APPLIED **M**ETHODS OF **S**TATISTICAL **A**NALYSIS.

SIMULATIONS AND STATISTICAL INFERENCE

BOOK OF ABSTRACTS of the International Workshop

NOVOSIBIRSK

20-22 September 2011

Applied Methods of Statistical Analysis. Simulations and Statistical Inference

Chairmen:

Narayanaswamy Balakrishnan,	McMaster University, Canada
Mikhail Nikulin,	University of Bordeaux, France
Aleksey Vostretsov,	Novosibirsk State Technical University, Russia
Boris Lemeshko,	Novosibirsk State Technical University, Russia
Evgeny Tsoy,	Novosibirsk State Technical University, Russia

Scientific Committee:

A. Antonov,	Institute of Nuclear Power Engineering, Russia
N. Balakrishnan,	McMaster University, Canada
V. Denisov,	Novosibirsk State Technical University , Russia
K.-H. Eger,	Technische Univesitat Chemnitz, Germany
B. Lemeshko,	Novosibirsk State Technical University , Russia
N. Limnios,	Universite de Technologie de Compiegne, France
V. Melas,	St. Petersburg State University, Russia
M. Nikulin,	University of Bordeaux, France
B. Ryabko,	Siberian State University of Telecommunications and Information Sciences, Russia
V. Rykov,	Gubkin University of Oil and Gas, Russia
V. Solev,	St.Petersburg Department of Steklov Mathematical Institute RAS, Russia
A. Vostretsov,	Novosibirsk State Technical University , Russia
N. Zagoruiko,	Sobolev Institute of Mathematics of the Siberian Branch of the RAS, Russia

Local Committee:

E. Chimitova, M. Vedernikova, N. Galanova, A. Gorbunova, A. Rogozhnikov

Part I
Simulation and Research of Probabilistic Regularities

Simulation and research of probabilistic regularities in motion of traffic flows

M.A. Fedotkin, E.V. Kudryavtcev, M.A. Rachinskaya

A random number of vehicles crossed a transverse line of a highway during an arbitrary interval of time and a random number of vehicles situated on an arbitrary part of the highway at a fixed instant of time form a complicated stochastic dependence. In case of bad weather and bad road conditions the spatial intervals between the neighbouring vehicles at a fixed instant of time (a spatial characteristic of the flow) are dependent and have different probability distributions. Under these conditions, the intervals between moments of crossing the transverse line of the highway by the consecutive vehicles (a temporal characteristic of the flow) are also dependent and have different probability distributions. In this paper some probabilistic regularities in motion of this kind of traffic flows are researched.

Keywords: flow of vehicles, traffic batch, system of Kolmogorov differential equations, limiting probability distribution, parameters estimate.

Real-time studying of statistic distributions of non-parametric goodness-of-fit tests when testing complex hypotheses

Boris Yu. Lemeshko, Stanislav B. Lemeshko, Andrey P. Rogozhnikov

In present work, a “real-time” ability to simulate and research the distributions of tests statistics in the course of testing the complex goodness-of-fit hypothesis (for distributions with estimated parameters) is implemented by the use of parallel computing. It makes it possible to make correct statistical inferences even in those situations when the distribution of the test statistic is unknown (before the testing procedure starts).

Keywords: goodness-of-fit test, composite hypotheses testing, Kolmogorov test, Cramer-Mises-Smirnov test, Anderson-Darling test, methods of statistical simulation.

Application of variance homogeneity tests under violation of normality assumption

Alisa A. Gorbunova, Boris Yu. Lemeshko

Classical tests for homogeneity of variances (Fisher's, Bartlett's, Cochran's, Hartley's, Neyman-Pearson's, Levene's, modified Levene's, Z-variance, Overall-Woodward modified Z-variance, O'Brien tests) and nonparametric tests (Ansari-Bradley's, Mood's, Siegel-Tukey's, Capon's and Klotz's tests) have been considered. Distributions of classical tests statistics have been investigated under violation of assumption that samples are normally distributed. The comparative analysis of power of classical tests with power of nonparametric tests has been carried out. Tables of percentage points for Cochran's test have been made for distributions which are different from normal. Software, that allows us to apply tests correctly, has been developed.

Keywords: homogeneity of variance test, power of test.

On random-number generator of given distribution

V. F. Pervushin, N. A. Sergeeva, A. V. Strelnikov

The algorithm generating the sample of random numbers of defined distribution and numerical characteristics (expectation, dispersion, etc.) is considered. The high accuracy of algorithm working on sample modeling is shown. The general principle of sample generation allows assigning the given approach in a category of random-numbers generation algorithms that combining such algorithms in group of "precision random-numbers generator".

Simulation study for the NRR chi-square test of goodness-of-fit for censored data

Ekaterina V. Chimitova, Angelika O. Tsivinskaya

This paper presents the investigation results for the Nikulin-Rao-Robson (NRR) chi-square type test. The distributions of the NRR test statistic have been investigated by means of computer simulation technique depending on the sample size, censoring distribution, proportion of censoring and number of intervals. Simulation studies of chi-square test statistic distributions have been shown for type I, II and random censoring. Using computer simulation we have studied the power of the NRR test for close competing hypotheses.

Keywords: Nikulin-Rao-Robson chi-square test, goodness-of-fit, censored samples, test power, Monte Carlo simulations.

Part II
Statistical Methods in Reliability and Survival analysis

Mathematical model of the residual lifetime of NPP equipment calculation based on operational information specific type

Alexander Antonov, Sergey Sokolov, Valeriy Chepurko

Probabilistic estimation method of the average straight residual lifetime for nuclear power plants (NPPs) systems and their constituent elements is considered. The mathematical model for calculating of this reliability characteristic for the objects to be recovered from the initial data on failures censored interval is presented. Besides, the issue of its accuracy estimating using the bootstrap method is considered.

