OP&PM SURVEYS
ON APPLIED AND INDUSTRIAL
Volume 25 MATHEMATICS Issue 2
2018

III INTERNATIONAL BALTIC SYMPOSIUM ON APPLIED AND INDUSTRIAL MATHEMATICS

A. K. Melnikov (Moscow, NTC InformInvestGroup, CJSC)**. Application of the calculation method of near-exact statistics probability distributions.**

Problems of computational complexity of the statistics probabilities $P_T\,\{S_n g\text{æ}\}$ exact distributions during processing text flows are considered in the author's work [1]. Creation of statistic criteria for processing texts with the length $n > 50$ within the character set with the power $N > 64$ requires use of exact distributions [2], because use of limit distributions leads to increasing of the number of false selected texts.

In the author's work [3] the parameter space (n, N) of possible calculation and, as a consequence, possible application of exact distributions was calculated on the base of assumption that the performance of the computational resource, available for calculation of exact distributions, is $\Pi_{\text{вс}} = 10^{16}$ operations per second, and the processing time is $T = 30$ days or 2 592 000 seconds. In the same work, a field of limit distributions application was created on the base of the R. A. Fisher's statement [4, p. 73] concerning possibility of use of limit distributions when $k \geqslant 5$, where $k = n/N$. Between the ranges of exact and limit distributions, the parameter space (n, N) was found. For this space it is impossible to calculate exact distributions, and limit distributions cannot be used due to loss of accuracy. As a result, the space was called the space of uncertainty.

Since it is impossible to calculate exact distributions of statistics for samples with the parameters (n, N) from the space of uncertainty, the author in the work [1] suggests to use distributions, which differ from the exact ones not more than by the value Δ, given in advance, or Δ-exact distributions. As an example of calculation of Δ-exact distributions, we selected a point from the space of uncertainty (50, 26). According to the calculation method for Δ-exact distributions, described in [5], Δ-exact statistics distributions with the accuracy $\Delta = 10^{-5}$ were calculated. Their limit distribution is $\chi^2(N-1)$-distribution with N–1 degrees of freedom. There were also calculated Δ-exact distributions of statistic χ_n probabilities

$$\chi_n = \sum_{i=1}^{N} \frac{(h_i - np_i)^2}{np_i},$$

suggested in [6], maximum likelihood statisticsλ

$$\lambda_n = 2\sum_{i=1}^{N} h_i \ln \frac{h_i}{np_i}$$

and Matusita statistic m_n [7]

$$m_n = 4n\sum_{i=1}^{N} \left(\sqrt{\frac{h_i}{n}} - \sqrt{p_i}\right)^2,$$

where h_i is the frequency (sample space) of the symbol a_i, n is the text size (sample size), N is the number of sample spaces of the polynomial scheme (the power of the character set A_n), and p_i is the probability of the a_i sample.

With the chosen accuracy $\Delta = 10^{-5}$ and the calculation method [5], the probability of the statistic with the maximum frequency

$$M_n = \max_{i=1}^{N} h_i$$

was obtained according to recurrent equations (8) – (10) from [5].

The obtained probability of the statistic with the maximum frequency $P\{M_{50} < 12\} = 0,9999992$ defined the limits for calculation of the considered statistics distributions χ_{50}, λ_{50} and m_{50}. The considered statistics distributions are calculated with the chosen accuracy $\Delta = 10^{-5}$ according to the equations:

$$\begin{aligned}
\chi_{50} &= \frac{26}{50}\sum_{\nu=0}^{12}\mu_\nu\left(\nu - \frac{50}{26}\right)^2,\\
\chi_{50} &= 2\sum_{\nu=0}^{12}\nu\,\mu_\nu\,\ln\left(\frac{26}{50}\,i\right)^2,\\
m_{50} &= 200\sum_{\nu=0}^{12}\mu_\nu\left(\sqrt{\frac{\nu}{50}} - \sqrt{\frac{1}{26}}\right)^2,
\end{aligned}$$

where μ_ν is defined as the number of positive integer solutions of the equation $h_1 + \cdots + h_n = n$, for which $h_i = \nu$. Owing to this fact, the calculations became possible and considerably more simple.

Conclusion. Owing to the calculation method of Δ-exact distributions, which differs from the exact ones not more than by the value Δ, given in advance, it is possible to calculate statistics distributions in such parameters space of texts, which has insufficient accuracy of limit distributions, and calculation of exact distributions is not possible. In order to increase effectiveness of statistical procedures of text flows processing it is reasonable to calculate Δ-exact distributions in required parameter spaces, where use of limit distributions is impossible, beforehand.

REFERENCES

1. *Melnikov A. K.*, *Ronzhin A. F.* Generalized statistical method of text analysis, based on calculation of statistics probability distributions. — Informatics Appl., 2016, v. 10, is. 4, p. 89–95. (In Russian.)
2. *Melnikov A. K.* Processing complexity for exact probability distributions of symmetrical additively partitioned statistics and application area of limit distributions. — TUSUR Comm., 2017, v. 20, №4, p. 126–130. (In Russian.)
3. *Zelyukin N. B.*, *Melnikov A. K.* Processing complexity for exact probability distributions of statistics and application area of limit distributions. In: Electronic Facilities and Control Systems: Proceedings of the XIIIth International Scientific and Practical Conference. (Tomsk, November 29th – December 1st, 2017.) Tomsk: V-Spektr, 2017, Part 2, p. 84–90. (In Russian.) https://storage.tusur.ru/files/115115/2017-2.pdf
4. *Fisher R. A.* Statistical Methods for Research Workers. 12 ed. Edinburgh–London: Oliver & Boyd, 1954, 370 p. (Russian translation: Moscow: Gosstatizdat, 1958, 257 p.)
5. *Melnikov A. K.* Methodology of calculation of statistics probability distributions, close to their exact distributions OP&PM Surveys Appl. Industr. Math., 2017, v. 24, is. 5. (In Russian.) http://tvp.ru/conferen/vsppmXVIII/kisso028.pdf
6. *Pearson K.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling. — Philos. Mag. Ser. 5, 1900, v. 50, №302, p. 157–170.
7. *Matusita K.* Decision rules, based on the distance, for problems of fit two samples, and estimation. — Ann. Math. Statist., 1955, v. 26, № 4, p. 631–640.