

Methodology and Application of One-way ANOVA

Eva Ostertagová¹, Oskar Ostertag^{2,*}

¹Department of Mathematics and Theoretical Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Nemcovej 32, 042 00 Košice, Slovak republic

²Department of Applied Mechanics and Mechatronics, Faculty of Mechanical Engineering, Technical University of Košice, Letná 9, 042 00 Košice, Slovak republic

*Corresponding author: oskar.ostertag@tuke.sk

Received October 15, 2013; Revised October 28, 2013; Accepted November 13, 2013

Abstract This paper describes the powerful statistical technique one-way ANOVA that can be used in many engineering and manufacturing applications and presents its application. This technique is intended to analyze variability in data in order to infer the inequality among population means. The application data were analyzed using computer program MATLAB that performs these calculations.

Keywords: one-way ANOVA test, normality tests, homoscedasticity tests, multiple comparison tests, MATLAB

Cite This Article: Eva Ostertagová, and Oskar Ostertag, "Methodology and Application of One-way ANOVA." *American Journal of Mechanical Engineering* 1, no. 7 (2013): 256-261. doi: 10.12691/ajme-1-7-21.

1. Introduction

Analysis of variance (ANOVA) is a statistical procedure concerned with comparing means of several samples. It can be thought of as an extension of the *t*-test for two independent samples to more than two groups. The purpose is to test for significant differences between class means, and this is done by analysis the variances.

The ANOVA test of the hypothesis is based on a comparison of two independent estimates of the population variance [3].

When performing an ANOVA procedure the following assumptions are required:

- The observations are independent of one another.
- The observations in each group come from a normal distribution.
- The population variances in each group are the same (homoscedasticity).

ANOVA is the most commonly quoted advanced research method in the professional business and economic literature. This technique is very useful in revealing important information particularly in interpreting experimental outcomes and in determining the influence of some factors on other processing parameters.

The original ideas of analysis of variance were developed by the English statistician Sir Ronald A. Fisher (1890-1962) in his book "Statistical Methods for Research Workers" (1925). Much of the early work in this area dealt with agricultural experiments [1].

2. One-way ANOVA Test Procedure

The simplest case is one-way ANOVA. A one-way analysis of variance is used when the data are divided into groups according to only one factor.

Assume that the data $x_{11}, x_{12}, x_{13}, \dots, x_{1n_1}$ are sample from population 1, $x_{21}, x_{22}, x_{23}, \dots, x_{2n_2}$ are sample from population 2, \dots , $x_{k1}, x_{k2}, x_{k3}, \dots, x_{kn_k}$ are sample from population k . Let x_{ij} denote the data from the i^{th} group (level) and j^{th} observation.

We have values of independent normal random variables X_{ij} , $i=1, 2, \dots, k$ and $j=1, 2, \dots, n_i$ with mean μ_i and constant standard deviation σ , $X_{ij} \sim N(\mu_i, \sigma)$. Alternatively, each $X_{ij} = \mu_i + \varepsilon_{ij}$ where ε_{ij} are normally distributed independent random errors, $\varepsilon_{ij} \sim N(0, \sigma)$. Let $N = n_1 + n_2 + \dots + n_k$ is the total number of observations (the total sample size across all groups), where n_i is sample size for the i^{th} group.

The parameters of this model are the population means $\mu_1, \mu_2, \dots, \mu_k$ and the common standard deviation σ .

Using many separate two-sample *t*-tests to compare many pairs of means is a bad idea because we don't get a *p*-value or a confidence level for the complete set of comparisons together.

We will be interested in testing the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k \quad (1)$$

against the alternative hypothesis

$$H_1 : \exists 1 \leq i, l \leq k : \mu_i \neq \mu_l \quad (2)$$

(there is at least one pair with unequal means).

Let \bar{x}_i represent the mean sample i ($i=1, 2, \dots, k$):

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad (3)$$

$\bar{\bar{x}}$ represent the grand mean, the mean of all the data points:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}, \tag{4}$$

s_i^2 represent the sample variance:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \tag{5}$$

and $s^2 = MSE$ is an estimate of the variance σ^2 common to all k populations,

$$s^2 = \frac{1}{N - k} \sum_{i=1}^k (n_i - 1) \cdot s_i^2. \tag{6}$$

ANOVA is centered around the idea to compare the variation between groups (levels) and the variation within samples by analyzing their variances.