Keywords: residual lifetime, system, element, reliability characteristic, operational data.

Statistical analysis of mortality-comorbidity links

Varvara V. Tsurko, Anatoly I. Michalski

In this paper we investigate dependences between associated diseases that a person has at the end of his live and the cause of death. We analyze public data about cause-specific mortality in conjunction with the problem of average risk estimation on empirical data. The use of the theory of Vapnik-Chervonenkis provides informative results about differences between distributions of associated diseases in group of people who died of cancer and group of people who died of another disease. This difference uncovers a relationship between some groups of associated diseases and risk of death of cancer.

Keywords: cancer mortality, distributions discrepancy, selection of associated diseases, the Vapnik-Chervonenkis dimension.

Testing goodness-of-fit with parametric AFT-model

Ekaterina V. Chimitova, Natalia S. Galanova

This paper is devoted to the problems of goodness-of-fit testing with parametric AFT-model. Modified nonparametric goodness-of-fit tests such as Kolmogorov test, Cramer-von Mises-Smirnov test and Anderson-Darling test by samples of residuals are investigated. The problem of baseline distribution selecting is considered.

Keywords: AFT-model, censored data, samples of residuals, Kolmogorov test, Creamer-von Mises-Smirnov test, Anderson-Darling test.

Inference for a simple step-stress model with progressive type II censoring and an extension of the exponential distribution

Firoozeh Haghighi, Mazaher Sohrabi Noor

In this work we consider a simple step-stress model under progressive Type-II censoring based on an extension of the exponential distribution, which provides a more flexible model than the exponential model. This new generalization of the exponential distribution has been recently introduced by Nadarajah and Haghighi (2010), and can be used for modeling lifetime data. For this simple step-stress model the maximum likelihood estimates of its parameters as well as the corresponding observed fisher information matrix are derived. A method for simulating data from an extension of the exponential distribution in the presence of progressive Type-II censoring is proposed. Using this method we conducted a simulation study for estimating the parameters of simple step-stress model and then provided asymptotic and bootstrap confidence intervals for the parameters.

Keywords: An extension of the exponential distribution, Bootstrap, Coverage probabilities, Cumulative exposure model, Fisher information matrix, Step-stress test, Type-II censoring.

Comparing prediction performances via IDI. Application to French Alzheimer data

C. Huber-Carol, S. Toaff-Gross

Motivation

It may happen that among the factors that have an impact on the occurrence of a certain event we want to predict, some of them are difficult to obtain. It can be due to their high cost or else to the time spent to get them. In that case, and if the purpose is purely predictive, and not at all explanatory, it may happen that dropping such factors has a very low cost in terms of predictive ability of the model. The aim of this paper is to derive the asymptotic properties of an estimator of an index of predictive ability, the IDI, when both the full model and the reduced model are estimated on the same data set, together with their IDI. Having thus a confidence interval for their comparative predictive ability, we have elements allowing us to conclude whether we can drop or not certain pertinent factors.

Framework

Let $X = (Y, \mathbf{Z})$ be a random variable such that Y is binary with values in $\{0, 1\}$, $P(Y = 1|\mathbf{Z} = \mathbf{z}) = \mathbf{p}(\mathbf{z})$, and $\mathbf{Z}$ is a k-dimensional real variable, with distribution $Q(\mathbf{z})$ with density $q(\mathbf{z})$ with respect to some measure μ. We have a data set $\mathbf{X} = (X_1, \cdots, X_n)$ consisting in n i.i.d. observations of X, and two models for predicting Y on the basis of $\mathbf{Z}$ are to be compared:

$$\begin{array}{lrcl} \text{Model 1} & P(Y = 1|\mathbf{Z} = \mathbf{z}) & = & p_1(\mathbf{z}) \\ \text{Model 2} & P(Y = 1|\mathbf{Z} = \mathbf{z}) & = & p_2(\mathbf{z}) \end{array}$$

while the unknown true distribution of X is given by

$$\begin{array}{lcl} P(Y=1|\mathbf{Z}=\mathbf{z}) & = & p(\mathbf{z}) \\ dQ(\mathbf{z}) & = & q(\mathbf{z})\mathbf{d}\mu(\mathbf{z}) \end{array}$$

This setting originates from the following special problem in epidemiology:
Y_i is the indicator of the occurrence of a certain disease for subject i. Occurrence of this event is to be predicted to happen within a fixed period of time, the prediction being based on the value $\mathbf{z_i}$ of $\mathbf{Z}$ observed on subject i. $\mathbf{Z}$ is a k-dimensional covariate, p_1 and p_2 are logistic models, denoted g_1 and g_2 in the sequel. While g_1 is including all k components of $\mathbf{Z}$, g_2 is obtained by throwing away Z_k which is a genetic feature. We consider the case when a test of fit of the full model g_1 shows that Z_k is a pertinent covariate, that is its coefficient is significantly different from 0. It may happen that, in spite of the fact that g_1 is a better model than g_2, the improvement in prediction is not significant, due to the fact that the coefficients of the remaining covariates are modified so as to fit better the data at the cost of giving misleading false effects for the remaining covariates. The reason for avoiding the last covariate, even though it is a pertinent one, may be, as is the case for a genetic feature, the fact that it is not available for all the subjects that could be involved in the study, or else it would be too expensive or too long to obtain its values in view of the magnitude of the small benefit it would provide.
We stress that, while $q(z)$ is the true distribution of $\mathbf{Z}$, we do not assume that the full model g_1 is the true model for the data.
The aim of this work is to derive the asymptotic properties of the IDI in order to obtain confidence intervals for IDI (Integrated Discrimination Improvement) and other related measures of comparison of prediction performances. Several indexes for comparing prediction ability of two models can be found in Pencina et al.[1]. Properties of M-esdtimators as can be found in C. Huber[2] are used for derivng the asymptotic properties of IDI.

