Investigation of dependence variance analysis statistical distributions on error and random factor distribution laws*

B.Yu. Lemeshko, V.M. Ponomarenko

NSTU
630092, Novosibirsk, K.Marks, 20, Russia

Abstract –**The distribution of statistics used in single-factor variance analysis for testing hypothesis about variance have been investigated by statistical simulation methods. The case of normality assumption failure has been considered.**

The feather of statistical criteria concerned with hypotheses about means is robustness to the normality assumption failure [1]. The recent researches, for example [2-3], sustain it. On the other hand the criteria concerned with hypotheses about variances are very sensitive to failure of assumption about normality of random variable including to variance analysis model [1].

In practical situations the normal law is not always the best model for error or random levels of factor distribution. In such cases using the classical variance analysis results may lead to not valid statistical conclusions.

The purpose of the paper is to investigate distribution of variance analysis statistics by statistical simulation methods for different error or random factor level distribution laws.

The balanced variance analysis model with one random factor is under consideration. This model is given with

$$y_{ij} = \mu + a_i + e_{ij}, \qquad (1)$$

where I - is the number of randomly extracted levels of factor A, $i = 1,\dots,I$; J - is the number of experiences at each level, $j = 1,\dots,J$; μ - is the average of factor A effect; $\{a_i\}$, $\{e_{ij}\}$ are independent in aggregate and have zero average.

In the classical statement the assumptions of error and random factor normality are made. This assumptions look like:

$$\Omega : \begin{cases} \{a_i\} & \textit{are distributed by } N(0,\sigma_A^2) \\ \{e_{ij}\} & \textit{are distributed by } N(0,\sigma_e^2) \end{cases}, \qquad (2)$$

where σ_A^2 is a variance of $\{a_i\}$ which are random factor A level effects, σ_e^2 is a variance of errors $\{e_{ij}\}$.

In the general case about the random factor A the hypothesis which looks like

$$H_A: \ \sigma_A^2 \le \theta_0 \sigma_e^2, \ \theta_0 \ge 0, \qquad (3)$$

where θ_0 is some defined constant, is tested.

Most frequently the hypothesis (3) at $\theta_0 = 0$ is tested:

$$H_A: \ \sigma_A^2 = 0. \qquad (4)$$

If the hypothesis (4) is not rejected using this criterion than the influence of the random factor A effects on output y are significant.

At $\theta_0 = 1$ the same result means that influence of the random factor A effects on output y not exceeds influence of the errors on output y.

Other presented in this work situations ($\theta_0 = 0.5$ and $\theta_0 = 2$) can be interpreted in the same way as case $\theta_0 = 1$.

To test the hypothesis (3) the statistics (5) is used

$$S = \frac{1}{(1+J\theta_0)} \frac{\overline{SS}_A}{\overline{SS}_e}, \qquad (5)$$

where

$$\overline{SS}_A = \frac{J}{I-1} \sum_{i=1}^{I} \left(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}\right)^2,$$

$$\overline{SS}_e = \frac{1}{I(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{y}_{ij} - \overline{y}_{i\bullet}\right)^2,$$

$$\overline{y}_{i\bullet} = \frac{1}{J} \sum_{j=1}^{J} y_{ij}, \ \overline{y}_{\bullet\bullet} = \frac{1}{I} \sum_{i=1}^{I} \overline{y}_{i\bullet}.$$

By Sheffe [1] the statistic (5) at the limit submits the Fisher's F-distribution with $I-1$ and $I(J-1)$ degrees of freedom when the normality assumptions (2) and conditions of a hypothesis

$$\sigma_A^2 = \theta_0 \sigma_e^2 \qquad (6)$$

are hold.

* The research is supported by the Ministry of Education and Science of the RF (project No. 15378).

