The accuracy of statistical computing of the standard deviation of a random variable

E V Lebed[1]

[1]Moscow State University of Civil Engineering, Yaroslavskoye Shosse 26, Moscow, 129337, Russia

E-mail: adamtsevichao@mgsu.ru

Abstract. Standard deviation is the main characteristic of statistical investigation of assembly errors of space framed structures. Due to this reason, it is important to have the instrument for the evaluation of the accuracy of the standard deviation of a random variable. The paper establishes the connection between the level of reliability and the accuracy of calculating the standard values. Histograms of the distribution of random numbers subject to the normal law are given for different sample sizes. The dependence of the accuracy of the calculation of the mean-square value on the sample size is shown. Using the tabulated values of chi-square distribution, the dependence is derived between the accuracy of the standard deviation and the sample size for different levels of reliability. These dependencies are presented in the form of diagrams.

1. Introduction

The literature on Mathematical Statistics adequately covers the issue of evaluating the accuracy of statistical parameters such as sample mean and sample variance [1 – 3]. Herewith the method of estimating average statistical is described quite extensively, from both theoretical and practical sides due to the high prevalence of statistical tasks in which the mathematical expectation of a random variable by its numerical value is much greater than its standard deviation [4 – 7].

At the same time, there are tasks where the most important parameter is the standard deviation. These, for example, include the problem of determining the possible errors in the manufacture of parts of mechanical systems and structures in mechanical engineering and construction. The standard deviation here characterizes possible deviation of the actual values of the geometric parameters of parts or structures from the nominal values. For example, the assembly process of space framed structures from single bar elements leads to the accumulation of errors and distortion of their specified geometric configurations. The real dimensions of individual elements are random values and are characterized by the normal law of distribution with mathematical expectation that corresponds to the nominal dimensions of the bars. Therefore, due to the law of large numbers, the errors in the dimensions of individual elements will lead to mathematical expectations of errors of the structure that approach zero. At the same time, the deviation of the actual geometric configuration of a structure from its design configuration will follow the normal distribution law. For statistical analysis of possible errors, it is necessary to determine their standard deviation, which becomes the most important evaluation characteristic. Methods recommended in the literature for variance estimation that can be used for the standard deviation are mostly of analytical nature and are not convenient for practical use [8 – 10]. For this reason, in this work the method for evaluation of the root mean square deviation (standard deviation) is implemented in the graphic form. This approach is suitable for the problems of computer modelling of actual coordinates of nodes of space framed structures (i.e. lattice domes and shells). For these problems, the standard deviation that allows to find the possible location of the node in space is the most important parameter. Modelling of the actual geometric configuration of framed structures is performed with the aid of special software products with the algorithms based on the Monte Carlo method. This algorithm allows modelling the assembly or the installation of space framed structures by assigning every individual element a random deviation of its nominal dimension [11, 12]. A single computer modelling of the actual form of a space framed structure does not provide an objective information because the data obtained is analogous to a single random event. To obtain reliable information about possible errors one must perform a series of computer modellings that correspond to a statistical sampling n that allows computing of a standard deviation S .

In statistical studies, to determine the law of the distribution of unknown random variables, it is common to construct histograms for the density of their distribution by classes in a given interval. Histograms visualize the distribution law of random variables and allow to perform its comparative analysis with a hypothetical or assumed distribution law. For geometric errors in the shape of complex spatial structures, the hypothetical distribution is the normal distribution law. Histograms also enable us to evaluate in a visual form how close the distribution approaches the normal law. In this case, the reliability of the hypothesis can be verified on the basis of the χ^2- distribution or χ^2 criterion. In addition, the appearance of the histograms allows to judge how closely the investigated distribution law corresponds to the hypothetical one. The larger the sample size n of random variables, the closer the match. Figure 1 shows several histograms of the distribution of random variables, obeying the normal law, depending on the number of investigated numbers, i.e. of the sample size n .

The histograms were obtained using a computer program that computes random normal distributed numbers with a mean square deviation of 1 and a mathematical expectation of m = 0. Such programs are called computer-aided random number generators.

