Proceedings of 8th Korea-Russia International Symposium on Science and Technology (KORUS 2004). Tomsk. Russia. June 26-July 3, 2004. – Vol.2. – P.149-152.

COMPUTER SIMULATION TECHNIQUE ON THE INVESTIGATION OF STATISTICAL REGULARITIES [1]

B.Yu. Lemeshko, S.N. Postovalov, E.V. Chimitova, S.S. Pomadin, V.M. Ponomarenko, A.V. Frantsuzov, E.P. Mirkin, S.B. Lemeshko

Novosibirsk State Technical University

Novosibirsk, Russia

e-mail: headrd@fpm.ami.nstu.ru

Abstract

The approach to investigate the fundamental regularities of the mathematical statistics using statistical simulation and computer analysis methods has been considered. The approach is proved to be the most effective for analysing the properties of statistics under non-standard conditions when classical results of mathematical statistics can't be used.

The most considerable results have been obtained by the approach in the research fields listed. These are: goodness-of-fit criteria in composite hypothesis testing, the correlation analysis of multidimensional random variables, the analysis of censored samples, robust estimation methods, robustness of different tests.

Keywords: Goodness-of-fit tests, grouped, censored samples, correlation analysis, variance analysis, normality tests.

1. Introduction

The classical technique of mathematical statistics is based, as a rule, on a number of assumptions, under which the corresponding methods are valid. The assumptions often can't be held under real conditions of applications. For example, errors of one-dimensional or multidimensional measurements don't submit to the normal distribution law; random variable observations can be represented as the following samples: censored, grouped, partially grouped or interval ones (it depends on the measuring conditions); the properties of estimates and statistics in case of limited sample size can essentially differ from the asymptotical ones.

The basic procedures, methods and algorithms in such areas as correlation analysis of multidimensional variables, regression analysis, variance analysis are based on the normality of multidimensional random variables under observation or on the normality of error terms. Under these assumptions the limiting distributions of most statistics used for testing hypotheses on model adequacy or on model parameters are obtained.

2. Computer simulation technique

As a rule, to determine the limiting distribution for a certain statistic by analytical methods turns out to be a highly complicated problem. That is why statistical simulation methods for analyzing regularities are becoming widely used. Within last several years we have been successfully developing the technique of computer simulation and the analysis of statistical regularities. The method has already brought a number of useful results, improving the apparatus of mathematical statistics. These results provide statistical conclusions to be correct when the classical procedures and methods can't be used.

This technique is not opposed to the analytical methods but supplements to them. It enables finding an approximate solution in the cases when it can't be found by the analytical methods. Using simulation results it is possible to come not only to asymptotical conclusions but also to observe changes in regularities with the sample size increasing as well as to investigate their changeability with different factors.

The computer simulation technique consists of two basic stages. The first stage is statistical simulation of a corresponding statistic (by means of appropriate calculation methods of mathematical statistics and software), that results in an empirical distribution of the statistic simulated.

On the basis of the empirical statistic distributions obtained at the second stage it is possible to create and improve the mathematical distribution models of the statistics investigated (on the basis of appropriate software) as well as to make tables of percentage points if necessary.

The technique is steadily being developed in a spiral-way: fundamental regularities obtained (models, describing them) improving the mathematical statistics methods are being built in software systems, that, in its turn, improves their ability to investigate probabilistic regularities.

The main idea of our research is to develop and steadily improve the numerical approach that provides to confirm theoretical results by calculation experiments.

[1] The research is supported by the Ministry of Education of the Russian Federation (project No. T02-3.3-3356)

3. Robust estimation methods

It has been shown that the maximum likelihood estimates (MLE) by grouped samples are highly robust to deviations from assumptions and to existence of anomalous observations. Grouping before estimating procedure allows obtaining robust estimates.

By asymptotically optimal grouping (AOG) that minimizes the Fisher information losses [1, 2] the quality of estimates can be increased. Optimal *L*-estimates of shift and scale parameters obtained on the basis of AOG-quantile estimates were proposed in [3] and investigated in [4]. The MLE for grouped data and optimal *L*-estimates of shift and scale parameters provide parametric procedures of screening anomalous observations to be effective.

