ISSN 0020-1685, Inorganic Materials, 2026, Vol. 62, No. 1, pp. 141–149. © Pleiades Publishing, Ltd., 2026.
Russian Text © The Author(s), 2025, published in Zavodskaya Laboratoriya, Diagnostika Materialov, 2025, Vol. 91, No. 11, pp. 93–102.

About the Sample Size

A. I. Orlov*

Bauman Moscow State Technical University, Moscow, 105005 Russia

**e-mail: prof-orlov@mail.ru*

Received February 24, 2025; revised March 28, 2025; accepted April 29, 2025

Abstract—It is often required to determine the sample size before engaging into studies. The rules for finding the required sample size are discussed by experts in sufficient detail both in fundamental domain, in engineering, economics and management, medicine, sociology, etc. At the same time sometimes only specific formulations of the problem are addressed lacking the substantiation of required sample size. Without solid ground these estimates can be erroneous. We examine the issue of applying the methods to calculate the required sample size available today and justify the corresponding algorithms from the standpoint of modern applied mathematical statistics. As many as four approaches are considered. The first model has the sample size determined using available resources (material, financial, personnel, etc.) that are sufficient to obtain the sample data (measurements, observations, analyses, experiments, tests, and surveys). The second method explores the issue of estimating the theoretical parameters of a random variable. The required sample size is determined through the specified estimation accuracy which implies the mean half-width of the confidence interval. The third approach is based on the theory of statistical hypothesis testing. Under this we mean selecting from the null and alternative hypotheses based on the level of significance and the strength of test criterion. One-sided alternatives are discussed in detail. The fourth technique is developed within the framework of the interval data statistics. It is based on the principle of equating the statistical and metrology errors. Based on the known accuracy of the initial measurements, we determine the rational sample size, which is considered sufficient under existing conditions. We analyze the rules to calculate the required sample size to estimate probabilities and mathematical expectations as basic examples. Limit theorems of probability theory and mathematical statistic tools are used as a base. It is promising to explore the rate of convergence in the obtained asymptotic expressions and consider the finite sample sizes in further research.

Keywords: applied statistics, sample size, confidence intervals, hypothesis testing, strength of test, statistics of interval data, limit theorems

DOI: 10.1134/S0020168526700664

INTRODUCTION

When conducting scientific and practical work, an integral part of which deals with the statistical analysis of data, quite often the researchers are confronted with the issue: what sample size should be chosen? Numerous approaches have been suggested to handle this issue and we will explore it here in a granular manner.

We discuss four approaches. The first one provides for determining the sample size based on available resources to proceed with measurements, observations, analyses, experiments, tests, and surveys. The second method is based on estimating the theoretical characteristics of a random variable. The required sample size is determined through specified estimation accuracy, which is understood as a half-width of the confidence interval. The third approach is hinged on the theory of statistical hypothesis testing. It boils down to choosing between the null and alternative hypotheses based on the given significance level and the strength of a particular statistical test. The fourth approach was developed within the framework of interval data statistics, which is a part of non-numerical data statistics and systemic fuzzy interval mathematics. A general approach based on the principle of equalizing statistical and metrological errors was considered. The rational sample size, which is necessary from the considered point of view, is determined by the known accuracy of the initial measurements.

The basic examples are given for calculating the required sample size when studying probability values and mathematical expectations. We use the limit theorems of probability theory and mathematical statistics here as a basic mathematical apparatus.

Studies Addressing the Issue of Sample Size

The rules of making above choice are discussed in the works of specialists in various areas of science, engineering and technology [1–4], economics and management [5–8], biology [9], sociology [10], linguistics [11] and others. The impressive scope of material is gathered in medical research [12–14], since collecting and analyzing statistical data makes an inherent part of medical survey. Teams of scientific workers [15–17] and individual researchers (see series of works by G.P. Tikhovoy [18, 19] and M.I. Shpitonkov [20, 21]) specialize in developing recommendations for determining the sample sizes, which is indicative of demand to provide these recommendations. On the other hand, almost all the work of specialists in the area of working out recommendations on estimating the required sample size has certain shortcomings.

It should be mentioned that almost all efforts to specify the required sample size have certain flaws making conclusions obtained less reliable. Electronic and published reports are often incomplete and/or feature inaccuracies and errors. Usually, only individual specific formulations of the problem are considered, and certain recommendations are given on selecting the sample size without proper justification. Moreover, in several cases these conclusions are erroneous. Even the highest quality research fell short on solving the issue of determining the correct sample size. Thus, in [4], only the approach based on a half-width of the confidence interval is analyzed, whereas [22] analyzes the issue from the standpoint of hypothesis testing theory.

At the same time, it should be admitted that mathematicians have overlooked this issue although it is very important for practice. No approach has been developed or explored, no method has been proposed to estimate the sample size.

We provide exact formulation of the methods to calculate the required sample size and justify the corresponding algorithms from the standpoint of modern applied mathematical statistics. The mathematical apparatus is based on the limit theorems of probability theory and on mathematical statistics. We will consider the four approaches mentioned in the introduction section.