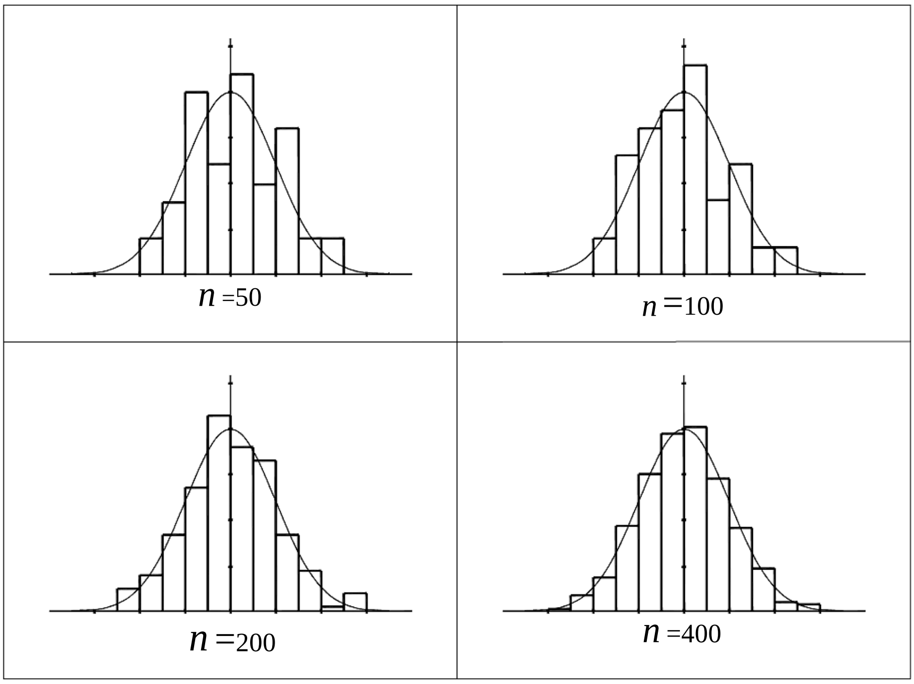

Figure 1. Histograms of the distribution of random numbers, subject to normal law, in dependence of the sample size n .

From these histograms it can be seen that with a small number of random numbers in the sample, the shape of the histogram is not similar to the normal distribution curve (Figure 2). However, when the number of random variables increase, the shape of the histograms begins to gradually approach the shape of the normal distribution curve. This becomes especially obvious if, for example, we compare the histogram with the sample size of 50 numbers with a histogram with a sample size of 400 numbers (Figure 1).

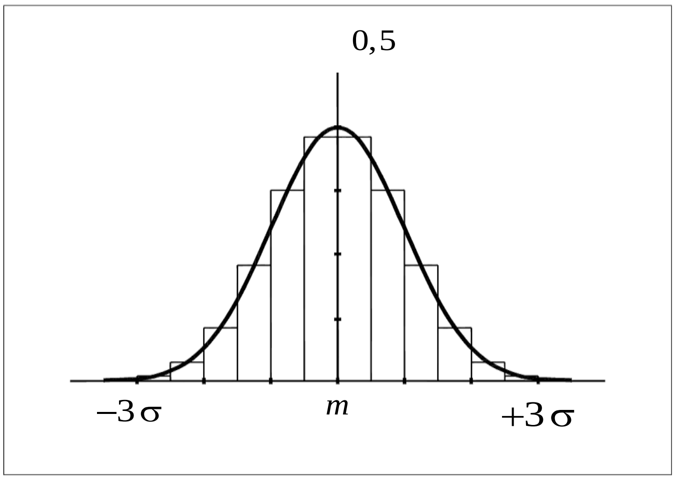

Figure 2: The density curve for the normal distribution of random numbers with σ =1 and m=0.

Therefore, the sample size has a significant effect on the qualitative properties of the diagram, characterized by its general outline. Exactly the same effect does the sample size have on the quantitative characteristic of the distribution - the standard deviation.

2. Methodology

As it is known, any calculation of the standard deviation s according to statistical sample of size n is an approximate estimate of the true value σ, and the difference between them diminishes as the sample size n increases. The dependence between the sample size n and the accuracy of computation of the standard deviation enables us to define beforehand the optimal sample size n needed for computing the standard deviation s, or to estimate the accuracy of the computed value of the standard deviation if the sample size is known. Let us establish the dependence between the sample size n and the accuracy of computation of the standard deviation s characterized by a parameter q for selected reliability levels. It is known from mathematical statistics that the level of reliability γ of the estimate of standard deviation is determined with the given accuracy q by the probability that the true value σ lies in the interval from $s-qs$ to $s+qs$ [2, 9], i.e.

$$\gamma = P\left(s - qs < \sigma < s + qs\right) = 1 - \alpha \qquad (1)$$

where α is the significance level of calculated criterion.

Fabrication errors in details and structures are the consequences of a large number of random factors, therefore, based on the Central Limit Theorem of probability it is possible to assume that their distribution is approaching the normal law.

Based on this assumption, the quantity $(n-1)\, s^2 / \sigma^2$ has the χ^2 – distribution, and the expression (1) can be written in the following form [1, 13]:

$$\gamma = P\left(\frac{\nu s^2}{\chi^2_{1-0,5\alpha}} < \sigma^2 < \frac{\nu s^2}{\chi^2_{0,5\alpha}} \right) = 1 - \alpha \tag{2}$$

where $\nu = n - 1$ is the number of degrees of freedom, and
$\chi^2_{1-0,5\alpha}$ and $\chi^2_{0,5\alpha}$ are the critical values of χ^2,
for which the following formula is true:

$$P_\nu\left(\chi^2_\gamma\right) = \int_0^{\chi^2_\gamma} p\left(\chi^2\right) d\chi^2 = \gamma \ .$$