Define the total sum of squares SST , sum of squares for error (or within groups) SSE , and the sum of squares for treatments (or between groups) SSC :

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2, \tag{7}$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1) \cdot s_i^2, \tag{8}$$

$$SSC = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^k n_i \cdot (\bar{x}_i - \bar{x})^2. \tag{9}$$

Consider the deviation from an observation to the grand mean written in the following way:

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}). \tag{10}$$

Notice that the left side is at the heart of SST , and the right side has the analogous pieces of SSE and SSC . It actually works out that:

$$SST = SSE + SSC. \tag{11}$$

The total mean sum of squares MST , the mean sums of squares for error MSE , and the mean sums of squares for treatment MSC are:

$$MST = \frac{SST}{df(SST)} = \frac{SST}{N - 1}, \tag{12}$$

$$MSE = \frac{SSE}{df(SSE)} = \frac{SSE}{N - k}, \tag{13}$$

$$MSC = \frac{SSC}{df(SSC)} = \frac{SSC}{k - 1}. \tag{14}$$

The one-way ANOVA, assuming the test conditions are satisfied, uses the following test statistic:

$$F = \frac{MSC}{MSE}. \tag{15}$$

Under H_0 this statistic has Fisher's distribution $F(k - 1, N - k)$. In case it holds for the test criteria

$$F > F_{1-\alpha, k-1, N-k}, \tag{16}$$

where $F_{1-\alpha, k-1, N-k}$ is $(1 - \alpha)$ -quantile of F -distribution with $k - 1$ and $N - k$ degrees of freedom, then hypothesis H_0 is rejected on significance level α [1,3].

The results of the computations that lead to the F -statistic are presented in an ANOVA table, the form of which is shown in the Table 1.

Table 1. Basic one-way ANOVA table

Variance source	Sum of squares SS	Degrees of freedom df	Mean square MS	F -statistic	Tail area above F
Between	SSC	$k - 1$	MSC	MSC/MSE	p -value
Within	SSE	$N - k$	MSE	—	—
Total	SST	$N - 1$	—	—	—

In statistical softwares is used to be in this table column with p -value. This p -value says the probability of rejection the null hypothesis in case the null hypothesis holds. In case $p < \alpha$, where α is chosen significance level, is the null hypothesis rejected with probability greater than $(1 - \alpha) \cdot 100\%$ probability.

3. Post Hoc Comparison Procedures

Post hoc comparisons (or post hoc tests, multiple comparison tests) are tests of the statistical significance of differences between group means calculated after ("post") having done ANOVA that shows an overall difference. Multiple comparison methods are designed to investigate differences between specific pairs of means. This provides the information that is of most use to the researcher.

One possible approach to the multiple comparison problem is to make each comparison independently using a suitable statistical procedure. For example, a statistical hypothesis test could be used to compare each pair of means, μ_I and μ_J , $I, J = 1, 2, \dots, k$; $I \neq J$, where the null and alternative hypotheses are of the form

$$H_0 : \mu_I = \mu_J, H_1 : \mu_I \neq \mu_J. \tag{17}$$

An alternative way to test for a difference between μ_I and μ_J is to calculate a confidence interval for $\mu_I - \mu_J$. A confidence interval is formed using a point estimate a margin of error, and the formula

$$(\text{point estimate}) \pm (\text{margin of error}). \tag{18}$$

The point estimate is the best guess for the value of $\mu_I - \mu_J$ based on the sample data. The margin of error reflects the accuracy of the guess based on variability in the data. It also depends on a confidence coefficient, which is often denoted by $1 - \alpha$. The interval is calculated by subtracting the margin of error from the point estimate to get the lower limit and adding the margin of error to the point estimate to get the upper limit [6].

If the confidence interval for $\mu_I - \mu_J$ does not contain zero (thereby ruling out that $\mu_I = \mu_J$), then the null hypothesis is rejected and μ_I and μ_J are declared different at level of significance α .

The multiple comparison tests for population means, as well as the F -test, have the same assumptions.

