The Second International Conference "Problems of Cybernetics and Informatics"
September 10-12, 2008, Baku, Azerbaijan. The plenary paper
www.pci2008.science.az/3/02.pdf

THE RESAMPLING APPROACH APPLICATION TO COMPLEX SYSTEMS ANALYSIS AND FORECASTING

Alexander Andronov

Riga Technical University, Riga, Latvia, *Aleksandrs.Andronovs@rtu.lv*

One can say about resampling methods by citation of Ph.Good from [17]: "The resampling methods – permutation, cross-validation, and the bootstrap – are easy to learn and easy to apply." And further: "Introduced in the 1930s, the numerous, albeit straightforward, calculations resampling methods require were beyond the capabilities of primitive calculators then to use. … Today, with a powerful computer on every desktop, resampling methods have resumed their dominant role ….".

We see that resampling methods include many of numerical statistical approaches. We restrict ourselves by the situation which can be described in the simplest manner by the following way.

Let us suppose that a considered task is to estimate an expectation θ of a function f of independent random variables $X_1,X_2,...,X_m : \theta = Ef(X_1,X_2,...,X_m)$. This function describes performance characteristics of a complex system. For independent variables $X_1, X_2,..., X_m$, sample populations $H_1, H_2,..., H_m$ are available as the primary data.

The traditional parametrical approach supposes three stages. 1) Hypotheses about distributions of random variables $\{X_i\}$ are made. 2) Unknown parameters of these distributions are estimated on the basis of observations $\{H_i\}$. 3) The estimated distributions are used for the expectation θ estimation. For complex systems it is difficult to get an analytical solution, so usually a simulation is used. For that, the random variables $\{X_i\}$ are generated using estimated distributions and the function f values are calculated. An average of the calculated values is used as the expectation θ estimate.

An alternate approach is based on the *resampling*. It supposes the following. 1) The first and the second stages are absent and the probabilistic distributions are not estimated. 2) At the third stage, immediate simulation of the function f is realised by means of the sample data direct usage. Here, the simulation procedure coincides with the traditional one with the only difference: random variables are not generated by random number generators in accordance with the estimated probabilistic distributions but extracted from the present samples at random.

Further, we will interpret the resampling method as the above described approach and its modifications. It is possible to distinguish resampling methods varieties via two directions: the technique of a resampling application and the aim of a statistical procedure. In the first case, there exists a resampling with and without replacement. In the second case, the problems of point (as above) or interval estimation, hypothesis testing, classification and so on can be considered.

A practical usage of the resampling methods is very simple and intuitive. The book [17] can be used as an excellent manual here. On the other hand, a theoretical investigation of the resampling methods properties is a complex problem [12, 13, 15]. Below some results of the author for the case of small samples will be reviewed.

Firstly, we consider a problem of *the point estimation.* One of the earliest works on the application of the simple resampling to system reliability point estimation belongs to V.Ivnitsky. *The hierarchical resampling* has been introduced in [4] for the case when a tree presents the function *f.* The resampling procedure includes k trials. During the v-th trial, $\nu = 1, 2,..., k$, some element $X_{j(i,\nu)}$ is extracted (with or without replacement) from sample H_i at random, i =1, 2, …, m. After k trials we calculate empirical mean as

$$\tilde{\theta} = \frac{1}{k} \sum_{v=1}^{k} f(X_{j(1,v)}, X_{j(2,v)}, ..., X_{j(m,v)}) . \quad (1)$$

The main problem is to calculate a variance of the corresponding estimate and to compare it with those of traditional estimates. The authors introduce a *notion of ω - pair* that turns out to be very useful. Let $M = \{1, 2, \ldots, m\}$ be the set of integers, $\omega \subset M$ – a subset of M, $\boldsymbol{X} = (X_1, X_2, ..., X_m)$ and $\boldsymbol{X'} = (X'_1, X'_2, ..., X'_m)$, where $X_i,\ X'_i \in H_i$, $i = 1, 2, \ldots, m$, be two sub-samples. The sub-samples $\boldsymbol{X}$ and $\boldsymbol{X'}$ are said to be a *ω-pair* if $X_i = X'_i$ for $i \in \omega$ and $X_i \neq X'_i$ otherwise. Let $cov(f(\boldsymbol{X}), f(\boldsymbol{X'})|\ \omega)$ be the conditional covariance for $f(\boldsymbol{X})$ and $f(\boldsymbol{X'})$ given the condition that $\boldsymbol{X}$ and $\boldsymbol{X'}$ union the ω-pair, $p(\omega)$ be a probability to union the ω-pair. Then,

$$Cov\big(f(X), f(X')\big) = \sum_{\omega} Cov\big(f(X), f(X')|\omega\big)p(\omega) , \quad (2)$$

where a sum is taken over a set of all ω-pairs.

Now, to calculate this covariance (and therefore, the variance of the estimate (1)) it is necessary to calculate the probability $p(\omega)$ for all possible ω. The corresponding technique depends on the used resampling procedure.