Required Sample Size under Limited Resources

From the point of view of classical mathematical statistics, the answer is quite clear: the larger the sample size, the better. However, in practice, we have to account for the funds required to collect the data, because as the sample size increases, the rate of growing accuracy of the estimate is coming down. Thus, when the sample size grows as a factor of k, the standard deviation from the estimated parameter decreases only by $1/\sqrt{k}$. The sample size has to be chosen based on a compromise between the accuracy and the cost of the research. For example, in sociology, the number of respondents polled is usually no higher than a few thousand (usually 500 – 1000) [23].

Required Sample Size in Probability Estimates

As already mentioned, the required sample size can be found by specifying the accuracy of statistical conclusions. According to [23, 24], one should start by choosing a probabilistic-statistical model. Let the analyzed data $X_1, X_2, \ldots, X_n$ be considered a sample, i.e., a set of independent (in total) identically distributed random variables. Let the sample elements take only two values (0 and 1, for example, 0 when the component made is good and 1 when it is faulty). Let's denote the probabilities $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. We should estimate the fraction p of responses 1 in the general total (the share of defective components). For an unknown parameter p the confidence interval (p_n, p_v) can be derived based on the sample size n and the number of responses 1 in the sample, equal to $X_1 + X_2 + \ldots + X_n = X$, as follows [23]:

$$p_n = p^* - C(\gamma)\sqrt{\frac{p^*(1-p^*)}{n}},\quad p_\text{v} = p^* + C(\gamma)\sqrt{\frac{p^*(1-p^*)}{n}}, \tag{1}$$

where p_n is the lower confidence limit; p_v is the upper confidence limit; $p^* = X/n$ is the share (the number of responses 1 in the sample); γ is the confidence probability;

$$C(\gamma) = \Phi^{-1}((1+\gamma)/2)$$

is the coefficient corresponding to the confidence probability ($\Phi^{-1}(x)$ is the inverse of the standard normal distribution function (with a mean of 0 and a variance of 1)). Typically, a confidence probability of $\gamma = 0.95$ is used, so $C(\gamma) = C(0.95) = 1.96$.

Note. The formulas (1) are derived in [23] based on Moivre–Laplace theorem under the assumption of an infinite growth in the sample size. When solving practical problems, it is usually assumed that this condition is met when a sample size $n \geq 10$.

Let the required accuracy of estimating the probability p be given by Δ. This means that the confidence interval should be completely contained within the interval $(p^* - \Delta; p^* - \Delta)$. According to the formulas (1), this means that the following inequality should be satisfied

$$C(\gamma)\sqrt{\frac{p^*(1-p^*)}{n}} \leq \Delta. \tag{2}$$

We obtain from formula (2)

$$n \geq \frac{p^*(1-p^*)}{\Delta^2}C^2(\gamma) \tag{3}$$

(since the sample size is a natural number, the right-hand side of the inequality in formula (3) should be increased to the nearest integer).

The formula (3) cannot be directly applied to determine the sample size because $p^* = X/n$, i.e., we should know the sample size to apply this formula. There are two ways to overcome this difficulty. Firstly, the value of p^* can be estimated from the results of previous studies or from the analysis of a small preliminary sample. Secondly, one can use the fact that $p^*(1 - p^*) \leq 1/4$, with equality achieved when $p^* = 1/2$. Therefore, instead of (3), we can use the formula

$$n \geq \frac{1}{4\Delta^2} C^2(\gamma). \quad (4)$$

We already mentioned that the value of the confidence probability most frequently used is $\gamma = 0.95$ we have $C(\gamma) = 1.96$, and this value can be replaced with 2 with sufficient accuracy for practice, we obtain formula (4) in the form

$$n \geq 1/\Delta^2. \quad (5)$$

According to formula (5), for estimating the probability with an accuracy of $\pm 10\%$ (i.e., $\Delta = 0.1$) the required sample size equals 100 and estimating with an accuracy of $\pm 5\%$ (i.e., $\Delta = 0.05$) already requires as many as 400 observations.

It is clear that formula (4) gives overrated volumes for probabilities other than 1/2. For example, if the probability estimate is about 0.1 (or 0.9), then $p^*(1 - p^*)$ is not 0.25, but 0.09, therefore, the required sample size according to formula (3) is reduced by 0.25/0.09 = 2.78 times compared to the right-hand side of formula (4).

Sample Size to Estimate the Mathematical Expectation

The model predicts that the sample elements X_1, X_2, ..., X_n are independent identically distributed random variables. Since the variables have equal distribution their mathematical expectations coincide, as do their variances: $M(X_i) = a$, $D(X_i) = \sigma^2$, $i = 1, 2, ..., n$. The point estimate of the mathematical expectation is the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

and the confidence interval (a_n, a_v) for the expected value a can be expressed as follows [23]:

$$a_n = \bar{X} - C(\gamma)\frac{s}{\sqrt{n}}, \quad a_v = \bar{X} + C(\gamma)\frac{s}{\sqrt{n}},$$

where

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

is the sample standard deviation (i.e., the statistical estimate of the theoretical standard deviation σ).