Application to French Alzheimer data

n = 4486 patients aged ≥ 65, included in a cohort between September 1999 and November 2000, are followed during several years. Covariates such as sex, age at inclusion, sociological, psychological and biological factors as well as three genetic factors are considered that could influence the occurrence of Alzheimer dsease. Among the 4486 patients, 162 became Alzheimer within 4 years. Only one of the three genetic factors is shown to be pertinent and is included in the best fitting logistic model for predicting this occurrence. Nevertheless, the IDI between the two models with and without the genetic factor is not significant. This allows us to think that the search for this costly factor could be avoided without loosing much as long as prediction only is concerned, and not structural explanation.

[1] Pencina M.J., D'Agostino R.B. Sr, D'Agostino R.B. Jr & Vasan R.S. (2007). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medecine*

[2] Huber C. (1985). Thorie de la Robustesse. *Lecture notes in Mathematics*, pp. 1-128, Springer Verlag.

Part III
Testing Statistical Hypotheses

Modified chi-squared goodness-of-fit test for Birnbaum-Saunders distribution

Mikhail Nikulin, N. Balakrishnan, Ramzan Tahir, Noureddine Saaidia

Two-parameter Birnbaum-Saunders distribution is widely used in industry for reliability. In this paper we give a modified chi-squared goodness-of-fit test for Birnbaum-Saunders distribution when the data are right censored. Random grouping intervals of the data function are used.

On validation of models in demography

Léo Gerville-Réache, Mikhail Nikulin, Ramzan Tahir

In demography, Gompertz and Makeham models have significant role in modeling and in analysis of mortality and ageing. Till the end of 20th century, researchers have been used the tables of mortalities for demographic analysis but in the end of 20th century due to the development in statistical methods of survival analysis one can treat the individuals data even with the information of censoring. Weibull model is considered the alternative for Gompertz model (Juckett and Rosenberg, (1993)). The Gompertz, Makeham, and Weibull distributions are compared with respect to the goodness-of-fit to the table of mortality and to the individuals data in presence of censoring. For data from the table of mortality, test statistic considered by Gerville-Reache and Nikulin (2000) is used. For censored individual data the test is based on the NRR-statistic where the choice of random grouping intervals is considered as given by Hjort (1990), Akritas (1988), Bagdonavicius, Kruopis and Nikulin (2010).

Keywords:Demography, Gompertz model, Makeham model, Weibull model, Composite hypothesis, ML estimators, Chi-square test, Censoring, NRR statistic.

Conditional distributions and scaling for categorial data

Henning Läuter

We consider continuous random variables, look at conditional distributions and find that the estimation of unknown conditions can be used for modeling categorical data. We introduce maximum likelihood estimates for the conditional parameters and so we find scaled values for the categories by maximum likelihood. With these methods the scaling for the levels of the factors in models of analysis of variance can be solved. Under normal distributions explicit solutions are given, for other distributions, e.g. for survival distributions, the scaling is described by copulas and different methods of estimates.

Keywords: Conditional distributions, modeling, categorical data, statistical scaling.

Goodness-of-fit testing, smoothing and resampling under censoring

Hannelore Liero

Goodness-of-fit tests using kernel estimators for the density and the hazard rate are proposed. The aim is to investigate these tests for data where censoring is present. A third test statistic, the continuous analogue of the χ^2-test statistic based on the counts of uncensored observations, is considered. Resampling methods for the realization of the test procedures are discussed.

Keywords: censoring; goodness of fit; kernel estimators; resampling

Bayesian Model specification: some problems related to model choice and calibration

Milovan Krnjajić, David Draper

In the development of Bayesian model specification for inference and prediction we focus on the conditional distributions $p(\theta|\mathcal{B})$ and $p(D|\theta, \mathcal{B})$, with data D and background assumptions $\mathcal{B}$, and consider *calibration* (an assessment of how often we get the right answers) as an important integral step of the model development. We compare several predictive model-choice criteria and present related calibration results. In particular, we have implemented a simulation study to compare predictive model-choice criteria LS_{CV}, a *log-score* based on cross-validation, LS_{FS}, a full-sample log score, with deviance information criterion, DIC. We show that for several classes of models DIC and LS_{CV} are (strongly) negatively correlated; that LS_{FS} has better small-sample model discrimination performance than either DIC, or LS_{CV}; we further demonstrate that when validating the model-choice results, a standard use of *posterior predictive tail-area* for hypothesis testing can be poorly calibrated and present a method for its proper calibration.

Keywords: log-score, deviance information criterion, posterior predictive tail areas, hypothesis testing.

Statistical analysis of Markov chains of conditional order

M.V. Maltsew, Yu.S. Kharin

A new mathematical model — Markov chain of conditional order — is proposed for statistical analysis of discrete time series with “long memory”. Statistical estimators for parameters of this model are constructed and their properties are analyzed. A statistical test for the values of parameters is proposed. Results of computer experiments are presented.

Keywords: Markov Chain, Markov Chain of Conditional Order, Maximum Likelihood, Hypothesis Testing.

Comparing predictive accuracy

Eva Ferreira, Winfried Stute

In this work we provide new tests for the difference in predictive accuracy of two prognostic factors X_1 and X_2 on a common output Y. Given a set of independent replicates of (X_1, X_2, Y), we split this sample into a learning part for estimating the unknown regression functions, and a validation part for which the residuals need to be computed. We show that the null distributions of our test statistics may be approximated by a normal. In simulations, the power is promising already for small to moderate sample sizes.

Keywords: Predictive accuracy; residuals; nonparametric test, data split.

Reduction of the average sample number in sequential scheme of testing hypotheses

Sergey Postovalov, Madina Shakhmametova

Problems of testing two simple hypotheses about the distribution of a random variable by the results of independent observations are considered. The main goal is finding critical values of SPRT, 2-SPRT and generalized optimal sequential Ayvazyan's test for testing hypotheses about normal and logistic distributions where probabilities of the first and second type errors have some specified values. It is shown that the use of the obtained critical values minimizes the average sample number as compared to the use of known theoretical approximate critical values.