Above, each random variable X_i had its unique sample population H_i. Later, a more general case has been considered, when some different random variables $\{X_i\}$ have the same sample population H_j. Another generalisation of considered positing concerns dependence in the sample data [9] and control of the extraction procedure [2].

Secondly, we consider a problem of *the interval estimation* for expectation $\theta = Ef(X_1, \ldots, X_m)$, that corresponds confidence probability γ. This problem is a main subject of the mathematical statistics and reliability theory. Often the bootstrap approach is applied for confidence interval construction [13], [14], [15]. In the papers [5, 7] the resampling method is used. One makes use of the following procedure. Series of experiments are produced. Each experiment includes k trials, which have been described above. After k trials we calculate empirical mean θ_l for the current, for instance the l-th, experiment. Then, we return all extracted elements into the corresponding sample populations and repeat the described experiment r times, getting a sequence of the estimates $\theta_1,\ \theta_2, ..., \theta_r$. It gives us the order statistics $\theta_{(1)},\ \theta_{(2)}, ..., \theta_{(r)}$ and corresponding α-quantile $\theta_{(\alpha r)}$ of estimate $\tilde{\theta}$ distribution (it is supposed αr is whole number). We set $\alpha = 1 - \gamma$ and accept $(\theta_{(\alpha r)}, \infty)$ as γ-confidence upper interval for the original value θ.

In the paper [5] a two-dimensional simple case is considered when $f(X_1, X_2)$ is the indicator function of the random event $\{X_1 < X_2\}$. For instance, X_1 means a value of shock, X_2 means a strength of a construction or X_1 means a term of the exploitation, X_2 means a lifetime of a construction. Our aim is to construct upper confidence interval $\big(\theta_{(\alpha r)},\ 1\big)$ for $\theta = P\{X_1 < X_2\}$ and to calculate a true probability of covering $R = P\big\{\theta_{(\alpha r)} \le \theta\big\}$. For that purpose a notion of *protocol* is used. Let $H_1 = \big\{x_1^{(1)}, x_1^{(2)}, ..., x_1^{(n}\big\}$ and $H_2 = \big\{x_2^{(1)}, x_2^{(2)}, ..., x_2^{(n')}\big\}$ are an order presentation of initial samples H_1 and H_2. The value $x_1^{(i)}$ is named as *the i-th point of a protocol*, $i = 0, 1, \ldots, n + 1$, $x_1^{(0)} = -\infty$, $x_1^{(n+1)} = \infty$. Let c_i be a number of values $x_2^{(j)}$ which satisfy the inequalities $x_1^{(i)} < x_2^{(j)} \le x_1^{(i)}$. Then, the $(n+1)$–dimensional vector $C = \big(c_0, c_1, ..., c_n\big)$ is said to be *a protocol*. Using the protocols simplifies a calculation of the covering probabilities. The following sequence of elements is used in the corresponding numerical procedure:

- the algorithm of enumeration of all protocols,

- the probability $P_{\boldsymbol{C}}$ to have fixed protocol $\boldsymbol{C}$,
- the conditional probability q_C of event $\{f(X_1, X_2) = 1\}$ by condition that protocol $\boldsymbol{C}$ is fixed,
- the conditional probability $\rho_{\boldsymbol{C}}$ of the event $\{\theta_l < \theta\}$ for the l-th experiment given $\boldsymbol{C}$,
- the conditional covering probability $R_{\boldsymbol{C}}$ by condition that protocol $\boldsymbol{C}$ is fixed.

Finally, the unconditional probability of the covering is calculated as

$$R = \sum_C P_C R_C .$$

The numerical results show that true probability of covering is close to appointee value.

A dissemination of this approach on multivariate case is described in the paper [7]. Here the function $f(x_1, ..., x_m)$ describes an efficiency of a logical system, when f is a predicate (with value 0 or 1), that contains real numbers as arguments $x_1, ..., x_m$, operations over them (such as minimum, maximum, order statistics), the sub-predicates < , >, = and so on.

In the paper [6] the resampling approach is used to statistical inferences of order statistics, those have an important role in various applications: insurance, reliability theory, storage control and so on. It is sufficient, for example, to refer to the well known *k-out-of-n-system* [16]. In fact, an elaborated method allows calculating the distributions of order statistic estimates and on this basis calculating the confidence intervals for quintiles of order statistics.

Thirdly, it is necessary to say about an application of the resampling method to a statistical analysis of stochastic processes. The total approach is here the same as earlier mentioned. To simulate one trajectory of a considered process, we extract (without a replacement) necessary random variables from the given samples at random and calculate the efficiency characteristics of interest θ_l (it corresponds to one trial). Then, we return all extracted values into the initial samples and repeat this procedure many times. As a result, the estimate sequence $\theta_1, \theta_2, ..., \theta_r$ gives a base for various statistical inferences. In a case of the multiple linear regressions [1], an estimate of unknown coefficients during one trial is performed using a part of the given observations. Now, to get a robustness estimate, we can take the median of getting coefficient estimates. Numerical results [1] testify an advantage of such an approach. A case of a nonparametric interval estimation of the regression function is considered in the paper [10].

