Interval Computations
No 1(3), 1992

INTERVAL STATISTICS

Alexander I. Orlov

A survey of basic concepts of statistical analysis of data that are intervals is given. The difference between interval and classical statistical inference is highlighted. The concept of rational sample size is discussed. The development of interval statistics in Russia in the period 1980-1990 is outlined.

ИНТЕРВАЛЬНАЯ СТАТИСТИКА

А.И. Орлов

В статье приводится обзор основных концепций статистического анализа данных, являющихся интервалами. Показано различие между статистическими выводами и классическими выводами. Обсуждена концепция рационального объема выборки. Показано развитие интервальной статистики в России за 80–90е годы.

Mathematics and statistics are related sciences. One may debate their correlation. Along these lines, we have claimed that applied statistics is branch of cybernetics [1]. But it is more important to find common ground for interaction.

In the period 1970–1990 in our country, interval mathematics and interval statistics developed independently. At the conference entitled "Problems of Applied Mathematics" held in Saratov, May, 1991, representatives of the two scientific fields met and initiated interesting scientific discussions. This paper is a survey of basic concepts and results of interval statistics.

One of the statistical research subjects in 1970 was the stability of statistical inference with respect to admissible deviations of initial data and with respect to premises of a model [2]. In this context, it is natural to consider fuzzy numbers instead of real numbers for initial data [3]. In that epoch, there was enthusiasm for Zadeh's theory of fuzziness [4]. Although it was immediately shown that the theory of fuzziness reduced [2] to the theory of random sets, that is, to one of the branches of probability theory, it was nonetheless tempting, for practical purposes, to replace real numbers by fuzzy numbers in our applications.

In practical realizations of this concept, the question of a membership function specification of fuzzy numbers arose. It is clear that the simplest possible specification is a step function taking the value 1 inside some interval and a value 0 outside this interval. In other words, an observation result is not a number, but an interval. In almost the same way, a number of scientists were attracted to interval statistics, a branch of applied mathematical statistics in which initial data are given as intervals but not as numbers.

Interval mathematics is not a fancy of a mathematician in search of new statements, but an answer to practical needs. There is good reason for its development not by academic and university scholars, but by those who work with real world problems. For example, when elaborating the state standard "Applied statistics. Determination rules for estimates and confidence limits for gamma distribution parameters" [5], data on the lifetime of cutting tools, up to limiting conditions, was analyzed. The data were given up to 0.5 hours, with a lifetime in the interval from 9 to 130 hours. The question of how this uncertainty in the data influences statistical inference arose. An appropriate theory came into being at the beginning of 1982. On this basis, the choice rules for an estimating technique, that is, the method of moments or the maximum likelihood method, are formulated. The aforementioned theory was published later in [6].

Further development of these concepts resulted in construction of a new statistical theory called "realistic statistics" [7,8]. Its results and recommendations differ in a principal way from those of classical mathematical statistics. In particular, they differ in the following points: there are no consistent estimates; it is not advisable to increase the number of observations beyond a certain limit, called a "rational sample size"; the method of moments may be better than the maximum likelihood method.

and so on. Consider the basic concepts of a "realistic statistics" in a very simple example of estimating expectation.

As in the classical case, let the independent equally distributed random variables $x_1, x_2, \ldots, x_n$ describe real phenomena. But a statistician knows different random variables

$$y_i = x_i + \varepsilon_i \tag{1}$$

distorted by measurement errors ε_i, $i = 1, 2, \ldots, n$.

Suppose that the absolute value of the errors is bounded by a known constant

$$|\varepsilon_i| < \Delta, \ i = 1, 2, \ldots, n. \tag{2}$$

In other words, an observation result is not a number y_i but an interval $[y_i - \Delta, y_i + \Delta]$. And, on the basis of observations of the set of intervals, one must draw a conclusion about the expectation $E(x_i)$ of the actual random variables $x_1, x_2, \ldots, x_n$.

How is a constant Δ found? It may often be found from the technical certificate of a measurement tool, or from the form of the data, as in the case of the lifetime of the cutting tools described above.

In classical applied mathematical statistics, the sampling arithmetic mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3}$$

is used as an estimate of the expectation.