Keywords: SPRT; 2-SPRT; generalized optimal sequential test; average sample number.

A chi-squared test for the family of inverse Gaussian distributions for censored data

N. Saaidia

In this paper we propose a Chi-squared type test based on the Rao-Robson-Nikulin statistic under random censoring for the inverse Gaussian family.

Keywords: Pearson's Chi-squared test, Modified Chi-squared test, Inverse Gaussian distribution, Rondam censored data, Estimation, Rao-Robson-Nikulin statistic, Goodness-of-fit test, Maximum likelihood.

Application of classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for censored samples

Ekaterina Chimitova, Hannelore Liero and Mariya Vedernikova

Problems of testing statistical goodness-of-fit hypotheses for censored data are considered in the paper. The application of the classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests for a complete sample obtained from original censored sample by using randomization is proposed. By means of computer simulation methods we have investigated the test statistic distributions and the power of considered tests for close competing hypotheses when testing simple and composite hypotheses.

Keywords: censored data, goodness-of-fit testing, Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test, randomization.

Part IV
Nonparametric Methods

Using FRiS-function for nine medical tasks solving

Nikolay Zagoruiko, Irina Borisova, Vladimir Dyubanov, Olga Kutnenko

In tasks of modern biology the quantity of features often on orders exceeds quantity of objects. For the decision of such tasks the method Data Mining based on use the new measure of similarity between objects in the form of Function of Rival Similarity (FRiS) is offered. On this basis the methods of a quantitative estimation of compactness of patterns, of a construction decision rules and of feature selection are developed. High efficiency of methods is illustrated by results of the decision of nine tasks of recognition of diseases on microarray dataset.

Keywords: feature selection, FRiS-function, compactness.

Non-parametric stochastic approximation in adaptive systems theory

A.V. Medvedev

The paper discusses the problem of parametric and non-parametric stochastic approximations according to the experimental information. In this respect the identification problem in a “wide” sense and working sample generation from the initial training one are considered. Some modifications of the known non-parametric estimations of the regression curve according to the observations are introduced; their use in adaptive identification and control problems in conditions of non-parametric uncertainty is analyzed. The results of numerical study are presented.

About regression characteristics nonparametric estimation in the identification problem

D. Bezmen, L. Golub, A. Medvedev

In many applications, the regression function restoration from observational data with random errors is directly related to the identification problem. In this regard, the case where by the "input-output" variables supervision of an object are significant errors such as "crude error" and the reducing the dimension of the input variables vector problem is most attractive. This report focuses on the these issues analysis.

Non-parametric H-models of thermal processes

R.S. Boyko, Ya.I. Demchenko

Some discrete-continuous processes identification and modeling tasks are considered. The new non-parametric estimations of probabilities distribution density function and regression curve according to observations are set, and also the convergence theorems. The new modifications of H-models are applied during the modeling of thermal process of oil decomposition. Key words: modeling, identification, non-parametric estimations of regression function, non-parametric estimations of probabilities distribution density functions, convergence theorems, linear models, H-models.

Test of goodness of fit in dose-effect model based on finite sample

Victor M. Kocheganov, Michail S. Tikhov

The central limit theorem is proved for nonparametric estimator of distribution function $F(x)$. It is shown that the given estimation approaches for convolution of $F(x)$ and kernel $K_h(x)$ is better. The direct way of estimation of the distribution function $F(x)$ based on consistent estimate of characteristic function if offered.

Keywords: distribution function kernel estimator, asymptotic normality, integrated square error, summarized square error, convolution.

Semi-recursive nonparametric algorithms of identification and forecasting

Irina L. Foox, Irina Yu. Glukhova, Gennady M. Koshkin

A class of semi-recursive kernel plug-in algorithms of identification and forecasting is considered. The main parts of the asymptotic mean square errors (AMSE) of the estimates are found. The algorithms of identification and forecasting are applied to investigate the dependence of Russian Federation's Industrial Production Index on the dollar exchange rate, direct investment, and export for the period from September 1994 till March 2004.

Keywords: identification, forecasting, kernel recursive estimator, mean square convergence.

Nonparametrical estimation of survival functions by censored data

Abdurahim Abdushukurov

Incomplete observations occur in survival analysis, especially in clinical trials and engineering when we partially observe death in biological organisms or failure in mechanical systems.

From statistical literature one can learn that incomplete observations arise in two ways: by censoring and truncation. Note that truncation is sampling an incomplete population, while censoring occurs when we are able to sample the complete population, but the individual values of observations below and/or above given values are not specified. Therefore censoring should not be confused with truncation. In this work we deal only with right censoring model which is easily described from methodological point of view.

Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables (i.i.d.r.v.-s) (the lifetimes) with common distribution function(d.f.) F. Let X_j be censored on the right by Y_j, so that observations available for us at the n-th stage consist of the sample $S^{(n)} = \{(Z_j, \delta_j), 1 \le j \le n\}$, where $Z_j = min(X_j, Y_j)$ and $\delta_j = I(X_j \le Y_j)$ with $I(A)$ meaning the indicator of the event A. Suppose that Y_j are again i.i.d.r.v.-s, the so-called censoring times with common d.f. G, independent of lifetimes X_j.