S statistic (5) distributions have been investigated for different model (1) error and random factor distributions: for the maximum value distribution

$$f\left(x,\theta_1,\theta_2\right)=\frac{1}{\theta_2}\exp\left\{-\frac{x-\theta_1}{\theta_2}-\exp\left(-\frac{x-\theta_1}{\theta_2}\right)\right\}$$

and for the symmetric distribution family (De) which density looks like

$$De(\lambda)=f(x,\theta_1,\theta_2,\lambda)=\frac{\lambda}{2\sqrt{2}\theta_2\Gamma(1/\lambda)}\exp\left\{-\left(\frac{|x-\theta_1|}{\sqrt{2}\theta_2}\right)^{\lambda}\right\}$$

The Laplace ($\lambda=1$) and normal ($\lambda=2$) distributions are the partial cases of the $De(\lambda)$ family.

While testing the hypothesis (4) the statistic (5) distributions have been simulated for different errors and random factor distributions for $I=5$, $J=6$. The resulting empirical distributions were compared with the Fisher's $F_{4,25}$ distribution to which the statistic distribution should submit in the normal case. In the Table 1 the high significance levels achieved as result of fitting test by several criteria are presented. The χ^2 Pearson's, Kolmogorov's, ω^2 Cramer-von-Mises-Smirnov's, Ω^2 Anderson-Darling's criteria are used. The hypothesis of fitting empirical statistic distribution, obtained by modeling under (4) validation and for different $\{e_{ij}\}$ distributions, with Fisher's $F_{4,25}$ distribution is tested. The high significance levels achieved, which are presented in table, are average for 10 experiments.

TABLE 1
THE RESULTS OF TESTING FITT EMPIRICAL STATIISTIC (5) DISTRIBUTIONS WITH $F_{4,25}$ DISTRIBUTION FOR HYPOTHESIS (4) UNDER TEST AND DIFFERENT ERROR DISTRIBUTIONS

Errors distribution	Fitting test	Significance levels achieved
De(1)	χ^2 Kolmogorov ω^2 Ω^2	0.046958 0.0743303 0.0646722 0.0385816
Maximum value	χ^2 Kolmogorov ω^2 Ω^2	0.0454898 0.0838117 0.0686183 0.0439787
Normal	χ^2 Kolmogorov ω^2 Ω^2	0.568523 0.601581 0.534741 0.494178
De(10)	χ^2 Kolmogorov ω^2 Ω^2	0.190879 0.385246 0.383669 0.294447

These results reveal that corresponding F distribution can be used to test hypothesis like (4) without danger of large mistakes in obtaining of significance levels achieved.

Fig. 1–3 demonstrate the statistic (5) distribution behavior when conditions of hypothesis (6) for $\theta_0=1$ are hold $\{a_i\}$ and $\{e_{ij}\}$ are not submit to normal law.

The errors and random factor effect distribution laws under which the statistic distribution approximation is receiving are specified in figure captions.

For example, $G(S(De(2),De(10)))$ means S statistic distribution approximation obtained in case when $\{e_{ij}\}$ are normally distributed, $\{a_i\}$ are distributed by $De(10)$ conditions of hypothesis under (4) validation.

Fig.1 illustrates case when $\{e_{ij}\}$ are normally distributed, $\{a_i\}$ are not normal. This figure shows that the statistic (5) distribution considerably differ from the classical $F_{4,25}$ distribution. By that the type I error increases when $\{a_i\}$ are distributed by $De(10)$. The type II error increases when $\{a_i\}$ are distributed by $De(1)$.

Fig.2 demonstrates the case when $\{a_i\}$ are normally distributed, $\{e_{ij}\}$ are not normal. In comparison with case presented in Fig.1 the kind of statistic (5) distribution behavior changes. Difference between received statistic distribution approximations and $F_{4,25}$ distribution is not so strong. But general statistical behavior regularities are kept.

Fig.3 illustrates case when $\{e_{ij}\}$ and $\{a_i\}$ distributions are the same.

One can see that statistic (5) distribution behaviors in this case and in case represented in Fig.1 are similar.