The analog of formula (2) is the ratio

$$C(\gamma)\frac{s}{\sqrt{n}} \leq \Delta.$$

Therefore, the required sample size is the minimum n such that

$$n \geq \frac{s^2}{\Delta^2} C^2(\gamma). \quad (6)$$

Similar to the probability estimation, we are faced with a problem since the value of the sample variance s^2 should be known before making measurements. This value can be estimated using previous studies or from the analysis of a small size preliminary sample.

The technical specifications of the measuring instrument usually specify the value of random error as its standard deviation σ. In this case, instead of (5), we can use the formula

$$n \geq \frac{\sigma^2}{\Delta^2} C^2(\gamma).$$

As with probability estimation, the confidence probability $\gamma = 0.95$ is usually used when $C(\gamma) = 1.96$. Occasionally the theoretical accuracy is indicated as $a \pm \sigma$, which corresponds to the sample estimate $\bar{X} \pm s$. In this case, $C(\gamma) = 1.00$, i.e. $\gamma = 0.68$.

We considered two estimation tasks. Many other estimation tasks are used in applied statistics. In all these tasks, the required sample size is found by equating the half-width of the symmetric confidence interval to the specified accuracy.

The Required Sample Size to Test a Statistical Probability Hypothesis

The construction of a probabilistic-statistical model [23, 24] begins with formulating two hypotheses—the null H_0 and the alternative H_1. Let's consider two typical formulations.

If random variables take two values, then the probabilistic-statistical model is described by two parameters: by the sample size n and by the probability of a certain value p.

Task 1. Let hypotheses be given

$$H_0\colon p = p_0, \quad H_1\colon p = p_1. \quad (7)$$

Consider a random variable X that has a binomial distribution with parameters n (sample size) and p (probability of having a certain value). Based on X and n, a decision should be made about which of the two hypotheses is correct. For example, in statistical quality control ([25], Ch. 10), based on the sample size and the number of defective units in it, it is necessary to determine the share of defective units in the batch p_0 or $p = p_1$.

Task 2. The second formulation deals with the mathematical expectation $M(X)$ of a random variable X. Hypotheses are given:

$$H_0: M(X) = a_0, \quad H_1: M(X) = a_1. \tag{8}$$

Based on the sample $X_1, X_2, \ldots, X_n$ a decision should be made as to which of the two hypotheses is correct. For example, in the statistical control of technological processes using the Shewhart control chart method, the hypothesis H_0 corresponds to the correct process state, and the hypothesis H_1 corresponds to disorganized process ([25], Ch. 10).

Statistical criteria [23] are used to test the statistical hypotheses, i.e., functions of statistical data (specifically: the sample size and the number of defective units—in task 1 and $X_1, X_2, \ldots, X_n$—in task 2). Each statistical test has two types of errors: a type I error (the probability that the null hypothesis is rejected when it is true) and a type II error (the probability that the null hypothesis is not rejected when the alternative hypothesis is true). The probability of a type I error is usually denoted by α, and the probability of a type II error is denoted by β. The value of α is called the significance level of the statistical test, and the value of $(1 - \beta)$ is called the strength of the test. The significance level is set by the researcher, and is usually set at $\alpha = 0.05$. The required sample size is found by setting the desired strength of the test.

It is recommended [23] to apply the statistics criterion to test the null hypothesis (8) in task 1 that is based on statistics

$$Y = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}.$$

As follows from the Moivre–Laplace theorem, if the null hypothesis is true, then for growing sample size the distribution of this statistics converges to a standard normal distribution with mathematical expectation 0 and variance 1:

$$\lim_{n\to\infty} P(Y < x) = \Phi(x). \tag{9}$$

The rule for making a decision based on the value of Y depends on the assumptions about the alternative hypotheses.

If it is not known in which direction the deviation from the null hypothesis can be expected (i.e., if it is not known which value is larger p_0 or p_1 a two-sided alternative is considered), then the rule to take decision is as follows [23]: if $|Y| \le k(\alpha)$, then the null hypothesis is accepted, if $|Y| > k(\alpha)$, then the null hypothesis is rejected and the alternative is accepted. Here, the coefficient $k(\alpha)$ depends on the significance level α. It is determined from the condition $P(|Y| > k(\alpha)) = \alpha$. If the null hypothesis is true, then in the case of $|Y| > k(\alpha)$, we make a type I error. From (9) it follows that

$$P(|Y| > k(\alpha)) = P(Y < -k(\alpha)) + P(Y > k(\alpha))$$
$$= \Phi(-k(\alpha)) + 1 - \Phi(k(\alpha)) = 2 - 2\Phi(k(\alpha))$$

(for a sufficiently large sample size). Therefore,

$$\alpha = 2 - 2\Phi(k(\alpha)), \quad \Phi(k(\alpha) = 1 - \alpha/2,$$
$$k(\alpha) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

If $\alpha = 0.05$, then $k(\alpha) = 1.96$.