The main problem consist of nonparametrically estimating F with nuisance G based on censored sample $S^{(n)}$, where r.v.-s of interest X_j-s observed only when δ_j=1. Kaplan and Meier (1958) were the first to suggest the product-limit (PL) estimator F_n^{PL} defined as

$$F_n^{PL}(t) = \begin{cases} 1 - \prod_{\{j:Z_{(j)} \le t\}} \left[1 - \frac{\delta_{(j)}}{n-j+1}\right], & t \le Z_{(n)}, \\ 1, & t > Z_{(n)}, \delta_{(n)} = 1, \\ undefined, & t > Z_{(n)}, \delta_{(n)} = 0, \end{cases}$$

where $Z_{(1)} \le ... \le Z_{(n)}$ are the order statistics of Z_j and $\delta_{(1)}, ..., \delta_{(n)}$ are the corresponding δ_j. In statistical literature there are different versions of this estimator. However, those do not coincide if the largest Z_j is a censoring time. Gill (1980) redefined the F_n^{PL} setting $F_n^{PL}(t) = F_n^{PL}(Z_{(n)})$ when $t > Z_{(n)}$. Further, we use the Gill's modification of PL-estimator. At present there is an enormous literature on properties of the PL-estimator (see, for example [3]-[9]) and most of work on estimating incomplete observation are concentrated on PL-estimator. However F_n^{PL} is not unique estimator of F.

The second, closely related with the F_n^{PL}, nonparametrical estimator of F is the exponential hazard estimator

$$F_n^E(t) = 1 - exp\Big\{ -\sum_{j=1}^{n} \frac{\delta_{(j)} I(Z_{(j)} \le t)}{n-j+1} \Big\}, -\infty < t < \infty.$$

F_n^E plays an important role in investigating the limiting properties of the estimator F_n^{PL}.

Abdushukurov (1998,1999) proposed another estimator for F of power type:

$$F_n(t) = 1 - (1 - H_n(t))^{R_n(t)} = \begin{cases} 0, & t < Z_{(1)}, \\ 1 - (\frac{n-j}{n})^{R_n(t)}, & Z_{(j)} \le t < Z_{(j+1)}, 1 \le j \le n-1, \\ 1, & t \ge Z_{(n)}, \end{cases}$$

where

$$H_n(t) = \frac{1}{n} \sum_{j=1}^{n} I(Z_j \le t)$$

is empirical estimator of d.f. $H(t) = P(Z_j \le t) = 1 - (1 - F(t))(1 - G(t))$ and

$$R_n(t) = \frac{-log(1 - F_n^E(t))}{\sum_{j=1}^{n} \frac{I(Z_{(j)} \le t)}{n-j+1}}.$$

As we see, estimator F_n is defined on whole line. Let

$$a_n(t) = \sum_{j=1}^{n} \frac{I(Z_{(j)} \le t)}{(n-j)(n-j+1)}.$$

Note that $\sup\{a_n(t), t \le T\} \le [n(1 - H_n(T))]^{-1} = \mathbb{O}(\frac{1}{n})$ with probability 1, where $T < Z_{(n)}$.

Following inequalities show that all three estimators are closely related (Abdushukurov [1-5]): *For $t < Z_{(n)}$ with probability 1*

(I) $\quad 0 < -log(1 - F_n^{PL}(t)) + log(1 - F_n^E(t)) < a_n(t);$

(II) $\quad 0 \le F_n^{PL}(t) - F_n^E(t) < \frac{1}{2} a_n(t);$

(III) $\quad 0 < -log(1 - F_n(t)) + log(1 - F_n^E(t)) < a_n(t);$

(IV) $\quad |-log(1 - F_n^{PL}(t)) + log(1 - F_n(t))| < a_n(t);$

(V) $\quad |F_n^{PL}(t) - F_n(t)| < a_n(t);$

(VI) $\quad |F_n^E(t) - F_n(t)| < a_n(t).$

Thus one can expect the stochastic equivalences of these estimators in the sense of their weak convergence to the same Gaussian process (Abdushukurov (1998)). Let d.f.-s F and G

be continuous and $T < T_H = \inf\{t : H(t) = 1\}$. Then one can define the sequence of Wiener processes $\{\mathbb{W}_n(x), 0 \leq x < \infty\}_{n=1}^{\infty}$ such that when $n \to \infty$

$$\sup_{t \leq T} |n^{\frac{1}{2}}(F_n^*(t) - F(t)) - (1 - F(t))\mathbb{W}_n(d(t))| \overset{a.s.}{=} O(n^{-1/2} \log n),$$

where F_n^* stands for one of estimators F_n^{PL}, F_n^E, F_n and

$$d(t) = \int_{-\infty}^{t} [(1 - F)^2(1 - G)]^{-1} dF.$$

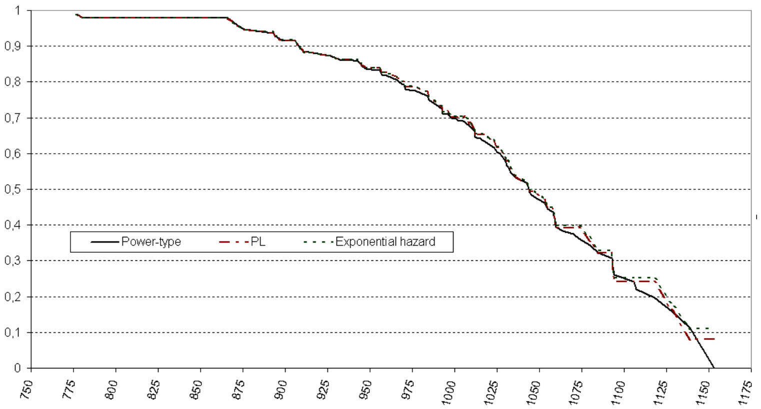
Here we state the convergence result in the form of strong approximation by the sequence of appropriate copies of limiting Gaussian process, with optimal rate. As consequence from here we obtain that $n^{\frac{1}{2}}(F_n^* - F)$ converges weakly in the Skorochod's space $\mathbb{D}(-\infty, T]$ to the mean-zero Gaussian process with covariance function $\sigma(t; s) = (1 - F(t))(1 - F(s))d(min(t, s)),\ t, s \leq T$. Thus we see that all three estimators are equivalent in the asymptotic sense. But as we see in [3-5] the estimator F_n has some peculiarities and even a better properties than F_n^{PL} and F_n^E do for all $n \geq 1$. Let's consider the following exponential representation for any right continuous d.f. (Gill (1980)):

