Distributed computing system for simulation of classical test statistic distributions under nonstandard conditions

E.V. Chimitova, B.Yu. Lemeshko, S.B. Lemeshko, S.N. Postovalov, A.P. Rogozhnikov
chim@mail.ru, lemeshko@fpm.ami.nstu.ru, skyer@mail.ru, postovalov@ngs.ru, rogozhnikov.andrey@gmail.com

Novosibirsk State Technical University, 20 Karl Marx Avenue, Novosibirsk, 630092, Russia

ABSTRACT: Some limitations of classical statistical methods are considered. General approach to usage of computer simulation in statistical research and mutual improvements of statistical methods and software are described. Wide range of statistical tests supported by the developed software system is enumerated. It is shown how distributed computing in statistical simulation widens the field of application of statistical tests.

1 INTRODUCTION

Applications of classical mathematical statistics are considerably based on an assumption that observations (or measurement errors) have normal distribution. Without that assumption satisfied, use of many statistical tests turns to be incorrect. In the same time, these methods are widely used in applications (either correctly or not). In many cases, a form of data registration (grouping, censoring, independent random censoring, interval representation) prevents researcher from correct using of classical apparatus. The apparatus, in turn, has numerous gaps in possibilities of use, e.g. classical results on testing simple hypotheses with the use of nonparametric goodness-of-fit tests do not extend to complex hypotheses.

2 COMPUTATIONAL APPROACH

Revealing fundamental statistical regularities under nonstandard conditions of applications is always a complicated problem. In the same time, 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 a numerical approach that is computer simulation of statistical regularities under conditions simulating some real situations of measurement taking. Then mathematical models approximating the obtained regularities 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 regularities analysis are becoming more and more popular.

Computer technologies based on efficient use of up-to-date multiprocessor and multiple core computers and on constantly improved software provide tools for developing and extending the methods of applied mathematical statistics. The extension of the methods contributes to extending abilities of the software and development of computer technologies of data analysis and investigation of statistical regularities.

When we develop the software and carry out research with the use of it, we pursue an object to provide correctness of statistical conclusions and practical decisions based on them under real conditions of applications in which the required classical assumptions are not satisfied. I.e. the research is directed to extension of methods of mathematical statistics to the field where classical assumptions are not satisfied.

Using the software system a researcher can carry out comparative analysis of properties and power of statistical tests intended for use in problems of the same type, and, after that choose the most preferable tests and those which use can’t be recommended because of certain shortcomings.

The results of the research (e.g. models of distributions of statistics obtained under conditions that do not agree with classical assumptions) will serve as extension for apparatus of applied mathematical statistics, thus providing the capability for making correct conclusions within the framework of wider assumptions.

3 SUPPORTED TESTS

Giving more details, the developed software system for computer data analysis and investigation of probabilistic (statistical) regularities enables:

– to estimate parameters of more than 30 distribution laws by complete, censored and grouped samples using various estimation methods;

– to test distribution models for goodness-of-fit using χ^2 tests (Pearson test, Rao-Robson-Nikulin test (Nikulin 1973a, 1973b, Rao & Robson 1974), Dzhaparidze-Nikulin test), nonparametric goodness-of-fit tests (Kolmogorov test, Cramer-Von Mises-Smirnov test, Anderson-Darling test) in case of simple and composite hypotheses.

The system enables to simulate and investigate statistic distributions (Lemeshko & Lemeshko 2009a, 2009b), (Lemeshko et al. 2010a), and the test power (Lemeshko et al. 2009) relative to various competing hypotheses, for example for the following groups of tests:

– χ^2 goodness-of-fit tests (Pearson test, Rao-Robson-Nikulin test, Dzhaparidze-Nikulin test) when testing simple and composite hypotheses (Lemeshko & Chimitova 2000, Lemeshko et al. 2001, Lemeshko & Chimitova 2002, Lemeshko & Chimitova 2003);

– nonparametric goodness-of-fit tests (Kolmogorov test, Cramer-Von Mises-Smirnov test, Anderson-Darling test) for simple and composite hypotheses;

– criteria for testing deviation of an empirical distribution from the normal law (Shapiro-Wilk test, Epps-Pulley test, D’Agostino test, Frosini’s test, Hegazy-Green’s test, Spiegelhalter’s test, Geary’s test, David-Hartley-Pearson’s test and others, more than 15 (Lemeshko & Lemeshko 2005a, Lemeshko & Rogozhnikov 2009));

– criteria for testing hypotheses about mathematical expectations and variances when samples of random variables belong to various distributions;

– criteria for testing homogeneity of parametric means (Tests to compare two sample means with known variances, student’s test for comparison of two sample means with unknown but equal variances, test to compare two sample means with unknown and unequal variances) and nonparametric means (Mann–Whitney’s test, Crustal–Wallis’s test, Fisher’s test (Lemeshko & Lemeshko 2008));

– criteria for testing homogeneity of parametric variances (Bartlett’s test, Cochran’s test, Hartley’s test, Levene’s test) and nonparametric variances (Ansari-Bradley’s test, Siegel-Tukey’s test, Mood`s test and others (Lemeshko et al. 2010b));

– tests of independence (Abbe’s test (Lemeshko 2006));

– tests for rough measurement errors (Grubbs’s test (Lemeshko & Lemeshko, 2005b));

– and so on.