Yet, in practice, the problem of hypothesis testing (7) is often considered when the probabilities p_0 and p_1 are known. Let $p_1 > p_0$ for the definiteness, i.e. consider a one-sided alternative. Then it is natural to use a different decision rule: if $Y \le m(\alpha)$, the null hypothesis is accepted, if $Y > m(\alpha)$, then the null hypothesis is rejected and the alternative is accepted. Here $m(\alpha)$ is a coefficient that depends on the significance level α. It is determined from the condition $P(Y > m(\alpha)) = \alpha$. If the null hypothesis is true, then in the case of $Y > m(\alpha)$ we make a type I error. From (9) it follows that $P(Y > m(\alpha)) = 1 - \Phi(m(\alpha))$ (for a sufficiently large sample size). So,

$$\alpha = 1 - \Phi(m(\alpha)), \quad m(\alpha) = \Phi^{-1}(1 - \alpha).$$

If $\alpha = 0.05$, then $m(\alpha) = 1.64$.

Let us proceed to study the strength of the test. It is equal to the probability that the null hypothesis is rejected under the alternative hypothesis being true. The identity holds

$$Y = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$
$$= \frac{X - np_1 + (np_1 - np_0)}{\sqrt{np_1(1 - p_1)}} \frac{\sqrt{np_1(1 - p_1)}}{\sqrt{np_0(1 - p_0)}},$$

from which it follows

$$Y = \frac{\sqrt{p_1(1 - p_1)}}{\sqrt{p_0(1 - p_0)}} \frac{X - np_1}{\sqrt{np_1(1 - p_1)}} + \sqrt{n}\frac{p_1 - p_0}{\sqrt{p_0(1 - p_0)}}.$$

Consider the case of a one-sided alternative $p_1 > p_0$. Let us introduce the notation:

$$A = A(p_0, p_1) = \frac{\sqrt{p_1(1 - p_1)}}{\sqrt{p_0(1 - p_0)}} > 0,$$
$$B = B(p_0, p_1) = \frac{p_1 - p_0}{\sqrt{p_0(1 - p_0)}} > 0.$$

Then

$$Y = A(p_0, p_1)\frac{X - np_1}{\sqrt{np_1(1 - p_1)}}$$
$$+ B(p_0, p_1)\sqrt{n} = A\frac{X - np_1}{\sqrt{np_1(1 - p_1)}} + B\sqrt{n}.$$

Since the random variable Y has a binomial distribution with parameter $p = p_1$. Then, as the sample size grows, the distribution of the centered and normalized random variable

$$\frac{X - np_1}{\sqrt{np_1(1 - p_1)}}.$$

Approaches the distribution of a random variable Z having a standard normal distribution with mean 0 and variance 1. The distribution of Y converges to the distribution of $W = AZ + B\sqrt{n}$, which allows us to calculate the strength of the statistical test in question. Since $P(Y > m(\alpha)) = P(AZ + B\sqrt{n} > m(\alpha))$, the (asymptotic) strength

$$P\left(Z > \frac{m(\alpha) - B\sqrt{n}}{A}\right) = 1 - \Phi\left(\frac{m(\alpha) - B\sqrt{n}}{A}\right) = 1 - \beta.$$

Consequently,

$$\frac{m(\alpha) - B\sqrt{n}}{A} = \Phi^{-1}(\beta),$$

$$n = \left(\frac{m(\alpha) - A\Phi^{-1}(\beta)}{B}\right)^2.$$

The reasoning for the two-tailed alternative hypothesis is similar.

Let's consider an example. Let $p_0 = 0.1$ and $p_1 = 0.5$. The required sample size

$$n = \left(\frac{1.64 + 1.67 \times 2.33}{1.33}\right)^2 = 19.14.$$

Since the sample size is a natural number, rounding up 19.14 to the nearest natural number, we get that the required sample size is 20.

Sample Size Required to Test a Statistical Hypothesis about the Mathematical Expectation

In the problem statement above, 2 hypotheses H_0 are given: $M(X) = a_0$. H_1: $M(X) = a_1$. Based on the sample $X_1, X_2, \ldots, X_n$ a decision should be made as to which of the two hypotheses is correct. The probabilistic-statistical model is usually adopted [23, 24], according to which, if the null hypothesis H_0 is true, sample elements have a distribution function $f(x)$ with variance σ_0^2. And if the alternative hypothesis H_1 is true the distribution function is different $G(x)$ and it has variance σ_1^2. If, however, the same measuring instrument is used in both cases to determine the value of the controlled parameter, then it is assumed that $\sigma_0 = \sigma_1$ and the distribution functions $f(x)$ and $G(x)$ differ only by a shift. For mathematical rigor, we assume that the first three central moments exist, and therefore the Central Limit Theorem (CLT) of probability theory holds for both hypotheses.