There are many different multiple comparison procedures that deal with these problems. Some of these

procedures are as follows: Fisher’s method, Tukey’s method, Scheffé’s method, Bonferroni’s adjustment method, Dunn-Šidák method. Some require equal sample sizes, while some do not. The choice of a multiple comparison procedure used with an ANOVA will depend on the type of experimental design used and the comparisons of interest to the analyst [8].

The Fisher (LSD) method essentially does not correct for the type 1 error rate for multiple comparisons and is generally not recommended relative to other options.

The Tukey (HSD) method controls type 1 error very well and is generally considered an acceptable technique. There is also a modification of the test for situation where the number of subjects is unequal across cells called the Tukey-Kramer test.

The Scheffé test can be used for the family of all pairwise comparisons but will always give longer confidence intervals than the other tests [6]. Scheffé’s procedure is perhaps the most popular of the post hoc procedures, the most flexible, and the most conservative.

There are several different ways to control the experiment-wise error rate. One of the easiest ways to control experiment-wise error rate is use the Bonferroni correction. If we plan on making m comparisons or conducting m significance tests the Bonferroni correction is to simply use α/m as our significance level rather than α . This simple correction guarantees that our experiment-wise error rate will be no larger than α . Notice that these results are more conservative than with no adjustment. The Bonferroni is probably the most commonly used post hoc test, because it is highly flexible, very simple to compute, and can be used with any type of statistical test (e.g., correlations), not just post hoc tests with ANOVA.

The Šidák method has a bit more power than the Bonferroni method. So from a purely conceptual point of view, the Šidák method is always preferred.

The confidence interval for $\mu_I - \mu_J$ is calculated using the formula:

$$\bar{x}_I - \bar{x}_J \mp t_{1-\alpha/2, N-k} \cdot \sqrt{s^2 \left(\frac{1}{n_I} + \frac{1}{n_J} \right)}, \quad (19)$$

where $t_{1-\alpha/2, N-k}$ is the quantile of the Student’s t -probability distribution, by Fisher method (LSD – Least Significant Difference);

$$\bar{x}_I - \bar{x}_J \mp q_{\alpha, k, N-k} \cdot \sqrt{\frac{s^2}{2} \left(\frac{1}{n_I} + \frac{1}{n_J} \right)}, \quad (20)$$

where $q_{\alpha, k, N-k}$ represents the quantile for the Studentized range probability distribution, by Tukey-Kramer method (HSD – Honestly Significant Difference);

$$\bar{x}_I - \bar{x}_J \mp \sqrt{(k-1) s^2 \left(\frac{1}{n_I} + \frac{1}{n_J} \right)} \cdot F_{1-\alpha, k-1, N-k}, \quad (21)$$

by Scheffé method;

$$\bar{x}_I - \bar{x}_J \mp t_{1-\alpha^*/2, N-k} \cdot \sqrt{s^2 \left(\frac{1}{n_I} + \frac{1}{n_J} \right)}, \quad (22)$$

where $\alpha^* = \frac{\alpha}{c}$, $c = \binom{k}{2}$ is the number of pairwise comparisons in the family, by Bonferroni method;

$$\bar{x}_I - \bar{x}_J \mp t_{1-\alpha^*/2, N-k} \cdot \sqrt{s^2 \left(\frac{1}{n_I} + \frac{1}{n_J} \right)}, \quad (23)$$

where $\alpha^* = 1 - (1 - \alpha)^{1/c}$ and $c = \binom{k}{2}$, by Dunn-Šidák method [2].

4. Tests for Homogeneity of Variances

Many statistical procedures, including analysis of variance, assume that the different populations have the same variance. The test for equality of variances is used to determine if thtion of equal variances is valid.

We will be interested in testing the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \quad (24)$$

against the alternative hypothesis

$$H_1 : \exists 1 \leq i, l \leq k : \sigma_i^2 \neq \sigma_l^2. \quad (25)$$

There are many testse assump of homogeneity of variances. Commonly used tests are the Bartlett (1937), Hartley (1940, 1950), Cochran (1941), Levene (1960), and Brown and Forsythe (1974) tests. The Bartlett, Hartley and Cochran are technically test of homogeneity. The Levene and Brown and Forsythe methods actually transform the data and then tests for equality of means.

Note that Cochran’s and Hartley’s test assumes that there are equal numbers of participants in each group.

