Computer methods for investigating statistical regularities in problems of statistical data analysis and reliability

Boris Yu. Lemeshko, Stanislav B. Lemeshko, Ekaterina V. Chimitova, Sergey N. Postovalov
Novosibirsk State Technical University
20 Karl Marx Prospekt, Novosibirsk 630092
Russia
Lemeshko@fpm.ami.nstu.ru

The practice of using statistical analysis methods in applications is full of various problems whose statements are not described within the framework of classical assumptions. A wide range of statistical methods are based on the assumption of measurement error normality. Under real conditions normality and often some other assumptions are not satisfied. The use of classical methods of mathematical statistics in such situations can turn out to be incorrect.

Many classical results are of an asymptotical nature. At the same time in practice one usually works with samples of a limited size. The application of asymptotical results is not always valid for limited size samples.

The form of data (measurements) registration doesn't often conform to complete samples considered in mathematical statistics textbooks. Actually samples of observations can be grouped, censored, partially grouped or interval samples. Mathematical techniques must give an ability to analyze data in any form and must take into account this form and not to neglect it.

As a rule revealing fundamental statistic regularities in nonstandard conditions is a complicated problem for researchers.

Analytical methods for investigating properties of statistical estimates and test statistic distributions are very difficult and as a result of their complexity don't allow researchers to solve a great number of problems. The best way out is to use the numerical approach that is computer modeling of statistical regularities under conditions simulating some real situations of measurement taking. Then mathematical models approximating the regularities obtained are constructed. Such an approach allows us to obtain good results in dealing with problems which are difficult to solve by analytical methods only.

That is why computer simulation methods for statistical regularity analysis are becoming more and more popular.

At present a great number of software systems for statistical data analysis, are widely used in various applications.

The base of CTI (The Computers in Teaching Initiative is an organization, consolidating the British Universities) contains more than 100 software packages of statistical data analysis. Software systems STATISTICA, SPSS, SAS are most often applied in Russia. Some of software systems for statistical analysis are universal systems designed for an extremely wide range of statistical methods, while others are intended for solving a comparatively narrow class of problems. As a rule all software systems provide some techniques for solving statistical analysis problems in various applications. But they don't enable researches to investigate regularities in mathematical statistics for developing mathematical instruments.

While analyzing publications in such journals as "Journal of Statistical Software", "Journal of Computational and Graphical Statistics", "Communication in Statistics", "Computational Statistics & Data Analysis" and others, we can see more and more papers in which numerical methods and in particular statistical simulation methods are used for investigating properties of estimates and statistics and for validating analytical results. In other words computer technologies are more and more frequently used for developing tools of applied mathematical statistics.

A number of useful practical results have been obtained using a computer approach, in particular:

1. The results of investigating statistic distributions and the power of nonparametric goodness-of-fit criteria when testing simple and composite hypotheses about distributions which are frequently used in practice [(1)-(3)] as well as constructed models of statistic distributions for various composite hypotheses and the tables of percentage points were included into recommendations for standardization R 50.1.037-2002 [(4)]. The recommendations are destined for eliminating cases of incorrect

application of goodness-of-fit tests when analyzing observation results in various applications. At the present time these results have been specified and extended [(5)-(8)].

2. The tables of asymptotically optimal grouping have been constructed for rather wide range of distributions most frequently used in practice. An application of asymptotically optimal grouping tables provides the maximal power of χ^2 tests for close competing hypotheses [(9)]. We have investigated the dependence of the test power on the number of grouping intervals. It has been shown for the first time that there is an optimal number of intervals depending on sample size, concrete alternatives and a way of grouping [(10)-(13)]. Some of these results were involved into the recommendations for standardization R 50.1.033–2001 [(14)].

3. The information containing in various sources about advantages of this or that goodness-of-fit test in certain situations is often ambiguous and inconsistent. Estimations of an asymptotical test power are difficult to be used because of the limited sample sizes with which it is necessary to deal in practice. The investigation of test power is a complicated problem because statistic distributions when a competing hypothesis is true are usually unknown.

   The results of investigation of the test power for close competing hypotheses, presented at works [(15)-(17)], enable to order criteria by power.

4. It has been shown that in some cases even for the considerable censoring degree the losses of the Fisher information induced by censoring samples are not large [(18)-(19)]. This enables to obtain rather good estimates of distribution parameters. The distributions of maximum likelihood estimates (MLE) of distribution parameters from censored samples have been investigated by computer simulation methods for various censoring degrees and various sample sizes. It has been shown that for the limited sample sizes the distributions of MLE turn out to be asymmetric and MLE are biased.