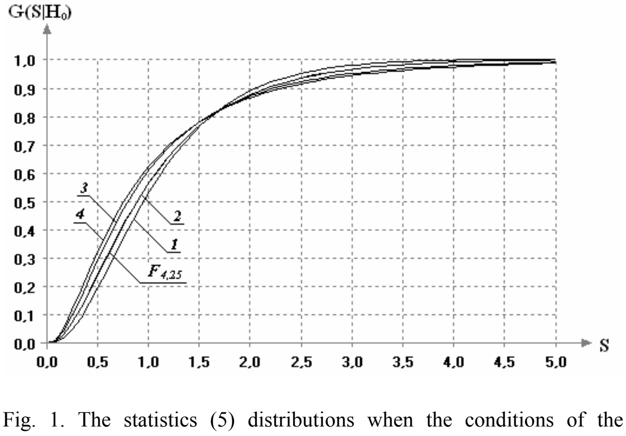

Fig. 1. The statistics (5) distributions when the conditions of the hypothesis (6) are hold and θ_0 = 1: 1 – *G(S(Norm,De(10)))*, 2 – *G(S(Norm,Norm))*, 3 – *G(S(Norm, Max))*, 4 – *G(S(Norm, De(1)))*

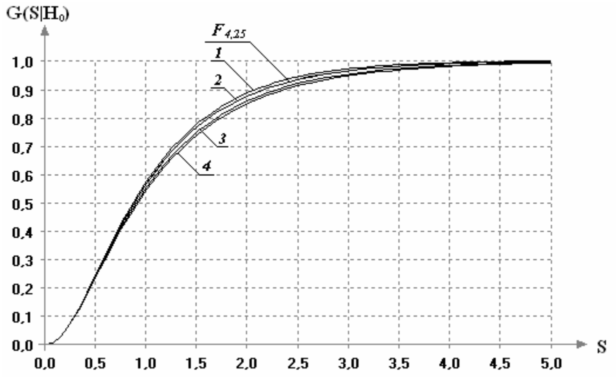

Fig. 2. The statistics (5) distributions when the conditions of the hypothesis (6) are hold and $\theta_0 = 1$: 1 – *G(S(De(10),Norm))*, 2 – *G(S(Norm,Norm))*, 3 – *G(S(Max,Norm))*, 4 – *G(S(De(1),Norm))*

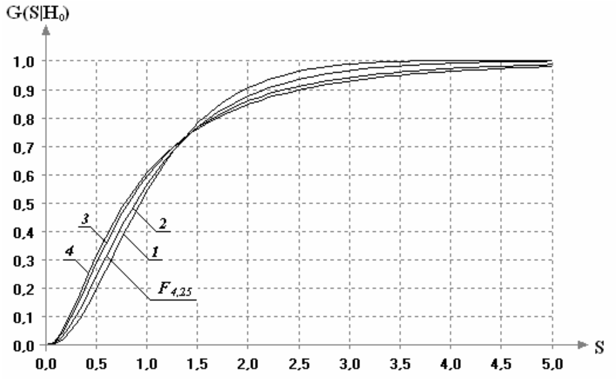

Fig. 3. The statistics (5) distributions when the conditions of the hypothesis (6) are hold and $\theta_0 = 1$: 1 – *G(S(De(10),De(10)))*, 2 – *G(S(Norm,Norm))*, 3 – *G(S(Max, Max))*, 4 – *G(S(De(1),De(1)))*

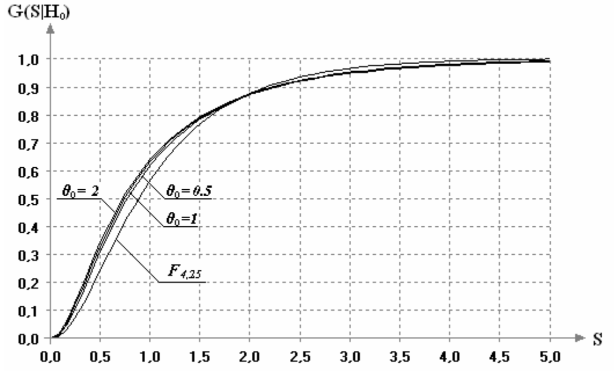

Fig. 4. Distributions *G(S(De(10),De(1)))* of statistics (5) in dependence on θ_0 when the conditions of the hypothesis (6) are hold

The statistic (5) distribution approximations for different combinations of $\{e_{ij}\}$ and $\{a_i\}$ distributions are received.