The tests of Bartlett, Cochran, Hartley and Levene may be applied for number of samples $k > 2$. In such situation, the power of these tests turns out to be different. When the assumption of the normal distribution holds for $k > 2$ these tests may be ranked by power decrease as follows: Cochran } Bartlett } Hartley } Levene. This preference order also holds in case when the normality assumption is disturbed. An exception concerns the situations when samples belong to some distributions which have more heavy tails than the normal law. For example, in case of belonging samples to the Laplace distribution the Levene test turns out to be slightly more powerful than three others [7].

Bartlett’s test has the following test statistic:

$$B = c^{-1} \left((N - k) \cdot \ln s^2 - \sum_{i=1}^k (n_i - 1) \cdot \ln s_i^2 \right), \quad (26)$$

where constant $c = 1 + \frac{1}{3(k-1)} \cdot \left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{N - k} \right)$ and meaning of all the others symbols is evident (see section 2). The hypothesis H_0 is rejected on significance level α , when

$$B > \chi_{1-\alpha, k-1}^2 \quad (27)$$

where $\chi_{1-\alpha, k-1}^2$ is the critical value of the *chi-square* distribution with $k - 1$ degrees of freedom.

Cochran’s test is one of the best methods for detecting cases where the variance of one of the groups is much larger than that of the other groups. This test uses the following test statistic:

$$C = \frac{\max s_i^2}{\sum_{i=1}^k s_i^2} \tag{28}$$

The hypothesis H_0 is rejected on significance level α , when

$$C > C_{\alpha,k,n-1} \tag{29}$$

where critical value $C_{\alpha,k,n-1}$ is in special statistical tables.

Hartley’s test uses the following test statistic:

$$H = \frac{\max s_i^2}{\min s_i^2} \tag{30}$$

The hypothesis H_0 is rejected on significance level α , when

$$H > H_{\alpha,k,n-1} \tag{31}$$

where critical value $H_{\alpha,k,n-1}$ is in special statistical tables [2].

Originally Levene’s test was defined as the one-way analysis of variance on $z_{ij} = |x_{ij} - \bar{x}_i|$, the absolute residuals $x_{ij} - \bar{x}_i$, $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, n_i$, where k is the number of groups and n_i the sample size of the i^{th} group. The test statistic has Fisher’s distribution $F(k-1, N-k)$ and is given by:

$$F = \frac{(N-k) \sum_{i=1}^k n_i \cdot (z_i - \bar{z})^2}{(k-1) \sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2} \tag{32}$$

where $N = \sum_{i=1}^k n_i$, $\bar{z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}$, $\bar{z} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} z_{ij}$.

To apply the ANOVA test, several assumptions must be verified, including normal populations, homoscedasticity, and independent observations. The absolute residuals do not meet any of these assumptions, so Levene’s test is an approximate test of homoscedasticity [5].

Brown and Forsythe subsequently proposed the absolute deviations from the median \tilde{x}_i of the i^{th} group, so is $z_{ij} = |x_{ij} - \tilde{x}_i|$.

5. Example from Technical Practice

What follows is an example of the one-way ANOVA procedure using the statistical software package, MATLAB.

One important factor in selecting software for word processing and database management systems is the time required to learn how to use a particular system. In order to evaluate three database management systems, a firm

devised a test to see how many training hours were needed for six of its word processing operators to become proficient in each of three systems [9]. The data from this experiment are in the Table 2. Using a 5 % significance level, is there any difference between the training time needed for the three systems?

Table 2. Experiment data in hours

System 1	20	17	15	19	14	13
System 2	18	17	14	20	13	12
System 3	23	25	20	21	19	20

5.1. Testing the Assumption of Normality

One of the first steps in using the one-way ANOVA test is to test the assumption of normality. Even if the distribution is somewhat different from normal, one-way ANOVA can still work good if the sample sizes are large enough. However, when sample sizes are small, one-way ANOVA can be unreliable if the data in one or more of the groups comes from a highly non-normal distribution.

