

Математические методы исследования

Mathematical methods of investigation

DOI: <https://doi.org/10.26896/1028-6861-2025-91-11-93-102>

ОБ ОБЪЕМЕ ВЫБОРКИ

© Александр Иванович Орлов

Московский государственный технический университет им. Н. Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., д. 5; e-mail: prof-orlov@mail.ru

*Статья поступила 24 февраля 2025 г. Поступила после доработки 28 марта 2025 г.
Принята к публикации 29 апреля 2025 г.*

При планировании статистических исследований часто возникает задача определения необходимого объема выборки. Правила нахождения необходимого объема выборки достаточно широко обсуждаются в работах специалистов фундаментальной и прикладной науки, техники, экономики и управления, медицины, социологии и др. Однако обычно рассматривают лишь частные постановки задачи, приводят без обоснования те или иные рецепты расчета необходимого объема выборки, причем зачастую ошибочные. В данной работе строго сформулированы методы расчета необходимого объема выборки, разработанные к настоящему времени, а также обоснованы соответствующие алгоритмы с позиций современной прикладной математической статистики. Рассмотрены четыре подхода. В первом из них объем выборки определяется имеющимися ресурсами (материальными, финансовыми, кадровыми и др.), которые можно использовать для получения элементов выборки (проведения измерений, наблюдений, анализов, опытов, испытаний, обследований). Во втором подходе рассматриваются задачи оценивания теоретических характеристик случайной величины. Необходимый объем выборки определяют по заданной точности оценивания, под которой понимают полуширину доверительного интервала. Третий подход основан на теории проверки статистических гипотез. Речь идет о выборе между нулевой и альтернативной гипотезами исходя из уровня значимости и мощности критерия. Подробно рассмотрены односторонние альтернативы. Четвертый подход разработан в рамках статистики интервальных данных. Он базируется на принципе уравнивания статистических и метрологических погрешностей. По известной точности исходных измерений определяют рациональный объем выборки, который и является с рассматриваемой точки зрения необходимым. Как базовые примеры рассмотрены правила расчета необходимого объема выборки при изучении значений вероятностей и математических ожиданий. В качестве основы математического аппарата использованы предельные теоремы теории вероятностей и математической статистики. В дальнейших исследованиях целесообразно изучить скорость сходимости в полученных асимптотических выражениях, рассмотреть конечные объемы выборок.

Ключевые слова: прикладная статистика; объем выборки; доверительные интервалы; проверка гипотез; мощность критерия; статистика интервальных данных; предельные теоремы.

ABOUT THE SAMPLE SIZE

© Alexander I. Orlov

Bauman Moscow State Technical University, 5, 2-ya Baumanskaya ul., Moscow, 105005, Russia; e-mail: prof-orlov@mail.ru

Received February 24, 2025. Revised March 28, 2025. Accepted April 29, 2025.

When planning statistical studies, it is often necessary to determine the required sample size. The rules for finding the required sample size are quite widely discussed in the works of specialists in various fields of science and practice — engineering and technology, economics and management, medicine, sociology, etc. However, usually only individual specific statements of the problem are considered, and certain recipes for calculating the required quantity are given without justification. sample size, and often erroneous. The article is devoted to the strict formulation of methods for calculating the required sample size, developed to date, as well as the justification of the corresponding algorithms from the standpoint of modern applied mathematical statistics. Four approaches are considered. In the first of them, the sample size is de-

terminated by the available resources (material, financial, personnel, etc.) that can be used to obtain sample elements (measurements, observations, analyses, experiments, tests, surveys). In the second, problems of estimating the theoretical characteristics of a random variable are considered. The required sample size is determined by the specified estimation accuracy, by which we mean the half-width of the confidence interval. The third approach is based on the theory of statistical hypothesis testing. We are talking about choosing between the null and alternative hypotheses, based on the level of significance and power of the criterion. One-sided alternatives are discussed in detail. The fourth is developed within the framework of interval data statistics. It developed a general approach based on the principle of equalizing statistical and metrological errors. Based on the known accuracy of the initial measurements, we determine the rational sample size, which is necessary from the point of view under consideration. As basic examples, the rules for calculating the required sample size when studying probability values and mathematical expectations are considered. Limit theorems of probability theory and mathematical statistics are used as the basis of the mathematical apparatus. In further research, it is advisable to study the rate of convergence in the obtained asymptotic expressions and consider finite sample sizes.

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

Введение

При проведении научных и практических работ, составной частью которых является статистический анализ данных, часто возникает вопрос: какой объем выборки выбрать? Для ответа на этот вопрос разработан ряд подходов, рассмотрению которых посвящена данная статья.