$$1 - F(t) = exp\Big\{ - \int_{-\infty}^{t} \frac{dF(u)}{1 - F(u-)} \Big\} \prod_{s \leq t} (1 - \Delta\Lambda(s)),$$

where $\Delta\Lambda(s) = (F(s) - F(s-))/(1 - F(s-))$ and $F(s-) = lim_{u \uparrow s} F(u)$. Then we see that F_n^{PL} is a natural estimator for $\prod_{s \leq t}(1 - \Delta\Lambda(s))$, that is a discrete d.f.. On the other side, F_n^E and F_n are a natural estimators for continuous d.f. $F(t) = 1 - exp\Big\{ - \int_{-\infty}^{t} (1 - F)^{-1} dF \Big\} = 1 - (1 - H(t))^{R(t)}$, where $R(t) = -log(1 - F(t))/[-log(1 - H(t))]$ - relative risk function. Obviously, the relative risk estimators $F_n(t)$ and $G_n(t) = 1 - (1 - H_n(t))^{1 - R_n(t)}$ of $F(t)$ and $G(t)$ satisfy the empirical analogy of equality $(1 - F(t))(1 - G(t)) = 1 - H(t),\ -\infty < t < \infty$, that is $(1 - F_n(t))(1 - G_n(t)) = 1 - H_n(t),\ -\infty < t < \infty$. But for exponential hazard estimators $F_n^E(t)$ and $G_n^E(t) = 1 - exp\{-\sum_{j=1}^{n}(1 - \delta_{(j)})I(Z_{(j)} \leq t)/(n - j + 1)\}$ of $F(t)$ and $G(t)$, we have

$$(1 - F_n^E(t))(1 - G_n^E(t)) = exp\Big\{ - \sum_{j=1}^{n} \frac{I(Z_{(j)} \leq t)}{n - j + 1} \Big\} \neq 1 - H_n(t).$$

Moreover, for $t \geq Z_{(n)}, F_n(t) = 1$, but $F_n^E(t) < 1$. Therefore F_n is a correct estimator of continuous d.f. F than F_n^{PL} and F_n^E. In picture below we demonstrate plots of estimators $1 - F_n$, $1 - F_n^{PL}$ and $1 - F_n^E$ of survival function $1 - F$ using well-known Channing House data of size n=97(see [3-5]). Here, thin-solid line stands for $1 - F_n^E$, medium-one for $1 - F_n^{PL}$ and thick-solid line stands for $1 - F_n$. In monographies [3-5] of author one can find several extensions of estimators F_n, F_n^{PL} and F_n^E with full asymptotical results theory (weak convergence, law of itherated logarithm type strong consistency, weak and strong approximation, empirical Bayes approach ...) in competing risks models with random censorship from the right and both sides.



[1] Abdushukurov, A.A. (1998). Nonparametric estimation of the distribution function based on relative risk function. *Commun. Statist.:Theory and Methods*, **27**, N.8, pp. 1991-2012.

[2] Abdushukurov, A.A. (1999). On nonparametric estimation of reliability indices by censored samples. *Theory Probab. Appl.*, **43**, N.1, pp. 3-11.

[3] Abdushukurov, A.A. (2009). *Statistics of Incomplete observations: Asymptotical Theory of Estimation for Nonclassical Models.* University Press. Tashkent.

[4] Abdushukurov, A.A.(2011). Nonparametric estimation based on incomplete observation. International Enciclopedia of Statistical Sciences. Springer. Part 14. pp. 962-964.

[5] Abdushukurov, A.A.(2011). Estimation of unknown distributions from incomplete data and its properties. LAMBERT Academic Publishing.

[6] Akritas, M.G. (2000). The central limit theorem under censoring. *Bernoulli, 6,* N.6, pp. 1109-1120.

[7] Csörgő, S. (1996). Universal Gaussian Approximations under Random Censorship, *Ann. Statist.*, **24**, N.6, pp. 2744-2778.

[8] Gill, R.D. (1980). *Censoring and Stochastic Integrals*, Mathematical Centre Tracts, **124**. Amsterdam.

[9] Gill, R.D. (1994). Glivenko-Cantelli for Kaplan-Meier. *Math. Meth. of Statist.*, **3**, N.1, pp. 76-87.

[10] Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from Incomplete observations. *J.Amer.Statist.Assoc.*, **58**, pp. 457-481.

Estimation of density from indirect observation

V.Solev

In this paper we consider an observational scheme, as data are interval censored. We suggest a simple nonparametric estimator $\widehat{f_n}$ for unknown density f and under some appropriate condition prove the consistency of this estimator.

Estimation of multivariate survival functions by dependent censored data

N.T.Dushatov, R.S.Muradov

The problem on estimation of multivariate survival functional is considered at random dependent censored data on the right. For construction of estimators are used Archimedean copula functions. The property of uniform consistency of estimators is proved.

Keywords: random censorship, survival function, Sklar's theorem, Archimedean copulas, counting processes, martingales.

Part V
Application of Statistical Methods

Gradient statistical attack at block cipher RC6

A.S. Lysyak

This work covers an experimental research of statistical methods in cryptoanalysis on the example of the block cipher RC6. The given gradient attack is based on the statistical test "book stack", developed by B.Ya. Ryabko. The attack's circuit allows to reduce considerably a labour of input private key finding. The earlier known variants of attacks based on the test of hi-square made the big complexity. In the given work the efficiency researches of gradient attacks are conducted, the limits of its modern practical and theoretical applicability are shown (up to 9 rounds of cipher RC6), the mathematical dependences between effectively cracked rounds and quantity of demanded computing resources are received. Also in this operation the trial and error method of optimal parameters for the test is given, and also their influence on attack is shown; the time estimation of an attack and its dependence on test and size parameters of cipher text is researched; theoretical requirements for the computational capabilities necessary for realization of attack are shown.