For evaluating normality there are graphical and statistical methods. For example normal probability plot is a graph specifically designed to check for normality. If the data comes from a normal distribution the points should form a line. The statistical methods include diagnostic hypothesis tests for normality, where the null hypothesis is that there is no significant departure from normality for each of the groups/levels. The alternative hypothesis is that there is a significant departure from normality. The main tests for the assessment of normality are Kolmogorov-Smirnov (K-S) test, Lilliefors test (corrected K-S test), Shapiro-Wilk test, Anderson-Darling test, Cramer-von Mises test, D’Agostino test and Jarque-Bera test.

For the above example we are using MATLAB with functions [4]:

- [h,p]=lillietest(x,0.05,'norm') for Lilliefors test,
- [h,p]=swtest(x,0.05) for Shapiro-Wilk test.

For example the Shapiro-Wilk test using significance level 0.05 give these results: $p = 0.6599$ for system 1, $p = 0.6643$ for system 2, and $p = 0.4044$ for system 3.

We would conclude that each of the levels of the independent variable are normally distributed.

5.2. Testing the Assumption of Homogeneity of Variances

We seek to test equality of variances (see part 4) and have run Bartlett’s test in MATLAB:

```
X=[20,17,15,19,14,13;18,17,14,20,13,12;23,25,20,21,...
19,20]';[p,stats]=vartestn(X)
```

From the following analysis in MATLAB, the p -value for Bartlett’s test (significance level 0.05 is here default) is $p = 0.7769 > \alpha = 0.05$.

Therefore, we would fail to reject the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2.$$

5.3. Hypothesis Testing Using ANOVA

Letting μ_1, μ_2 , and μ_3 be the mean for the three systems, the null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$. The alternative is $H_1 : \mu_i \neq \mu_l$ for at least one i, l pair ($i, l = 1, 2, 3$).

In MATLAB we use command:

```
[p,tbl,stats]=anova1(X)
```

This will return an ANOVA table, showing the value of the F -statistic and p -value, and a boxplot of three different groups. The results of the calculations for this case are summarized in Table 3 and Figure 1.

Table 3. Summary table of the one-way ANOVA for experiment data

Variance source	Sum of squares SS	Degrees of freedom df	Mean square MS	F -statistic	p -value
Between	115.1111	2	57.5556	7.5731	0.0053
Within	114.0000	15	7.6000	—	—
Total	229.1111	17	—	—	—

Since p -value is less than given significance level 0.05 for this problem, we reject the null hypothesis. There is a difference between the mean learning times for at least two of the three database management systems.

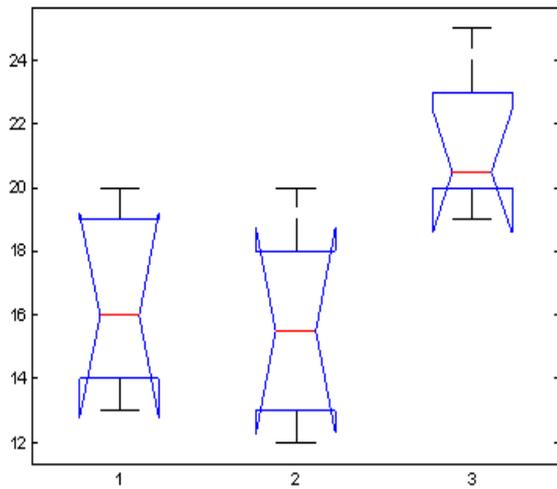


Figure 1. Boxplot of three different group

5.4. Pair-wise Comparison

When the null hypothesis is rejected using the F -test in ANOVA, we want to know where the difference among the means is. To determine which pairs of means are significantly different, and which are not, we can use the multiple comparison tests.

MATLAB implements for example the Tukey-Kramer procedure, the Bonferroni procedure, Dunn-Šidák procedure and reports the results in terms of the confidence interval.

Now we can make the 95 % confidence interval for differences in pair of population group means $\mu_I - \mu_J$, $I, J = 1, 2, 3$; $I \neq J$.

In MATLAB we use the following series of commands:
 multcompare(stats,'alpha',.05,'ctype','tukey-kramer')
 multcompare(stats,'alpha',.05,'ctype','bonferroni')
 multcompare(stats,'alpha',.05,'ctype','dunn-sidak')

The statistical outputs are, respectively, shown in Table 4, Table 5, Table 6 and Figure 2.