During the statistic (5) distribution behavior investigation in the normality assumption failure case it was revealed that influence of random factor effect distribution deviation from normal law is greater than influence of error distribution deviation from normal law.

The figure 4 illustrates the change in statistic (5) distributions depending on the θ_0 value parameter change. During the investigation it was revealed that θ_0 value parameter change not essentially influences on statistic distribution.

Within the framework of the model (1) the strength of test with statistic (5) and hypothesis like

$$H_A: \ \sigma_A^2 = \sigma_e^2 \tag{7}$$

under test have been investigated.

The alternative to hypothesis under test is

$$H_1: \ \sigma_A^2 = C\sigma_e^2, \ C > 1. \tag{8}$$

Table 2 represents the power of investigating test for the different alternative (different values of C) and the type I error α when normality assumption are hold.

TABLE 2

POWER OF TEST WITH STATISTIC (5) FOR DIFFERENT α VALUE AND ALTERNATIVE (8) IN THE NORMAL CASE

α	C			
	1.05	1.2	1.5	2
0.1	0.111403	0.148523	0.223937	0.347356
0.05	0.056803	0.080752	0.136446	0.23737

Table 3 and table 4 represent power of investigating test for the different $\{e_{ij}\}$ and $\{a_i\}$ distributions when conditions of alternative hypothesis (8) at $C = 1.2$ are hold. Table 3 and table 4 contain power of test for the type I error $\alpha = 0.05$ and $\alpha = 0.1$ respectively. It is shown that increasing of λ parameter of $\{e_{ij}\}$ or $\{a_i\}$ distribution implies the increasing of power of investigating test.

TABLE 3

POWER OF TEST WITH STATISTIC (5) FOR DIFFERENT $\{e_{ij}\}$ AND $\{a_i\}$ DISTRIBUTIONS AT $\alpha = 0.05$ AND $C = 1.2$

		Error distributions			
		De(1)	Maximum value	Normal	De(10)
Random factor level effect distributions	De(1)	0.072044	0.071396	0.073892	0.075679
	Maximum value	0.071229	0.076016	0.074033	0.072781
	Normal	0.078336	0.086783	0.081353	0.083217
	De(10)	0.084805	0.090454	0.087468	0.090868

TABLE 4

STRENGT OF TEST WITH STATISTIC (5) FOR DIFFERENT $\{e_{ij}\}$ AND $\{a_i\}$ DISTRIBUTIONS AT $\alpha = 0.1$ AND $C = 1.2$

		Error distributions			
		De(1)	Maximum value	Normal	De(10)
Random factor level effect distributions	De(1)	0.134456	0.134081	0.136636	0.139667
	Maximum value	0.135248	0.141331	0.138363	0.1376
	Normal	0.144109	0.155027	0.149217	0.151361
	De(10)	0.152905	0.161304	0.157872	0.161344

REFERENCES

[1] Sheffe, H. 1959. The analysis of Variance. New York: Wiley

[2] Lemeshko B.Yu., Ponomarenko V.M. The problems of classical variance analyses methods applications in the technical, economics and scientifical fields // Proceedings of the regional conference (with participating foreign scientists) "The probability ideas in the science and philosophy ". – Novosibirsk: Inst-t of philosophy and law SB RAS / NSU 2003. – P. 106-109. (In Russian)

[3] Lemeshko B.Yu., Ponomarenko V.M. Statistical Hypotheses Testing In Variance Analysis In Case Of Classical Assumptions Failure // Proceedings of the Seventh International Conference "Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods", September 6-10, 2004, Minsk. Vol. 1. – P. 110-113.