Planning seismic networks

O. Omelchenko

This paper is dealt with a brief statement of a basis of the theory of optimal planning of seismic networks. Some concepts of such a planning of seismic networks are given. Some specific formulations of problems of planning of seismic networks are presented.

Keywords:optimal planning, seismic networks, hypocenters of earthquakes.

Practical application of forecasting method based on universal measure

P. Pristavka, B. Ryabko

In this article we describe and experimentally investigate a method to construct forecasting algorithms for stationary and ergodic processes based on universal measures (or the so-called universal data compressors). By the example of predicting the sunspot numbers and some other solar characteristics we show that the precision of thus obtained predictions is higher than for known methods.

Keywords: time series, nonparametric methods, universal measure, universal coding, solar activity, sea level, cross-rates.

Applying statistical methods to text steganography

I. Nechta, A. Fionov

This paper presents a survey of text steganography methods used for hiding secret information inside some covertext. Widely known hiding techniques (such as translation based steganography, text generating and syntactic embedding) and detection are considered. It is shown that statistical analysis has an important role in text steganalysis.

Keywords: Steganography, steganalysis, linguistic stegosystem, statistical attacks.

Non-linear probability models and problems of their application

V.V. Gubarev

Concepts of the "linear" and "non-linear" random signals, as phisical information carriers, and their mathematical models in the form of random variables, vectors, the continuous in time processes and the discrete series are considered. Concepts are coordined with methods of defining and characteristics of static and dynamic signals models. The examples underlining unfitness or small suitability of the "linear" characteristics in a non-linear reality are resulted; characteristics and models, suitable for non-linear situations and also examples of their application for identification of non-linear systems and estimation of communication parameters of multidimentional distributions of random vectors, processes, series.

Part VI
Robust Methods of Statistical Analysis

Robust estimation of qualitative response regression models

Alexander A. Kalinin, Daniil V. Lisitsin

Qualitative response regression models such as logistic regression are typically estimated by the maximum likelihood method. To improve its robustness, two special cases of the M-estimation based approach for quantitative continuous random variables were extended to the variant of qualitative and mixed variables modeling. Expressions of the score functions for polytomous regression models were derived. In according to results of the research some conclusions and practical recommendations were given.

Keywords: qualitative response, Bayesian dot contamination, polytomous regression, robust estimation, influence function

Statistical forecasting for censored autoregressive time series

Yu.S. Kharin, I.A. Badziahin

Problems of optimal statistical forecasting are considered for autoregressive time series observed under distortions generated by interval censoring. If the model parameters are unknown, then the maximum likelihood estimators are found and the "plug-in" forecasting statistic can be constructed. Numerical results are given.

Keywords: Autoregression, censoring, log-likelihood function, mean-square risk.

Robust estimation of count response regression models

Svetlana Yu. Dovgal, Daniil V. Lisitsin

This paper is concerned with models of event counts, particularly with the Poisson regression model examination. Robust methods of M-estimation parameters were researched. Expressions of the score function for Poisson model were given. Maximum likelihood estimation and M-estimation for model's parameters were compared by simulation.

Keywords: count data, Poisson regression, robustness, M-estimation, influence function.

Part VII
Statistical Simulation of Natural Processes

Modeling of nonstationary processes with periodic properties on basis of Markov chains

Nina A. Kargapolova, Lev Ya. Saveliev, Vasily A. Ogorodnikov

In this paper heterogeneous Markov model with two states and periodic transition probability matrix is considered. Expressions for limiting probabilities and distributions of long-term identical value runs are obtained. On basis of real data, model is applied to investigation of air temperature's long-term overshoots.

Keywords: heterogeneous Markov chain, limiting probabilities, air temperature.

Effective coefficients of Maxwell's equations with multiscale isotropic random conductivity and permittivity

O.N. Soboleva, E.P Kurochkina

The effective coefficients in Maxwell's equations are calculated for a multiscale isotropic medium by using a subgrid modeling approach. The correlated fields of conductivity and permittivity are mathematically represented by a Kolmogorov multiplicative continuous cascade with a lognormal probability distribution. The scale of solution domain is assumed to be large as compared with the scale of heterogeneities of the medium.

Keywords: Maxwell's equations, effective coefficients; subgrid modeling; multiscale random conductivity and permittivity.

Construction of "modelled" probabilistic densities

Anton V. Voytishek, Irena E. Graifer

In this study, we represent some recommendations for construction the probabilistic densities allowing efficient numerical realization of the sample values.

Keywords: Monte Carlo methods, numerical statistical simulation, numerical realization of sample value of random variable, probabilistic densities, method of inverse distribution function, numerical simulation of stochastic vectors, superposition method, majorant rejection method

Numerical analysis of SDE on supercomputers

S.S. Artemiev, V.D. Korneev

This paper deals with some problems of accuracy of algorithms for the numerical solutions of stochastic differential equations (SDEs) versus the size of the ensemble of trajectories simulated and on the mesh size of integrating the generalized Euler method. The problems of accuracy arise in estimating functionals of SDE solutions with increasing variance, highly asymmetric distributions, and an indefinite time of arrival of trajectories of solutions at the boundaries of given domains. Some ways of parallelization of statistical algorithms on a multiprocessor cluster are described. Results of numerical experiments performed on a supercomputer available at the Siberian Supercomputer Center are presented.

Keywords: stochastic differential equations, statistical algorithms, parallelization, supercomputer, cluster, van der Pol equation, phase trajectory, stochastic oscillators.