4 DISTRIBUTED COMPUTING

Generally speaking, distributions of tests statistics are dependent on every variable parameter, what results in infinite count of variables combinations and, consequently, infinite count of distributions. It is impossible to foresee and provide corresponding models for every use case of a specific test. Moreover, building sufficiently accurate models based on statistical simulation requires a lot of calculations (tens and hundreds of hours on a typical PC).

Distributed computing in the data analysis software system enables to simulate unknown test statistic distributions in real time, i.e. while testing statistical hypotheses. Acceleration of calculations almost directly depends on count of computing units. Researcher can obtain required models in a reasonable amount of time by using more units and varying required accuracy.

5 CONCLUSION

The developed software system enables to carry out data analysis and research statistical regularities under conditions when standard assumptions are not satisfied. Distributed computing accelerates corresponding statistical simulations and makes it possible to obtain models of distributions without waiting for large amounts of time.

This research have been performed with partial support from the Russian Foundation for Basic Research (Grant 09-01-00056-a) and the Federal Agency on Education of the Ministry of Education and Science of the Russian Federation within the framework of the analytical departmental targeted program "Development of higher school scientific potential" (Project 2.1.2/3970) and federal targeted program of the Ministry of Education and Science of the Russian Federation "Scientific and scientific-educational personnel of innovation Russia".

6 REFERENCES

Nikulin, M.S. 1973a. On a chi-squared test for continuous distributions. *Theory of Probability and its Applications*, V.18, N 3: 638-639. [In Russian]

Nikulin, M.S. 1973b. Chi-squared test for continuous distributions with shift and scale parameters. *Theory of Probability and its Applications*, V.18, N 3: 559-568. [In Russian]

Rao, K.C., Robson, D.S. 1974. A chi-squared statistic for goodness-of-fit tests within the exponential family. *Commun. Statist.*, V.3: 1139-1153.

Lemeshko, B.Yu. & Lemeshko S.B. 2009a. Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part I, *Measurement Techniques* 52(6): 555–565.

Lemeshko, B.Yu. & Lemeshko, S.B. 2009b. Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II, *Measurement Techniques* 52(8): 799–812.

Lemeshko B.Yu., Lemeshko S.B., and Postovalov S.N. 2009. Comparative Analysis of the Power of Goodness-of-Fit Tests for Near Competing Hypotheses. I. The Verification of Simple Hypotheses, *Journal of Applied and Industrial Mathematics* 4(4): 462–475.

Lemeshko, B.Yu. Lemeshko, S.B. & Postovalov, S.N. 2010a. Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses, *Communications in Statistics - Theory and Methods* 39(3): 460–471.

Lemeshko, B.Yu., Chimitova, E.V. 2000. Maximization of the power of chi-squire tests. *Papers of Siberian branch of the Academy of Sciences of higher school* 2: 53-61. [In Russian]

Lemeshko, B.Yu., Postovalov, S.N., Chimitova, E.V. 2001. On statistic distributions and the power of the Nikulin chi-squire test. *Ind. Lab.*, V.67, N 3: 52-58.

Lemeshko, B.Yu., Chimitova, E.V. 2003. On the choice of the number of intervals in chi-squire goodness-of-fit tests. *Ind. Lab. Mat. Diag.*, V.69, N 1: 61-67.

Lemeshko, B.Yu., Chimitova, E.V. 2002. Errors and incorrect procedures when utilizing χ^2 fitting criteria. *Measurement Techniques*, V.45, N 6: 572-581.

Lemeshko, B.Yu., Lemeshko, S.B. 2005a. Comparative analysis of criteria of testing deviation from the normal law. *Metrology*, 2: 3-24. [In Russian]

Lemeshko, B.Yu., Rogozhnikov, A.P. 2009. The investigation of features and the power of some tests for normality. *Metrology*, 4: 3-24.

Lemeshko B.Yu., Lemeshko S.B. 2008. Power and robustness of criteria used to verify the homogeneity of means. *Measurement Techniques*, V. 51, N 9: 950-959.

Lemeshko B.Yu., Lemeshko S.B., Gorbunova A.A. 2010b. About using and power of variance homogeneity tests. P. I. *Measurement Techniques*. 3: 10-16.

Lemeshko, S.B. 2006. The Abbe independence test with deviations from normality. *Measurement Techniques*, V.49, N 10: 962-969.

Lemeshko, B.Yu., Lemeshko, S.B. 2005b. Extending the application of Grubbs-type tests in rejecting anomalous measurements. *Measurement Techniques*, V.48, N 6: 536-547.