal of the American Statistical Association*, *74*, 877-880.

Author's address:

Karl Moder
Department of Landscape, Spatial and Infrastructure Sciences
Institute for Applied Statistics and Computing
University of Natural Resources and Applied Life Sciences
Gregor-Mendel-Strasse 33
A-1180 Vienna
Austria

Tel. +43 1 47654 / 5062
Fax +43 1 47654 / 5069
E-mail: karl.moder@boku.ac.at