The test statistics for testing the null hypothesis according to [23] is given by

$$Y = \sqrt{n}\frac{\overline{X} - \alpha_0}{s},$$

where s is the sample standard deviation. According to the CLT, the distribution of the statistics Y converges to a standard normal distribution with mean 0 and variance 1 as the sample size grows. As is often the case in practice [25], let it be known that a_0 is less than a_1, i.e. the deviation can only grow. In this case, the decision rule corresponding to significance level α is as follows: if $Y \leq q(\alpha)$, then the null hypothesis is accepted, if $Y > q(\alpha)$, then the null hypothesis is rejected and the alternative is accepted. Here, under the null hypothesis, $P(Y > q(\alpha)) = 1 - P(Y \leq q(\alpha)) = \alpha$. According to the Central Limit Theorem 1—$\Phi(q(\alpha)) = \alpha$, $\Phi(q(\alpha)) = 1 - \alpha$, $q(\alpha) = \Phi^{-1}(1 - \alpha)$. For the most common value of the significance level $\alpha = 0.05$, we have $q(0.05) = \Phi^{-1}(0.95) = 1.64$.

The equality holds

$$Y = \sqrt{n}\frac{\overline{X} - \alpha_1 + (\alpha_1 - \alpha_0)}{s} = \sqrt{n}\frac{\overline{X} - \alpha_1}{s} + \sqrt{n}\frac{\alpha_1 - \alpha_0}{s}. \quad (10)$$

The distribution of the first term in (10) converges to the distribution of a standard normal random variable W. For the second, the limiting relation holds

$$\lim_{n \to +\infty} \frac{\alpha_1 - \alpha_0}{s} = \frac{\alpha_1 - \alpha_0}{\sigma_1} = B. \quad (11)$$

Therefore, the distribution of Y converges to the distribution of a random variable $W = Z + B\sqrt{n}$.

According to (10) and (11), we find the strength of the test. If the alternative hypothesis is true and the sample size is large enough

$$P(Y > q(\alpha)) = P\left(Z + B\sqrt{n} > q(\alpha)\right) = P\left(Z > q(\alpha) - B\sqrt{n}\right) = 1 - \Phi\left(q(\alpha) - B\sqrt{n}\right).$$

The required sample size for a given test strength is found from the equation $1 - \Phi(q(\alpha) - B\sqrt{n}) = 1 - \beta$. It follows that

$$\Phi\left(q(\alpha) - B\sqrt{n}\right) = \beta, \quad q(\alpha) - B\sqrt{n} = \Phi^{-1}(\beta),$$

$$n = \left(\frac{q(\alpha) - \Phi^{-1}(\beta)}{B}\right)^2.$$

If we assume the significance level to be 0.05 and the probability of the second kind error to be 0.01, then $\Phi^{-1}(0.01) = -2.33$ and

$$n = \left(\frac{1.96 + 2.33}{B}\right)^2 = \frac{18.4}{B^2} = \frac{18.4\sigma_1^2}{(a_1 - a_0)^2}. \quad (12)$$

Let's consider a numerical example. Let the difference in mathematical expectations $a_1 - a_0$ be equal to 1.0 and the standard deviation under the validity of the alternative hypothesis σ_1 be equal to 1.5. Then $B = 1.0/1.5 = 0.67$ and using formula (12) we find $n = 41.4$. Since the sample size is a natural number, then, rounding up 41.4 to the nearest natural number, we get that the required sample size is 42.

The reasoning for the two-tailed alternative hypothesis is similar.

Estimating Sample Size Using Interval Data Statistics

A number of methods for calculating the required sample size based on the width of the confidence interval, significance level, and test strength have been presented above. The choice of specific values for these characteristics is made when formulating the problem and is, therefore, largely subjective. The method based on interval data statistics, incorporating metrological considerations proves to be interesting and useful.

If the accuracy of the measurements is known, interval data statistics can answer the question about the required sample size ([23], Ch. 12). It is based on the equal error principle, according to which the statistical error should be equal to the metrological one. The statistics of interval data corresponds to the new paradigm of applied statistics [26].

In various problems of mathematical statistics, we usually start from the basic concept of a sample $X_1, X_2, \ldots, X_n$ and consider the values of certain functions $f(X_1, X_2, \ldots, X_n)$ of the sample elements. In the interval data statistics, the inevitable measurement errors are explicitly taken into account. These errors require processing not the elements of the sample X_i, but distorted values Y_i, $i = 1, 2, \ldots, n$, and, therefore, statistical conclusions are made not on the basis of $f(X_1, X_2, \ldots, X_n)$, but on the basis of the distorted $f(Y_1, Y_2, \ldots, Y_n)$. Here $Y_i = X_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ is the error vector.

The case is important when $|\varepsilon_i| \le \Delta$, $i = 1, 2, \ldots, n$ where the notation $X_i \pm \Delta$ applies. It means that the unknown true value of the sample element belongs to the (closed) interval $[X_i - \Delta, X_i + \Delta]$. The sample elements are not numbers X_i, $i = 1, 2, \ldots, n$, but intervals $[a_i, b_i]$, $i = 1, 2, \ldots, n$, where $a_i = X_i - \Delta$, $b_i = X_i + \Delta$. The methods to process these interval data are addressed in the statistics of interval data ([23], Ch. 12)—a relatively new area of modern applied mathematical statistics (it has been developed since the 1980s, while classical mathematical statistics was formed by the middle of the 20th century).