4. Investigations of goodness-of-fit tests

A lot of papers are devoted to investigation of goodness-of-fit tests. Nevertheless the practice of their application has proved them to be ineffective or incorrect in many cases. This concerns both the χ^2-type criteria and nonparametric goodness-of-fit tests.

The limiting distributions of nonparametric test statistics of the Kolmogorov type, ω^2 Mises and Anderson-Darling type for composite hypothesis testing depend on the true distribution law being tested, on the number and type of estimated parameters, on the estimation method used and sometimes on the parameter values. But most of even advanced users of statistical methods don't take it into account.

The effectiveness of χ^2 tests depends on the method of splitting a random variable domain into intervals and on the number of intervals. For example, using of AOG maximizes the power of χ^2 Pearson and likelihood ratio tests with respect to close alternative hypotheses. The power of these tests depends on the interval number. The optimal interval number depends on the sample size and certain pair of competing hypotheses. The optimal interval number often turns out to be much smaller than the values, recommended by different papers or given by empirical formulas. The maximal power of the tests for some given sample size is often achieved at the minimally possible or rather small interval number. Non-optimal choice of the interval number and grouping method results in increasing of the beta error probability.

The results of our investigations of nonparametric test statistics [5-9] as well as χ^2 statistic distributions [10-16] for testing simple and composite goodness-of-fit hypotheses were included into recommendations developed for standardization of the State Standard of the Russian Federation [17, 18]. The distributions most often used in practice were considered as the hypotheses under test. The recommendations include the models of statistic distributions created for different composite hypotheses; tables of percentage points; asymptotically optimal grouping tables providing the maximal power of χ^2 tests for close alternatives. The recommendations are based on the authors' results and devoted to eliminating the cases of incorrect using of goodness-of-fit tests during statistical manipulations in different applications.

5. Censored data processing

The problems of distribution parameter estimation for observed values and testing of goodness-of-fit hypotheses for strongly censored samples are very complicated. The Fisher information losses caused by sample censoring have been investigated. In certain cases the censored sample has turned out to contain quite a lot of information even for great censoring degree that enables to obtain rather good parameter estimates. By computer simulation technique MLE distributions have been investigated for different censoring degrees and sample sizes. When sample sizes being limited, the asymptotically effective maximum likelihood estimates have turned out to be biased and their distributions have become asymmetrical [19]. Further investigations have shown that it is possible to make the bias corrections for MLE (as empirical functions of full sample size and censoring degree) on the basis of statistical regularities obtained. The investigations of the Renyi goodness-of-fit test statistic distributions for simple hypotheses testing have showed their poor convergence to the limiting law. A modified statistic lack of this disadvantage has been proposed. The Kolmogorov tests for censored data have been shown to be more preferable.

6. Correlation analysis in non-normality case

Correlation analysis statistic distributions for multidimensional random variables in case of non-normal distribution laws (more peaked or more flat-topped symmetrical distributions) have been investigated. It has been shown that the limiting statistic distributions change insignificantly if the observable law is not normal. The empirical statistic distributions has proved to be in good agreement with the corresponding limiting distributions obtained in the classical correlation analysis under the assumption of observed data normality [20]. So, methods of classical correlation analysis can be valid in a wider area of applications. But this conclusion doesn't take into account the criteria for testing hypotheses on covariance matrixes.

7. Variance analysis in case of non-normal error distributions

The distributions of likelihood ratio statistic used in variance analysis for testing hypotheses on model

parameters have been investigated. The investigations have shown that the level of influence of error non-normality on the considered statistic distribution strongly depends on the method of model parameter estimation. In case of using least squires method, the influence on the limiting statistic distribution is not large. Considerable deviations have been observed only for heavy-tailed error distributions.

When using maximum likelihood method the limiting statistic distribution essentially depends on the error distribution. The approximations of the limiting statistic distributions for certain observation error laws have been constructed.

8. Testing adequacy of nonparametric models

In nonparametric statistics the problems of testing nonparametric models adequacy to the true law are fully omitted.

The investigations have shown that it is possible to use goodness-of-fit criteria for testing adequacy of nonparametric models of distribution laws. When nonparametric estimates are used the distributions of goodness-of-fit test statistics have been shown to be influenced by a number of factors defining a composite hypothesis. These are: the true distribution law of observed variables; the type of a kernel function used; the sample size; the estimation method for fuzziness parameters [21].