5. The distributions of classical statistics used for testing hypotheses about mathematical expectations and variances have been investigated by statistical simulation technique. It has been shown that when testing hypotheses about mathematical expectations the application of classical results turns out to be correct even in cases of considerable deviation from the normal law [(20)]. This result is true for the parametric Student t-tests used for testing hypotheses about equality of means by two samples [(21)]. We have also investigated stability and the power of the Abbe test used for testing hypotheses about the trend absence [(22)].

6. The tables of percentage points have been constructed for statistics used in criteria for testing hypotheses about variances when an observed distribution is described with the exponential family of distributions [(20)]. We have obtained the tables of percentage points for the Bartlett and Cochran tests which can be correctly used when an observed distribution law is described with the family of exponential laws [(23)].

7. The tables of percentage points for the Grubbs type criteria have been obtained for testing simultaneously 3 maximal (or 3 minimal) sample values and simultaneously minimal and maximal values in a sample to be outliers. The distributions of the Grubbs test statistics used for rejecting outliers have been investigated by statistical simulation methods when an observed law is not normal [(24)].

8. The power of Smirnov and Lehmann–Rosenblatt homogeneity tests for two samples has been investigated. A correction for the Smirnov statistic which improves the convergence of statistic distributions to the limiting law has been suggested [(25)].

9. The tools for modeling and investigating distributions of an arbitrary functions of random variables and functions of random variable systems as well as the tools for constructing approximation models for these distributions have been developed [(26)].

10. We have investigated statistic distributions and the power of a number of criteria for testing deviation from the normal law. These tests were compared by power with goodness-of-fit tests. Disadvantages of some popular tests have been shown [(27)].

11. Simulation methods for statistic distributions of multidimensional random variables have been developed. The distributions of statistics for multidimensional random variables are being investigated [(28)].

The computer technologies of data analysis and investigation of probabilistic and statistical regularities is a powerful technique for developing and improving an applied mathematical statistics apparatus.

The investigations carried out are based on the developing software system. On the basis of obtained investigation results and the software system we have developed the course "Computer technologies of data analysis and investigation of statistical regularities" [(29), (30)] for students of the faculty of applied mathematics and computer science of Novosibirsk state technical university.

This research was partly supported by Russian Foundation for Basic Research (project N 06-01-00059-a) and by Federal Education Agency of Russian Federation Ministry of Education in the Analytical departmental purposeful program framework "Development of Higher School Potential", (project N 2.1.2/3970).

References

[1] Lemeshko B.Yu., Postovalov S.N. Statistical distributions of nonparametric goodness-of-fit tests as estimated by the sample parameters of experimentally observed laws // Industrial laboratory (Ind. lab.). 1998. V.64, N 3. - P. 197-208.

[2] Lemeshko B.Yu., Postovalov S.N. // Methods of quality management. Reliability and quality control. - 1999. N 11.

[3] Lemeshko B.Yu., Postovalov S.N. Application of the nonparametric goodness-of-fit Tests in testing composite hypotheses // Optoelectronics, Instrumentation and Data Processing. 2001. - N 2. - P. 76-88.

[4] 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. (in Russian)

[5] Lemeshko B.Yu., Maklakov A.A. Nonparametric Test in Testing Composite Hypotheses on Goodness of Fit Exponential Family Distributions // Optoelectronics, Instrumentation and Data Processing, 2004. V.40, N 3. - P.3-18.

[6] Design of experiments and statistical analysis for grouped observations: Monograph / V.I. Denisov, K.-H. Eger, B.Yu. Lemeshko, E.B. Tsoy. – Novosibirsk: NSTU Publishing house, 2004. – 464 p.

[7] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N. Statistic distribution models for some nonparametric goodness-of-fit tests in testing composite hypotheses // Communications in Statistics. 2009 (in print).

[8] Lemeshko B.Yu., Lemeshko S.B. Statistic distribution models for nonparametric goodness-of-fit criteria when testing composite hypotheses using maximum likelihood estimates. Part I and II // Measurement Techniques. 2009 (in publishing).