For the study of the properties of statistical data analysis methods (specifically for interval data statistics), the basic concept of notna (the maximum possible deviation) is of great importance.

$$N_f(X_1, X_2, \ldots, X_n) = \sup_{\{\varepsilon\}} |f(X_1, X_2, \ldots, X_n) - f(Y_1, Y_2, \ldots, Y_n)|,$$

here the supremum is taken over the set of all possible errors $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$.

If all errors satisfy the constraint $|\varepsilon_i| \le \Delta$ and the maximum possible error Δ is small, then for a sufficiently smooth function $f(X_1, X_2, \ldots, X_n)$ (it is sufficient for this function to have second continuous partial derivatives) up to higher-order infinitesimals, the equality holds true

$$N_f(X_1, X_2, \ldots, X_n) = \left(\sum_{i=1}^{n} \left|\frac{\partial f(X_1, X_2, \ldots, X_n)}{\partial X_i}\right|\right)\Delta. \quad (13)$$

In terms of metrology $f(X_1, X_2, \ldots, X_n)$ this is an indirect measurement based on direct measurements of X_i, $i = 1, 2, \ldots, n$. The error of indirect measurement is equal to $N_f(X_1, X_2, \ldots, X_n)$, in other words, the result of indirect measurement has the form $f(X_1, X_2, \ldots, X_n) \pm N_f(X_1, X_2, \ldots, X_n)$.

The formula (13) allows to find the error of indirect measurement. When solving data analysis problems, it can usually be proven that there is a limit

$$\lim_{n \to +\infty} \left(\sum_{i=1}^{n} \left|\frac{\partial f(X_1, X_2, \ldots, X_n)}{\partial X_i}\right|\right) = C,$$

where C is some positive constant, its own for each problem under consideration. Then, for sufficiently large sample sizes n, the notna is equal to $C\Delta$. Therefore, the error in calculating the statistics $f(X_1, X_2, \ldots, X_n)$ does not tend to 0. It follows that there are no valid estimates (in terms of classical mathematical statistics) existing in interval data statistics.

When the common statistical methods are used, the considered statistic $f(X_1, X_2, \ldots, X_n)$ is asymptotically normal in the majority of cases, i.e. there are constants a and σ such that

$$\lim_{n \to +\infty} P\left(\sqrt{n}\frac{f(X_1, X_2, \ldots, X_n) - a}{\sigma} < x\right) = \Phi(x),$$

where $\Phi(x)$ is the standard normal distribution function with mean 0 and variance 1. Usually, the limiting relations are satisfied

$$\lim_{n\to+\infty} \sqrt{n}(Mf(X_1, X_2, \ldots, X_n) - a) = 0,$$
$$\lim_{n\to+\infty} nDf(X_1, X_2, \ldots, X_n) = \sigma^2.$$

According to them, in classical mathematical statistics, the average square error of the considered statistical estimate is

$$\begin{aligned} &M(f(X_1, X_2, \ldots, X_n) - a)^2 \\ &= (Mf(X_1, X_2, \ldots, X_n) - a)^2 \\ &+ Df(X_1, X_2, \ldots, X_n) = \sigma^2/n, \end{aligned} \tag{14}$$

to an infinitesimal of a higher order.

In the statistics of interval data (which takes into account the presence of observation errors), instead of formula (14) for the square of the error, we usually obtain the expression

$$\begin{aligned} &\sup_{\{\varepsilon\}} M(f(X_1, X_2, \ldots, X_n) - a)^2 \\ &= \sigma^2/n + N_f^2(X_1, X_2, \ldots, X_n) + o(\Delta^2 + 1/n), \end{aligned} \tag{15}$$

where the supremum is taken over the set of all possible errors $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$. The first term on the right-hand side of (15) corresponds to the statistical error, and the second—to the metrological one.

According to (15), as the sample size increases, the average square of the error does not tend to 0 (as in classical mathematical statistics), but always remains greater than some positive constant (namely, the square of the notna). Therefore, there is no point in increasing the sample size without limit. According to the theory of stability of economic-mathematical methods and models [27], we use the principle of error equalization here, according to which it is advisable to equalize the values of different errors. Applied to (15), the principle of error equalization leads to the relationship

$$\sigma^2/n = N_f^2(X_1, X_2, \ldots, X_n). \tag{16}$$

Formula (16) allows finding the required sample size:

$$n = \frac{\sigma^2}{N_f^2(X_1, X_2, \ldots, X_n)}. \tag{17}$$

For practical application of formula (17), it is necessary to replace the theoretical values of the variance σ^2 in considered statistics $f(X_1, X_2, \ldots, X_n)$ and the square of its notna N_f^2 $(X_1, X_2, \ldots, X_n)$ using sample data estimates. The calculation algorithms are developed for solving various problems of applied statistics: for estimating the characteristics and parameters of distributions, for hypothesis testing, linear regression analysis, discriminant analysis, cluster analysis, as well as for estimating the errors of the characteristics of financial flows of investment projects [23]. As for the statistics of interval data, formula (17) defines the rational sample size. This term is used in the same sense as the term necessary sample size used in this paper.