9. Investigation of the normality tests

Normality testing is an obligatory procedure during measuring, monitoring and testing. The international standard ISO 5479-97 "Statistical methods. Test for departure of the probability distribution from the normal distribution" doesn't answer such questions as: which of the criteria is preferable, which of them is the most powerful and against what alternatives, for what sample sizes some certain test has an advantage or disadvantage?

The distributions of the following test statistics, included into the standard ISO 5479-97 have been investigated at the paper by means of statistic simulation technique. These are the symmetric property test, the test on kurtosis, Shapiro-Wilk, Epps-Pulley and the modified Shapiro-Wilk tests. A number of criteria which were not included to the standard have been investigated as well.

It has been shown that the Shapiro-Wilk, Epps-Pulley and some other tests have a low power with respect to the flat-topped distributions in comparison with the normal distribution (they cannot distinguish such distributions).

Distributions of the modified Shapiro-Wilk test statistic, included to the standard ISO 5479-97, have been shown to converge to the limiting law very poorly. So using of this test results in increasing of alpha error probability.

The test proposed by D'Agostino, which wasn't included to ISO 5479-97, has been shown to be the most preferable normality test (by the power). This criterion doesn't have the mentioned disadvantage.

9. Testing hypotheses on mathematical expectations and variances

The numerical investigations of classical statistic distributions used for testing hypotheses on mathematical expectations have confirmed their high stability to the deviation of the observed distribution from the normal law. The empirical distributions of corresponding statistics are in a good agreement with the limiting laws obtained on the basis of the assumption of the observed distribution normality. This enables using classical results correctly in practice, when the laws under observation essentially differ from the normal distribution. Obtained results underscore the common regularity: the tests concerning hypotheses of mathematical expectations are robust to deviations of the observations from the normal distribution. This was shown in [20] during investigation of statistics used for testing hypotheses on mathematical expectation vector and correlation coefficient for multidimensional distribution laws.

On the other hand, the distributions of statistics, which concern testing of the hypotheses of variances value, essentially depend on the observed law. If observed distribution is significantly different from the normal law, then the classical results are not valid as using of them will inevitably result in incorrect conclusions.

The distributions of statistics, used in criteria of testing hypotheses of variances value have been investigated with different distributions under the observation. The tables of percentage points have been constructed for given test statistics. The tables are valid for the observed distribution laws, described by the exponential distribution family.

10. The investigation of Bartlett, Cochran and F- tests robustness

The Bartlett and Cochran tests are intended for testing of the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$. The limiting distribution law of the Bartlett statistic for normal random variables is the χ^2_{m-1}-distribution. Bartlett statistic distributions have been numerically shown to be strongly dependent on the distribution under observation. The Cochran statistic distributions have been investigated in a similar way. As a result of investigations the percentage points for the Bartlett and Cochran statistics have been made in case when observed varibles submit the exponential distribution family with different values of the form parameter. The investigations of *F*-test statistic distributions used for testing of the hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ have shown that the observed data distribution practically doesn't influence the statistic distributions.

More detailed information on these investigations is available at the web-site: http://www.ami.nstu.ru/~headrd/seminar/start.htm.

References

[1] Lemeshko B.Yu. (1997) The Robust Methods of Evaluation and Rejection of the Abnormal Measurements. *Zavodskaya Laboratoriya*. V.63. № 5. - pp. 43-49. (in Russian)

[2] Lemeshko B.Yu. (1997) Grouping of observations as the way to obtain robust estimates. *Reliability and quality control*. № 5. - pp. 26-35. (in Russian)

[3] Lemeshko B.Yu., Chimitova E.V. (2001) The development of optimal *L*-estimates for shift and scale distribution parameters by sample quantiles. *Sibirskiy journal industrialnoi mathematiki*. V.4. - № 2. - pp. 166-183. (in Russian)

[4] Lemeshko B.Yu., Chimitova E.V. (2004) The optimal *L*-estimates for shift and scale distribution parameters by sample quantiles. *Zavodskaya Laboratoriya. Diagnostika materialov*. V.70. № 1. (in Russian)

[5] Lemeshko B.Yu., Postovalov S.N. (1998) About nonparametric goodness-of-fit test statistic distributions when observed distribution parameters are estimated from samples. *Zavodskaya Laboratoriya*. V. 64. - № 3. - pp. 61-72. (in Russian)

[6] Lemeshko B.Yu., Postovalov S.N. (1999) On the rules of testing experimental distribution's goodness-of-fit to the theoretical law. *Methods of management of quality. Reliability and quality control*. № 11. - pp. 34-43. (in Russian)

[7] Lemeshko B.Yu., Postovalov S.N. (2001) Application of the nonparametric goodness-of-fit criterions in testing composite hypotheses. *Optoelectronics, Instrumentation and Data Processing*. 2001. - № 2. - pp. 76-88.