Table 4. Results using Tukey-Kramer method

pairs I, J	difference $\bar{x}_I - \bar{x}_J$	lower limit	upper limit
1, 2	0.6667	-3.4676	4.8009
1, 3 *	-5.0000	-9.3185	-0.8657
2, 3 *	-5.6667	-9.8009	-1.5324

Table 5. Results using Bonferroni method

pairs I, J	difference $\bar{x}_I - \bar{x}_J$	lower limit	upper limit
1, 2	0.6667	-3.6208	4.9541
1, 3 *	-5.0000	-9.2875	-0.7125
2, 3 *	-5.6667	-9.9541	-1.3792

Table 6. Results using Dunn-Šidák method

pairs I, J	difference $\bar{x}_I - \bar{x}_J$	lower limit	upper limit
1, 2	0.6667	-3.6073	4.9406
1, 3 *	-5.0000	-9.2740	-0.7260
2, 3 *	-5.6667	-9.9406	-1.3927

Figure 2 represents an interactive figure. By clicking on the group symbol at the bottom, in part of the figure is displayed the group from which the selected one statistically differs.

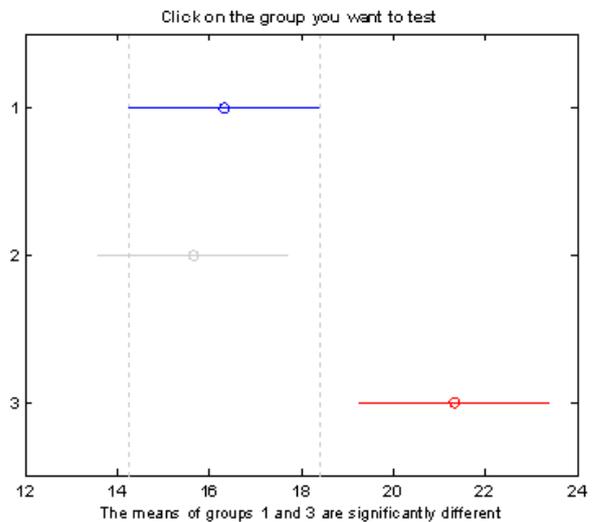


Figure 2. The interactive figure

Using all of the three multiple comparison methods, we discover that system 3 takes significantly longer to learn than systems 1 and 2 which are similar.

6. Conclusion

In many statistical applications in business administration, psychology, social science, and the natural sciences we need to compare more than two groups. For hypothesis testing more than two population means scientists have developed ANOVA method.

The ANOVA test procedure compares the variation in observations between samples (sum of squares for groups,

SSC) to the variation within samples (sum of squares for error, SSE). The ANOVA F -test rejects the null hypothesis that the mean responses are equal in all groups if SSC is large relative to SSE .

The analysis of variance assumes that the observations are normally and independently distributed with the same variance for each treatment or factor level [3]. If the normality assumption of the one-way ANOVA F -test is not met, we can use the Kruskal-Wallis rank test.

Acknowledgement

This article was created by implementation of the grant project VEGA no. 1/0102/11 Experimental methods and modeling techniques in-house manufacturing and non manufacturing processes.

References

- [1] Aczel, A.D., *Complete Business Statistics*, Irwin, 1989.
- [2] Brown, M., Forsythe, A., "Robust tests for the equality of variances," *Journal of the American Statistical Association*, 364-367. 1974.
- [3] Montgomery, D.C., Runger, G.C., *Applied Statistics and Probability for Engineers*, John Wiley & Sons, 2003.
- [4] Ostertagová, E., *Applied Statistic* (in Slovak), Elfa, Košice, 2011.
- [5] Parra-Frutos, I., "The behaviour of the modified Levene's test when data are not normally distributed," *Comput Stat*, Springer, 671-693. 2009.
- [6] Rafter, J.A., Abell, M.L., Braselton, J.P., "Multiple Comparison Methods for Means," *SIAM Review*, 44 (2). 259-278. 2002.
- [7] Rykov, V.V., Balakrishnan, N., Nikulin, M.S., *Mathematical and Statistical Models and Methods in Reliability*, Springer, 2010.
- [8] Stephens, L.J., *Advanced Statistics demystified*, McGraw-Hill, 2004.
- [9] Taylor, S., *Business Statistics*.www.palgrave.com.