As an example, let's calculate the required sample size for estimating the sample $X_1, X_2, \ldots, X_n$ for mathematical expectation of its elements using the sample arithmetic mean. In this case

$$f(X_1, X_2, \ldots, X_n) = \bar{X} = \frac{X_1, X_2, \ldots, X_n}{n}.$$

Since

$$\frac{\partial f(X_1, X_2, \ldots, X_n)}{\partial X_i} = \frac{1}{n},$$

then according to formula (13) the notation of the sample mean is

$$\begin{aligned} &N_f(X_1, X_2, \ldots, X_n) \\ &= \left(\sum_{i=1}^{n} \left|\frac{\partial f(X_1, X_2, \ldots, X_n)}{\partial X_i}\right|\right)\Delta = \left(\sum_{i=1}^{n} \frac{1}{n}\right)\Delta = \Delta. \end{aligned}$$

Following the Central Limit Theorem of probability theory

$$\lim_{n\to+\infty} P\left(\sqrt{n}\frac{\bar{X} - a}{\sigma} < x\right) = \Phi(x),$$
$$a = M(X_i), \quad \sigma = \sqrt{D(X_i)}, \quad i = 1, 2, \ldots, n.$$

Therefore, the required sample size is

$$n = \frac{\sigma^2}{N_f^2(X_1, X_2, \ldots, X_n)} = \sigma^2/\Delta^2. \tag{18}$$

For example, if $\sigma = 1$, $\Delta = 1/6$ (i.e. the standard deviation of the measurement result is six times the measurement precision), then the required sample size is 36. If the measurement precision improves, i.e. Δ decreases, then according to (18) the required sample size increases. And, conversely, when using a measuring instrument with low precision (i.e., with a large Δ), it is sufficient to use a small sample size. If the theoretical variance σ^2 used in formula (18) is unknown, then it should be estimated either from previous studies of the same phenomenon or process, or using the sample variance found from a small sample size.

CONCLUSIONS

The research dealing with gathering and analyzing the statistical data quite often involves the issue of what sample size should be used. We developed rules for calculating the required sample size using an example of studying the values of probabilities and mathematical expectations of a random variable. For each of these statistical tasks, two approaches were proposed within the framework of classical mathematical statistics. The first one considers tasks of statistical estimation with a given accuracy, which is understood as a half-width of the confidence interval. The second deals with the verification of statistical hypotheses and with choosing between the null and alternative hypotheses based on the given significance and the strength of the statistical criterion.

If the accuracy of the measurements is known, then the statistics of interval data allows to answer the question about the required sample size. A general approach based on the principle of equalizing the indicators of statistical and metrological errors was addressed. The mathematical expectation estimation was analyzed as an example of applying this approach.

In further research, it is advisable to study the convergence rate of obtained asymptotic expressions, in other words, to consider finite sample volumes. The method of statistical tests (proposed in [28]) can be instrumental in estimating the errors of the limit theorems obtained.

FUNDING

The research was funded from the by a grant from the University budget.

CONFLICT OF INTEREST

The author of this work declares that he has no conflicts of interest.

ADDITIONAL INFORMATION

No artificial intelligence tools were used in the work on the article.