If the x_i have a variance, the sampling arithmetic mean $\bar{x}$ is asymptotically normally distributed:

$$E(\bar{x}) = E(x_i), \quad D(\bar{x}) = \frac{\sigma^2}{n}, \quad \sigma^2 = D(x_i). \tag{4}$$

Thus, $\bar{x}$ is a consistent estimate of $E(x_i)$, a confidence interval for $E(x_i)$ with a fiducial probability γ of the form

$$\left[\bar{x} - u(\gamma)\frac{\zeta}{\sqrt{n}}; \bar{x} + u(\gamma)\frac{\zeta}{\sqrt{n}}\right], \tag{5}$$

for large n. Here, ζ^2 is a sample variance,

$$\zeta^2 = \frac{1}{n-1} \sum_{1 \le i \le n} (x_i - \bar{x})^2, \tag{6}$$

and $u(\gamma)$is a quantile of a normal distribution of order $(1+\gamma)/2$, that is, a root of the equation

$$\int_{-\infty}^{u(\gamma)} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{z^2}{2}\}dz = \frac{1+\gamma}{2}.$$

If a distribution of x_i is known, the asymptotic confidence interval **(5) may** be replaced by **the exact** one. For example, if x_i has a normal **distribution**, then in (5), **a** quantile of a normal distribution must be **replaced** by a quantile of a Student distribution.

Let us move on to estimation of an expectation within the framework **of realistic** statistics. Suppose for this purpose that one uses a sampling **mean of** available arithmetic data

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}. \tag{7}$$

If the x_i have a variance, then by (2) the random variables y_i also have a variance. Thus, $\bar{y}$ has an asymptotically normal distribution as $n \to \infty$, and

$$E(\bar{y}) = \frac{E(y_1) + E(y_2) + \cdots + E(y_n)}{n}. \tag{8}$$

Relation (1) implies that

$$E(x_i) - \Delta \le E(\bar{y}) \le E(x_i) + \Delta, \tag{9}$$

where the left and the right boundaries are attained by the values Δ or $-\Delta$ in the case of systematic error. Regarding the variance of a random variable y_i, one can show that for small Δ, this variable differs little from the variance of $\bar{x}$. The proof proceeds by arguments similar to those of [6] in the technically more complicated case of estimating parameters of a gamma distribution. Therefore,

$$D(\bar{y}) = \frac{\sigma^2}{n} + o(\frac{\Delta}{n}). \tag{10}$$

Formulas (9) and (10) imply

$$\sup E(\bar{y} - E(x_i))^2 = \Delta^2 + \frac{\sigma^2}{n} + o(\frac{\Delta}{n}), \tag{11}$$

where the supremum is taken over all ε_i satisfying condition (2).

Relation (11) allows us to draw several non classical conclusions. First, the left-hand side is always greater than Δ^2, and therefore does not tend to 0 as the size of the sample increases. This means that there is an objective threshold for the accuracy of the estimating parameters (estimating expectation here). In other words, a sampling mean is not a consistent estimate of an expectation.

This conclusion is well known by itself. Yet, Gnedenko and Hinchin discussed this topic in "An Elementary Introduction to Probability Theory" [9, p.p. 120–121]. Unfortunately, their reasoning was not based on mathematical statistical theory. That reasoning is not new to the theory of metrology, nor is it new to metrologists [10].

Eliasberg, a specialist in the field of cosmic research, wrote: "With good reason, consistency of statistical estimates can be considered as one of the "myths of the twentieth century" [11]. However, courses in mathematical statistics presently treat consistency, unbiasedness and efficiency, while the theory of realistic statistics, taking into account errors in observations, is just being created.

A second conclusion from relation (11) is connected to the concept of "rational sample size." To what extend should we increase n? A principle of "equating errors" was suggested in [2]. According to this principle, it is advisable to equate errors due to measurement uncertainty according to (1), and due to statistical uncertainty according to (10). In other words, it is proposed to equate the two principal members of the sum in relation (11), and to find a rational sample size n_{rat} by the formula

$$n_{rat} = (\frac{\sigma}{\Delta})^2 \approx (\frac{s(y)}{\Delta})^2, \quad s^2(y) = \frac{1}{n-1} \sum_{1 \le i \le n} (y_i - \bar{y})^2. \qquad (12)$$

Therefore, everything depends on the ratio of the mean square deviation σ of a random variable to the maximum measurement error Δ in the values of this random variable. For example, if $\Delta = 0.1\sigma$, then $n_{rat} = 100$.