[9] Denisov V.I. Lemeshko B.Yu., Tsoy E.B. Optimal grouping, parameter estimation and design of regression experiments. In 2 parts / Novosibirsk: Publishing house of NSTU. 1993. - 347 p. (in Russian)

[10] Lemeshko B.Yu., Postovalov S.N. Limit distributions of the Pearson χ^2 and likelihood ratio statistics and their dependence on the mode of data grouping // Industrial laboratory, 1998. V.64, N 5. - P.56-63. (in Russian)

[11] Lemeshko B.Yu., Chimitova E.V. Maximization of the power of χ^2 tests // Papers of Siberian branch of the Academy of Sciences of higher school, 2000. - N 2. - P. 53-61. (in Russian)

[12] Lemeshko B.Yu., Postovalov S.N., Chimitova E.V. On statistic distributions and the power of the Nikulin c2 test // Industrial laboratory. 2001. V. 67. - N 3. - P. 52-58. (in Russian)

[13] Lemeshko B.Yu., Chimitova E.V. Errors and Incorrect Procedures When Utilizing χ^2 Fitting Criteria // Measurement Techniques, 2002. V.45, N 6. – P.572-581.

[14] 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. (in Russian)

[15] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N. The power of goodness of fit tests for close alternatives // Measurement Techniques, 2007. V.50, N 2. – P. 132-141

[16] Lemeshko B.Yu., Lemeshko S.B, Postovalov S.N. // Siberian journal of industrial mathematics. 2008. - V.11. - N 2(34). - P. 96-111.

[17] Lemeshko B.Yu., Lemeshko S.B, Postovalov S.N. // Siberian journal of industrial mathematics. 2008. - V.11. - N 4(36) . - P. 78-93.

[18] Lemeshko B.Yu. On estimation of distribution parameters and hypothesis testing from censored samples // Methods of quality management. 2001. - N 4. - P. 32-38. (in Russian)

[19] Lemeshko B.Yu., Gildebrant S.Ya., Postovalov S.N. Estimation of reliability parameters from censored samples // Industrial laboratory. 2001. V. 67. - N 1. - P. 52-64. (in Russian)

[20] Lemeshko B.Yu., Pomadin S.S. Testing hypotheses about mathematical expectations and variances in problems of metrology and quality control when probabilistic distributions differ from the normal law. // Metrology. 2004. – N 3.- P. 3-15. (in Russian)

[21] Lemeshko B.Yu., Lemeshko S.B. Power and robustness of criteria used to verify the homogeneity of means // Measurement Techniques. 2008. Vol. 51, N 9. - P.950-959.

[22] Lemeshko S.B. The Abbe independence test with deviations from normality // Measurement Techniques, 2006. V. 49, N 10. – P. 962-969.

[23] Lemeshko B., Mirkin E. Bartlett and Cochran tests in measurements with probability laws different from normal // Measurement Techniques, 2004. Vol. 47, N 10. – P. 960-968.

[24] Lemeshko B.Yu., Lemeshko S. B. Extending the Application of Grubbs-Type Tests in Rejecting Anomalous Measurements // Measurement Techniques, 2005. V.48, N 6. – P.536-547.

[25] Lemeshko B.Yu., Lemeshko S.B. Statistical distribution convergence and homogeneity test power for Smirnov and Lehmann–Rosenblatt tests // Measurement Techniques, 2005. V. 48, N 12. – P.1159-1166.

[26] Lemeshko B.Yu., Ogurtsov D.V. Statistical modeling as an effective instrument for investigating the distribution laws of functions of random quantities // Measurement Techniques, 2007. V.50, N 6. – P. 593-600.

[27] Lemeshko B.Yu., Lemeshko S.B. Comparative analysis of the criteria for testing deviations from the normal law // Metrology. 2005. N 2. – P. 3-24. (in Russian)

[28] Lemeshko B.Yu., Pomadin S.S. Correlation analysis of observations of multidimensional random variables with deviations from normality // Siberian journal of industrial mathematics. 2002. - V.5. - N 3. - P.115-130. (in Russian)

[29] Lemeshko B.Yu., Postovalov S.N. Computer technologies of data analysis and investigation of statistical regularities: Textbook. – Novosibirsk: Publishing house of NSTU. 2004. – 119 p. (in Russian)

[30] Lemeshko B.Yu., Postovalov S.N. Computer technologies of data analysis and investigation of statistical: Tutorial for laboratory works. – Novosibirsk: Publishing house of NSTU. 2007. – 71 p. (in Russian)