REFERENCES

1. Aleskandrovskaya, L.N. and Solonnikov, Yu.I., Planning the size of statistical tests, *Tr. NPTsAP, Sist. Prib. Upravl.,* 2009, no. 4, pp. 33–36.
2. Feigenbaum, Yu.M., The minimum required size of flight information processing to assess typical operating conditions and load of an aircraft instance, *Nauch. Vestn. Mosk. Tekh. Univ. Grazhd. Aviats.,* 2012, no. 175, pp. 25–29.
3. Agamirov, L.V., Agamirov, V.L., and Vestyak, V.A., Algorithms for planning fatigue tests of aircraft materials, *Progr. Prod. Sist.,* 2014, no. 4, pp. 205–210.
4. Shleenko, A.V., Volkova, S.N., and Sivak, E.E., Optimization of sampling for setting up a scientific experiment of the construction process, *Byull. Stroit. Tekh.,* 2018, no. 11 (1011), pp. 46–48.
5. Larina, T.N., Justification of the required sample size when studying the quality of life of the rural population of Russia, *Internet Zh. Naukoved.,* 2015, vol. 7, no. 1 (26), p. 21.
6. Popov, A.M., Stability of the method for testing the hypothesis about the equality of the means of two samples, *Nauka Biznes: Puti Razvit.,* 2017, no. 6 (72), pp. 22–24.
7. Prokopchina, S.V., Determining the required size of sample data to create a scale with dynamic constraints, *Ekon. Upravl. Probl. Resh.,* 2018, vol. 4, no. 6, pp. 109–114.
8. Levine, D.M., Stephan, D., and Berenson, M.L., *Statistics for Managers Using Microsoft Excel,* 4th ed., Englewood Cliffs, NJ: Prentice-Hall, 1998.
9. Makarova, M.N., Shekunova, E.V., Rybakova, A.V., and Makarov, V.G., Sample size of laboratory animals for experimental studies, *Farmatsiya,* 2018, vol. 67, no. 2, pp. 3–8.
10. Koshevoi, O.S. and Karpova, M.K., Determining the size of the sample population when conducting regional sociological research, *Izv. Vyssh. Uchebn. Zaved., Povolzh. Reg., Obshchestv. Nauki,* 2011, no. 2, pp. 98–104.
11. Savchenko, V.V., Determining the size of the control sample under conditions of a priori uncertainty based on the principle of guaranteed results, *Nauch. Vedom. Belgor. Univ., Ser. Ekon. Inform.,* 2015, no. 1 (198), pp. 74–78.
12. Motrenko, A.P., Estimation of the required sample size of patients when predicting cardiovascular diseases, *Mash. Obuch. Anal. Dann.,* 2012, vol. 1, no. 3, pp. 354–366.
13. Rebrova, O.Yu. and Gusev, A.V., Sample size calculation for clinical trials of clinical decision support systems with binary response, *Sovrem. Tekhnol. Med.,* 2022, vol. 14, no. 3, pp. 6–14.
14. Narkevich, A.N. and Vinogradov, K.A., Methods for determining the minimum required sample size in medical research, *Sots. Aspekty Zdor. Naselen.,* 2019, vol. 65, no. 6, p. 10.
15. Grzhibovskii, A.M., Gorbatova, M.A., Narkevich, A.N., and Vinogradov, K.A., Sample size for correlation analysis, *Morsk. Med.,* 2020, vol. 6, no. 1, pp. 101–106.
16. Grzhibovskii, A.M., Gorbatova, M.A., Narkevich, A.N., and Vinogradov, K.A., Required sample size to compare arithmetic means in two independent groups, *Morsk. Med.,* 2020, vol. 6, no. 2, pp. 106–113.
17. Grzhibovskii, A.M., Gorbatova, M.A., Narkevich, A.N., and Vinogradov, K.A., Required sample size to compare proportions in two independent groups, *Morsk. Med.,* 2020, vol. 6, no. 3, pp. 76–83.
18. Tikhova, G.P., We are planning a clinical trial. Question No. 1: How to determine the required sample size?, *Reg. Anest. Lech. Ostr. Boli,* 2014, vol. 8, no. 3, pp. 57–63.
19. Tikhova, G.P., We are planning a clinical trial. Question No. 1: how to determine the required sample size?, *Zdravookhranenie,* 2016, no. 9, pp. 47–53.

20. Shpitonkov, M.I., Sample size modeling in clinical trials for binary data, *Issled. Oper. (Mod., Sist., Reshen.),* 2018, no. 4 (13), pp. 28–35.

21. Shpitonkov, M.I., *Modeling sample size in non-inferiority clinical trials, Issled. Oper. (Mod., Sist., Reshen.),* 2020, vol. 6, pp. 39–45.

22. Vetrov, L.G., Kuznecova, A.A., and Sunchalina, A.L., *Prikladnaya statistika: metodicheskie ukazaniya k vypolneniyu laboratornykh rabot* (Applied Statistics: Guidelines for Performing Laboratory Works), Moscow: MGTU im. N.E. Baumana, 2017.

23. Orlov, A.I., *Prikladnoi statisticheskii analiz* (Applied Statistical Analysis), Moscow: IPR MEDIA, 2022.

24. Orlov, A.I., On the requirements for statistical methods of data analysis (generalising article), *Zavod. Lab. Diagn. Mater.,* 2023, vol. 89, no. 11, pp. 98–106. https://doi.org/10.26896/1028-6861-2023-89-11-98-106

25. Orlov, A.I., *Iskusstvennyi intellekt: statisticheskie metody analiza dannykh* (Artificial Intelligence: Statistical Methods for Data Analysis), Moscow: IPR MEDIA, 2022.

26. Orlov, A.I., The new paradigm for applied statistics, *Zavod. Lab. Diagn. Mater.,* 2012, vol. 78, no. 1, part 1, pp. 87–93.

27. Orlov, A.I., *Ustoichivye ekonomiko-matematicheskie metody i modeli* (Stable Economic and Mathematical Methods and Models), Moscow: IPR MEDIA, 2022.

28. Orlov, A.I., Statistical simulations method in applied statistics, *Zavod. Lab. Diagn. Mater.,* 2019, vol. 85, no. 5, pp. 67–79. https://doi.org/10.26896/1028-6861-2019-85-5-67-79

Translated by V. Vetrov