We have considered a very simple example. This comparison of results of classical and realistic statistics has been done for a number of algorithms in [6–8], and will later be carried out for all principal algorithms of applied statistics. Software for realistic statistics, an interactive

system REST, is being developed by author jointly with Dr.Los'. There is a "parallelism" in statistics: a classical algorithm is developed jointly with a realistic one. In particular, a confidence interval (5) is displayed together with the realistic confidence interval

$$[\bar{y} - u(\gamma)\frac{s(y)}{\sqrt{n}} - \Delta; \bar{x} + u(\gamma)\frac{s(y)}{\sqrt{n}} + \Delta], \qquad (13)$$

The research program of developing realistic statistics [7] is far from completion. A current task is to enlist new specialists for this research.

We make the following remarks concerning the problem setting in realistic statistics. Besides restrictions on error, restrictions on relative error have been considered [5–8]. Furthermore, the form of the restrictions itself can be different. For example, in [6], the behavior of algorithms when the ε_i are all independent random variables was analyzed. Roughly speaking, it turned out that an expectation of an ε_i played the part of Δ. There are some unnecessary simplified models in which is assumed that x_i and ε_i are independent, or the ε_i are normally distributed. Such models would not be applied in real problems [10].

Lapidus, Rozno [12, 13], and Leifer [14] from the Gorky branch of the State Institute of Normalization in Technology (now, Nizhny Novgorod Branch of VNIISOT) have shown a constant interest in the effect of uncertainty of measurement results on statistical inference. Their work can be directly applied like Sher's work (Vladivostok). The latter researcher represented the estimates, obtained from captains of a fishing fleet in the form of relations (1) and (2), then processed them for managing the fleet [15].

The most important center of interval statistics is the Department of Automatics of the Moscow Power Engineering Institute. There, a research group has been active since the early 1980's under the guidance of Professor Voshchinin. Still headed by Prof. Voshchinin, this group now includes tens of specialists from CIS, Bulgaria, China, and other countries. One of the fundamental problems treated by this group is the search for a function describing the dependence due to interval data. Another fundamental problem studied there is that of optimizing on the basis of the dependence found [16, 17].

The following problem is important in practice. Let y be a quantity to be increased (for example, the yield of the target product in an industrial

chemical process). This quantity y depends on quantities $x_1, x_2, \ldots, x_n$ that can be controlled. There are initial data $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$, $i = 1, 2, \ldots, n$, and the model

$$y_i = f(x_{1i}, x_{2i}, \ldots, x_{ki}; a_1, \ldots, a_m) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \qquad (14)$$

where $a_1, \ldots, a_m$ are unknown parameters. After estimating the parameters from the statistical data, we can find a solution for an optimization problem and obtain $x_1, x_2, \ldots, x_k$ such that y attains the maximum.

If the ε_i in (14) are independent and equally normally distributed, then there is a regression analysis, generally nonlinear, in the ordinary setting. But if ε_i the satisfy the restriction (2), we have a problem of interval statistics, considered in [16,17].

Fragmentation in the development of interval statistics is due to the fact that specialists and research groups have worked independently. Different approaches came together in the discussions on the paper of Voshchinin, Bochkov and Sotirov [17] with participation of Borodyuk, Demidenko, Letsky, Orlov, Legostaeva and Kuznetsov. The first three of these related an interval approach to classical regression analysis. They concluded that it is advisable to continue study in interval statistics. We have analyzed the development of applied statistics in the nineteenth and twentieth centuries, and have shown that interval statistics deserves to be developed. Formulas for confidence intervals and rational sample size for estimating expectation and variance (see below) resulted from these studies. Together with [5–8], a comment [17, p.p. 86–89] represents an important publication on the approach studied.

The approach of Shiryaev and Blagoveshchensky et al was considered in the comment by Legostaeva [17, p.p. 90–93]. In models represented by (14), intervals arise due not to restrictions of the form (2) imposed on the ε_i, but due to the fact that a function f specifies the true regression only to a certain precision known by a statistician. With this approach, one succeeds in finding minimax estimates [18], for example, such that a mean square error reaches the minimum (see (11)). Preferring to study ordinary algorithms, we have not worked in such a setting. It is clear that interval statistics may stimulate many analyses of the type of [18] dealing with resolution of hard minimax problems. It is important that settings have significant applications.

In the above scheme, we began with ordinary probability theory (axioms of Kolmogorov). One also can construct probability theory on

the basis of interval expectations. This approach has been chosen by Kuznetsov [19].

Up to this point, interval statistics has been developed practically without interaction with interval mathematics or with the theory and practice of interval computations [20]. In [7], measurement errors were contrasted with computational ones, while in [6], they were studied jointly. There is a hope that the accumulated experience in interval mathematics will be useful in interval statistics, which will become an important applied branch of interval mathematics.