[8] Lemeshko B.Yu., Postovalov S.N. (2001) On the dependence of nonparametric test statistic distributions and their power on parameter estimation method used. *Zavodskaya Laboratoriya. Diagnostika materialov*. T. 67. - № 7. - C. 62-71. (in Russian)

[9] Lemeshko B.Yu., Postovalov S.N. (2002) Nonparametric criterions in testing composite hypotheses of goodness-of-fit to the Johnson distribution law. *Doklady sibirskogo otdeleniya akademii nauk vysshey shkoly*. № 1(5). - pp.65-74. (in Russian)

[10] Lemeshko B.Yu. (1997) Asymptotically optimal grouping of observations provides the maximal test power. *Reliability and quality control*. № 8. - pp. 3-14. (in Russian)

[11] Lemeshko B.Yu. (1998) Asymptotically optimal grouping of observations in goodness-of-fit tests. *Zavodskaya Laboratoriya*. V. 64. - №1. - pp.56-64. (in Russian)

[12] Lemeshko B.Yu., Postovalov S.N. (1998) About dependence of the Pearson χ^2 and likelihood ratio statistics limiting distributions on data grouping method. *Zavodskaya Laboratoriya*. V. 64. - № 5. - pp.56-63. (in Russian)

[13] Lemeshko B.Yu., Chimitova E.V. (2000) The maximization of χ^2 tests' power. *Doklady sibirskogo otdeleniya akademii nauk vysshey shkoly*. № 2. - pp. 53-61. (in Russian)

[14] Lemeshko B.Yu., Postovalov S.N., Chimitova E.V. (2001) On statistic distributions and the power of Nikulin χ^2 test. *Zavodskaya Laboratoriya. Diagnostika materialov*. V. 67. - № 3. - pp. 52-58. (in Russian)

[15] Lemeshko B.Yu., Chimitova E.V. (2002) About mistakes and incorrect operations made in using χ^2 goodness-of-fit tests. *Izmeritelnaya Tech-nika*. № 6. - pp. 5-11. (in Russian)

[16] Lemeshko B.Yu., Chimitova E.V. (2003) On the number of intervals' choice in χ^2 goodness-of-fit tests. *Zavodskaya Laboratoriya. Diagnostika materialov*. V.69. - № 1. - pp.61-67. (in Russian)

[17] R 50.1.033-2001. *Recommendations for standardization. Applied statistics. Rules of check of experimental and theoretical distribution of the consent. Part I. Goodness-of-fit tests of a type chi-square*. - Moscow: Publishing house of the standards. 2002. - 87 p. (in Russian)

[18] R 50.1.037-2002. *Recommendations for standardization. Applied statistics. Rules of check of experimental and theoretical distribution of the consent. Part II. Nonparametric goodness–of–fit test*. - Moscow: Publishing house of the standards. 2002. - 64 p. (in Russian)

[19] Lemeshko B.Yu Gildebrant S.Ya., Postovalov S.N. (2001) On reliability parameters estimation from censored samples. *Zavodskaya Laboratoriya. Diagnostika materialov*. V. 67. № 1. - pp. 52-64. (in Russian)

[20] Lemeshko B.Yu., Pomadin S.S. (2002) The correlation analysis of multidimensional random variables observations in the failure of normality assumption. *Sibirskiy journal industrialnoi mathematiki*. V.5. № 3. - pp.115-130. (in Russian)

[21] Lemeshko B.Yu., Postovalov S.N., Francuzov A.V. (2002) On the usage of nonparametric goodness-of-fit criterions for testing adequacy of nonparametric models. *Avtom*

etriya. № 2. - pp.3-14. (in Russian)