From the comparison of the expressions (2) and (1) it follows that

$$s(1-q) = \sqrt{\frac{\nu s^2}{\chi^2_{1-0,5\alpha}}}, \qquad s(1+q) = \sqrt{\frac{\nu s^2}{\chi^2_{0,5\alpha}}} \ . \tag{3}$$

Subtracting the first equation of the system (3) from the second one and multiplying both parts by $1/s$ we obtain the formula for computing the accuracy q of the standard deviation s :

$$q = \frac{1}{2}\left(\sqrt{\frac{\nu}{\chi^2_{0,5\alpha}}} - \sqrt{\frac{\nu}{\chi^2_{1-0,5\alpha}}} \right) . \tag{4}$$

Based on equation (4) and using data from the statistical tables for critical values χ^2 [14, 15], the values of the accuracy q of standard deviations S were obtained for different values of degrees of freedom ν. The results of computations for the reliability levels $\gamma = 0,99$ and $\gamma = 0,95$ are presented in the form of diagrams in Figure 3.

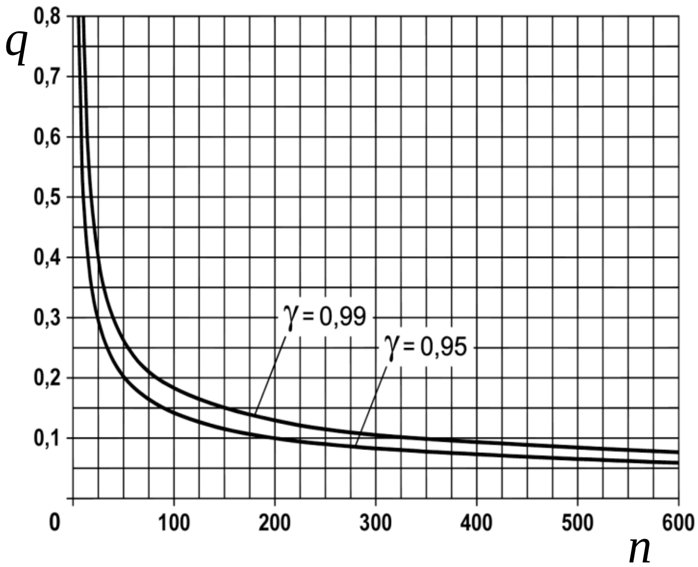

Figure 3. Dependence between the accuracy q and the sample size n for statistical standard deviations s of different levels of reliability γ

The diagrams demonstrate good correlation with particular data available in the literature on statistics [9, 14, 15]. The diagrams show that to obtain the confidence data on the errors the sample size must be not less than 500.

3. Conclusions

The obtained dependences allow for fast and convenient evaluation of the accuracy of the standard deviation of the random variable computed from the experimental results, or for estimation of the sample size required to calculate the standard deviation of a random variable with a given accuracy. It should be noted that the distribution of the desired random variable should approach a normal distribution.

References

[1] Smirnov N V and Dunin-Barkovskiy I V 1965 Course of Theory of Probability and Mathematical Statistics for technical applications (Moscow, Nauka) 512 p
[2] Pugachev V S 1979 Theory of Probability and Mathematical Statistics (Moscow, Nauka) 496 p
[3] Larsen Richard J 2012 An Introduction to Mathematical Statistics and its Applications (Boston, Pearson Education) 757 p
[4] Daren Starnes Dan Yates, David Moore 2012 The Practice of Statistics (New York, W.H. Freeman and Company) 898 p
[5] Chai T and Draxler R R 2014 Root mean square error (RMSE) or mean absolute error (MAE)? *Arguments against avoiding RMSE in the literature* (USA, Geoscientific Model Development) **7** pp 1247–1250
[6] Makkonen Lasse, Pajari Matti and Tikanmäki Maria 2013 Closure to "Problems in the extreme value analysis" (Struct. Safety. 2008:30:405-419) *Structural Safety Elsevier* **40** January Pages 65–67
[7] Meng Xiangrui 2013 Scalable Simple Random Sampling and Stratified Sampling *Proceedings of the 30th International Conference on Machine Learning* (ICML-13) (USA) Pp 531–539
[8] Zaks L 1976 Statistical Evaluation (Moscow, Statistika) 598 p
[9] Solonin I S 1972 Mathematical Statistics for Technology of Machine Building (Moscow, Mashinostroenie) 216 p
[10] Kremer N Sh 2007 Theory of Probability and Mathematical Statistics (Moscow, UNITI-DANA) 551 p
[11] Sobol I M 1973 Numerical Methods of Monte-Carlo (Moscow, Nauka) 312 p
[12] Law Averill M 1991 Simulation Modelling and Analysis (London, McGraw-Hill) 754 p
[13] Yanko Ya 1961 Mathematical-Statistical Tables Translation from Chech (Moscow, Gosstatizdat) 243 p
[14] Ouen D B 1973 Collection of Statistical Tables Translated from Engl (Moscow, Computing Centre of the Academy of Science of the USSR) 586 p
[15] Bol’shev L N and Smirnov N V 1965 Tables of Mathematical Statistics (M, Nauka) 464 p