References

1. A.I. Orlov, *On Development of Applied Statistics. Modern Problems of Cybernetics (Applied Statistics)*, Znanie, Moscow, 1981, pp. 3–14. (In Russian)
2. A.I. Orlov, *Stability in Social-Economic Models*, Nauka, Moscow, 1979. (In Russian)
3. P.B. Shoshin, *Fuzzy Numbers as a Means of Describing Subjective Quantities. Statistical Methods for Analysis of Expert Estimates*, Nauka, Moscow, 1977, pp. 234 – 250. (In Russian)
4. L. Zadeh, *The Concept of a Linguistic Variable and its Application to Approximate Reasoning*, Elsevier, New York, 1973.
5. GOST 11.011-83, *Applied Statistics. Rules of Determination of Estimates and Confidence Limits for Gamma-distribution Parameters*, Izdatelstvo Standartov, Moscow, 1984. (In Russian)
6. A.I. Orlov, *On the Influence of Observation Errors on the Properties of Statistical Procedures (in the case of the gamma distribution)*, Statistical Methods for Estimating and Testing of Hypotheses, Perm State University, Perm, 1988, pp. 45–55. (In Russian)
7. A.I. Orlov, *On Development of Realistic Statistics. Statistical Methods for Estimating and Testing of Hypothesis*, Perm' State University, Perm', 1990, pp. 89–99. (In Russian)
8. A.I. Orlov, *Some Algorithms of Realistic Statistics. Statistical Methods for Estimating and Testing of Hypothesis*, Perm' State University, Perm', 1991. (In Russian)
9. B.V. Gnedenko and I.Ya. Hinchin, *Elementary Introduction to Probability Theory*, Nauka, Moscow, 1976.
10. P.V. Novitsky and I.A. Zograf, *Error Estimating for Measurement Results*, Energoatomizdat, Leningrad, 1985. (In Russian)
11. P.E. Eliasberg, *Measurement Information. How much of it is necessary, how should it be processed?*, Nauka, Moscow, 1983. (In Russian)
12. V.A. Lapidus and M.I.Rozno, *Models and Methods of Accounting for Uncertainties*

in a Result of Control and Trials, Estimating Quality Characteristics of Complex Systems and Systems Analysis. Theses for IX Interdepartmental (II All-Union) Technical-Scientific Seminar, Academy of Science, Moscow, 1980, pp. 104–105. (In Russian)

13. M.I. Rozno, *Study, Development and Standardization of Methods of Decision Making According to Trial Results, which Guarantee Production Quality. Autoreferat*, VNIINMASH, Moscow, 1986. (In Russian)
14. L.A. Leifer, *Crude Methods of Search for Functional Dependence*, Statistics. Probability. Economic, Nauka, Moscow, 1985, pp. 345–354. (In Russian)
15. A.P. Sher, *Study of Testing Methods for Diagnostics and Constructing Algorithms for Processing Oceanographic Information on this Basis for Problems of Fishery Forecasting. Autoreferat*, Vladivostok, 1984. (In Russian)
16. A.P. Voshchinin and G.R.Sotirov, *Optimization under Incertainty Conditions*, Moscow Power Engineering Institute, Publishing House Technika (Bulgaria), Moscow–Sofia, 1989.
17. A.P. Voshchinin, A.F. Bochkov and G.R. Sotirov, *A Method to Analyse Data in the Presence of Interval Nonstatistical Error*, Industrial Laboratory 1990 **56** no. 7, 76–81; *Comments by Borodyuk*, 81–83; *Comments by Demidenko*, 83–84; *Comments by Letsky*, 84–86; *Comments by Orlov*, 86–89; *Comments by Legostaeva*, 90–93; *Comments by Kuznetsov*, 93–95. (In Russian)
18. I.L. Legostaeva, *Minimax Estimation of the trend of Stochastic Process. Autoreferat*, Vilnus University, Vilnus, 1986. (In Russian)
19. V.P. Kuznetsov, *Interval Statistical Models*, Radio i Svyaz, Moscow, 1991. (In Russian)
20. A.G. Yakovlev, *Interval Computations – Subject of Research and Useful Tool*, Interval Computations no. 1 (1991), 10–26.

Russian Department
of Intercultural Open University
International Department MPEI
Krasnokazarmennaya 14
Moscow E-250, 105835
Russia