Such successful results have place for an analogous application to the estimates of the renewal function [8], system reliability [11], efficiency characteristics of queuing systems and so on. Specifically it was showed that resampling estimates of the renewal function have less bias in comparison with usual estimates that use the empirical distribution function of time between renewals and its successive convolution [8]. In the paper [11] the following known reliability problem [16] was considered. The model supposes two types of failures – *initial* and *terminal failures*. Initial failures (or damages) appear according to homogeneous Poisson process with the rate λ. Each initial failure degenerates into a terminal failure after a random time B. So if an initial failure appears at time τ_i then a terminal failure appears at the instant $B_i + \tau_i$. The terminal failure and the corresponding initial failure are eliminated instantly. We assume that $\{B_i\}$ is mutually independent identical distributed random variables, independent on $\{\tau_i\}$. We take an interest in the number of initial failures $Y(t)$ at time t (which have not degenerated to the terminal failures) and the number of terminal failures $Z(t)$ that have occurred till time t. Let $EY(t)$ and $EZ(t)$ be the corresponding expectations, $P_i(t) = P\{Y(t) = i\}$, $R_i(t) = P\{Z(t) = i\}$ be the corresponding probability distributions, i = 0, 1, … . Our aim is to estimate these characteristics on the basis of the sample $H_1 = \left\{X_1^{(1)}, X_1^{(2)}, ..., X_1^{(n}\right\}$ of the intervals between initial failures appearances and the sample of $H_2 = \left\{B_1, B_2, ...B_{n'}\right\}$.

The authors testify that the proposed resampling-approach is a good alternative to the traditional plug-in estimation. It is especially remarkable increasing the size of the given samples.

Literature

1. H. Afanasyeva, A. Andronov. On Robustness of Resampling Estimators for Linear Regression Models. *Communication on Dependability and Quality Management*, Volume 9, Number 1 (2006), pp. 5–11.
2. A. Andronov, Yu.Merkuryev. Controlled Bootstrap Method and Its Application in Simulation. In *Proceedings of the 11th European Simulation Multiconference*, SCS, Istanbul, Turkey (1997), pp. 160–164.
3. A. Andronov, M. Fioshin. Simulation technology under small samples for unknown distributions. In: *Proceedings of 10 GI/ITG Special Interest Conference "Measurement, Modelling and Evaluation of Computer and Communication Systems",* Trier (1999), pp. 153–162.
4. A. Andronov, Yu. Merkuryev. Optimisation of Statistical Sample Sizes in Simulation. *Journal of Statistical Planning and Inference*, 85 (2000), pp. 93–102.
5. A. Andronov. On Resampling Approach to a Construction of Approximate Confidence Intervals for System Reliability. In: *Third International Conference on Mathematical Methods in Reliability. Methodology and Practice*. Trondheim, Norway, Norwegian University of Science and Technology (2002), pp. 39–42.
6. A. Andronov, H. Afanasyeva. Resampling-based nonparametric statistical inferences about the distribution of ordered statistics. In: *Transactions of the XXX1V International Seminar on Stability Problems for Stochastic Models.* Jurmala, Latvia (2004), pp. 300-307.
7. A. Andronov, M. Fioshin. (2004). Applications of resamping approach to statistical problems of logical systems. *ACTA et Commentationes Universitatis Tartuensis de Mathematica*, 2004, Vol. 8 (2004), pp. 63–71.
8. A. Andronov. Resampling-Estimators of the Renewal Function. In: *Transactions of the XXV International Seminar on Stability Problems for Stochastic Models.* Maiori (Salerno), Italy (2005), pp. 17–24.
9. A. Andronov. On nonparametric estimation of expectation of random variables function under dependent observations. *Journal of statistical planning and inference*, 137 (2007), pp. 3828–3837.
10. A. Andronov. On nonparametric interval estimation of a regression function based on the resampling. *Computer Modeling and New Technologies*, Vol. 11, No. 1 (2007), pp. 47-54.
11. A. Andronov, H. Afanasyeva, M. Fioshin. Statistical Estimation for a Failure Model with the Accumulation of Damages. *Journal of statistical planning and inference* (to be published).
12. Yu.K. Belyaev. Computer Intensive Method Based on Resampling in Analysis of Reliability and Survival Data. In *Resent Advances in Reliability Theory: Methodology, Practice and Inference*, N.Limnios, M.Nikulin (Eds.). Birkhauser, Boston (2000), 183-198.
13. A.C. Davison, d.V. Hinkley. Bootstrap methods and Their Applications. Cambridge, Cambridge, Press University (1997).
14. T.J. DiCiccio, B.Efron. Bootstrap Confidence Intervals. *Statistical Science*, 11, (1996), pp. 189–228 pp.
15. B. Efron, R.J.Tibshirani. Introduction to the Bootstrap. Chapman&Hall, London (1993).
16. I. Gertsbakh. Reliability Theory: With Applications to Preventive Maintenance. Springer, Berlin (2000).
17. P. Good. Resampling method: a practical guide to data analysis. Second edition, Birkhauser, Boston (2001), 238 p.
18. C.F.J. Wu. Jackknife. Bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14, 3 (1986), pp. 1261–1295.