Monte Carlo modeling of the radiation transfer in stochastic scattering media

Boris A. Kargin

The problems of statistical simulation of light propagation in stochastic scattering media as applied to the problems of optics of aerosol cloudy atmosphere are considered. A set of Monte Carlo algorithms, allowing the construction of numerical models for the field of multiply scattered optical radiation in the aerosol atmosphere and stochastic cloudiness has been provided for the purpose. A special attention has been paid to solving the problem of optimization of Monte Carlo algorithms. The optimization is based on the method of ”dependent trials”.

Keywords: stochastic media, transfer equation, Monte Carlo method.

Application of the modified method of the maximum section for statistical modeling of systems with a separated time

Tatyana A. Averina

The algorithm for statistical modeling of systems with a separated time, which can be described as a system with a distributed change of structure has been constructed. The offered algorithm is based on numerical methods of the solution to the stochastic differential equations and uses the modified maximum cross section method when the intensity of transition depends on a vector of state.

Keywords: Numerical methods, stochastic differential equations, systems with random structure, systems with a separated time, maximum cross section method.

Monte Carlo modeling in problems of lidar remote sensing of crystal clouds from satellites

Boris A. Kargin, Arsenii B. Kargin, M.V. Lavrov

Laser sensing is an effective way of studying optical properties of various atmospheric structures. If we consider strongly scattering media, like clouds, there arises the necessity of taking into account the effects of multiple scattering which changes the space and time characteristics of the light pulse. The Monte Carlo method is the most convenient one for obtaining practical results in such problems. In this paper two problems were solved. One is constructing an adequate optical model of crystal clouds taking into account optical anisotropy of the medium. The other is Monte Carlo modeling of laser radiation transfer in such a medium. The form and duration of light pulses reflected by clouds (lidar returns) are obtained by the Monte Carlo method in the case of single layer continuous crystal cloud and double layer continuous cloudiness (a crystal cloud of highest level is located above a drop cloud).

Keywords: Monte Carlo method, transfer equation, laser radiation, crystal clouds, optical anisotropy.

Sensitivity of a diffusion process to the moving boundary parameters

Sergey A. Gusev

A 2D parabolic boundary problem with moving boundary is considered in the paper. The moving part of the boundary is approximated by a broken line. A statistical modeling method for estimation of the solution of this problem and its parametric derivatives is proposed. Desired estimates are obtained as results of numerical simulation of trajectories of the corresponding to the problem diffusion process and its derivatives with respect to parameters determining the boundary motion. We set a biunique correspondence at any time point between the moving boundary domain and a fixed domain which coincides with the initial state of the moving boundary domain. In calculations the numerical simulation of the diffusion process is performed in the fixed domain.

Keywords: moving boundary problem, stochastic differential equations, statistical modeling, Euler method .

Optimization of a time-sharing queueing process in random environment with means of computer simulation

Andrei V. Zorine

A service of conflict flows with time-sharing algorithm with readjustments in random environment is considered. A mathematical model is constructed as a homogeneous denumerable discrete-time Markov chain. Conditions for the stationary distribution existence are found. A computer simulation model is also built. With means of computer simulation a switching function can be found which minimizes several cost-type objective functionals.

Keywords: Conflict flows, time-sharing algorithm with readjustments, optimal switching.

Numerical simulation of the sea surface and extreme ocean waves with stochastic spectral models

Sergei M. Prigarin, Kristina V. Litvenko

In this paper, we make an attempt to apply stochastic spatial-temporal conditional spectral models of the sea surface undulation to study features of formation and development of the ocean extreme waves.

Keywords: numerical simulation, sea surface undulation, extreme waves, rogue waves, spectral models, random fields.

Stochastic models of broken clouds (A few simulation examples)

Sergei M. Prigarin

This paper deals with numerical simulation of stochastic indicator fields corresponding to satellite images of broken clouds in the atmosphere. Numerical spectral models and a method based on thresholds of the Gaussian functions are used to simulate the indicator fields. An additional example of simulation of a binary pattern of granules on the photosphere of the Sun is presented as well.

Keywords: Broken clouds, stochastic geometry, numerical simulation, random fields, threshold and spectral models, Sun photosphere granules.

Statistical modeling method for kinetic traffic flow model with acceleration variable

Aleksandr Burmistrov, Mariya Korotchenko

We consider a kinetic vehicular traffic flow (VTF) model with acceleration variable and study evolution of the N-particle systems, which are governed by a homogeneous Boltzmann-like equation. For this model we obtain a linear integral equation of the second kind and suggest to solve it by the statistical modeling method. The numerical results show that the approach to simulation suggested by the authors is reasonable to apply to the vehicular traffic problems. Moreover, authors managed to exclude from simulation an external to the initial model parameter – a discrete time interval, which was used previously. It resulted in a simpler simulation process.

Keywords: N-particle system, vehicular traffic flow, Monte Carlo method.

Stochastic models of the price series for trade algorithms

Mikhail A. Yakunin

We investigate stochastic models of series of price increments for trade algorithms. The models are constracted based on stochastic differential equations. The estimates of unknown parameters of the model of price increments with jumps are obtained by using the method of moments.

Keywords: price increments, probability density, estimates of parameters.

Monte Carlo modeling the radiation heat transfer with temperature correction

Sergei A. Brednikhin, Boris A. Kargin

In the paper the results of modeling of radiation heat transfer in systems of dust protoplanetary clouds with the Monte Carlo method using an algorithm of temperature correction [1] with the NMC code are presented. The results of a series of standard calculations are presented. Also within the bounds of the paper an attempt is made to apply the mentioned modeling method with use of external iterations to planet atmospheres by means of introducing a gas component into the modeled system. The results of calculations of altitude distribution of temperature in a simplified model of Venus' atmosphere without taking into account convection and reradiation with the gas component are presented. The obtained results show a qualitative correspondence with experimental data.

Keywords: Radiation heat transfer, dust circumstellar clouds, Venus, Monte Carlo method.