Обсудим четыре подхода. В первом из них объем выборки определяют на основе имеющихся ресурсов для проведения измерений, наблюдений, анализов, опытов, испытаний, обследований. При втором подходе рассматривают задачи оценивания теоретических характеристик случайной величины. Необходимый объем выборки определяют по заданной точности оценивания, под которой понимают полуширину доверительного интервала. Третий подход основан на теории проверки статистических гипотез. Речь идет о выборе между нулевой и альтернативной гипотезами исходя из заданных уровня значимости и мощности того или иного статистического критерия. Четвертый подход разработан в рамках статистики интервальных данных, являющейся частью статистики нечисловых данных и системной нечеткой интервальной математики. Рассмотрен общий подход на основе принципа уравнивания статистических и метрологических погрешностей. По известной точности исходных измерений определяют рациональный объем выборки, который и является с рассматриваемой точки зрения необходимым.

В качестве базовых примеров приведены правила расчета необходимого объема выборки при изучении значений вероятностей и математических ожиданий. В качестве основы математического аппарата в данной работе используем предельные теоремы теории вероятностей и математической статистики.

Многообразие работ, посвященных выбору объема выборки

Правила такого выбора обсуждаются в работах специалистов различных направлений

науки, техники и технологий [1 – 4], экономики и управления [5 – 8], биологии [9], социологии [10], лингвистики [11] и др. Особенно много работ касается научных медицинских исследований [12 – 14], поскольку типовой раздел в них — сбор и анализ статистических данных. Коллективы научных работников [15 – 17] и отдельные исследователи (см. серии работ Г. П. Тиховой [18, 19] и М. И. Шпитонкова [20, 21]) специализируются на разработке рекомендаций по определению объемов выборок, что отражает значительную востребованность подобных рекомендаций.

Однако практически все работы специалистов отмеченных областей по выбору необходимого объема выборки имеют те или иные недостатки, снижающие обоснованность данных в них рекомендаций. Затрагивающие эту тему электронные и бумажные источники часто неполны, в них встречаются неточности и ошибки. Обычно рассматривают лишь отдельные частные постановки задачи, приводят без обоснования те или иные рецепты выбора объема выборки, причем зачастую ошибочные. Даже в наиболее квалифицированных работах рассмотрены лишь отдельные подходы к решению задачи определения необходимого объема выборки. Так, в [4] проанализирован только подход, исходящий из полуширины доверительного интервала, а в [22] даны рецепты в рамках теории проверки гипотез.

Одновременно приходится констатировать, что математики не разбирались глубоко в рассматриваемой проблеме, весьма важной для прикладников, не разрабатывали и не изучали подходы и методы определения объема выборки.

В данной работе строго сформулированы методы расчета необходимого объема выборки и обоснованы соответствующие алгоритмы с позиций современной прикладной математической статистики. Математический аппарат — предельные теоремы теории вероятностей и матема-

тической статистики. Рассмотрим последовательно четыре подхода, отмеченные во введении.

Необходимый объем выборки при ограниченных ресурсах

С точки зрения классической математической статистики ответ ясен: чем больше объем выборки, тем лучше. Однако на практике приходится учитывать размеры средств, необходимых для сбора данных, поскольку при увеличении объема выборки точность оценивания улучшается всё меньше. Так, при увеличении объема выборки в k раз среднее квадратическое отклонение от оцениваемого параметра уменьшается лишь как $1/\sqrt{k}$. Объем выборки приходится выбирать на основе компромисса между точностью и стоимостью исследования. Например, в социологии число опрошенных, как правило, не превышает несколько тысяч (обычно 500 – 1000) [23].

Необходимый объем выборки при оценивании вероятности

Как уже отмечалось, необходимый объем выборки можно найти, задав точность статистических выводов. Согласно [23, 24], начать надо с выбора вероятностно-статистической модели. Пусть анализируемые данные $X_1, X_2, \dots, X_n$ рассматриваются как выборка, т.е. набор независимых (в совокупности) одинаково распределенных случайных величин. Пусть элементы выборки принимают только два значения — 0 и 1 (например, 0, когда деталь годная, а 1 — когда она дефектная). Обозначим вероятности $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. Необходимо оценить долю p ответов 1 в генеральной совокупности (долю дефектных единиц). Доверительный интервал (p_n, p_v) для неизвестного параметра p строится по объему выборки n и числу ответов 1 в выборке, равному $X_1 + X_2 + \dots + X_n = X$, следующим образом [23]:

$$\begin{aligned} p_n &= p^* - C(\gamma) \sqrt{\frac{p^*(1-p^*)}{n}}, \\ p_v &= p^* + C(\gamma) \sqrt{\frac{p^*(1-p^*)}{n}}, \end{aligned} \quad (1)$$

где p_n — нижняя доверительная граница; p_v — верхняя доверительная граница; $p^* = X/n$ — выборочная доля (доля ответов 1 в выборке); γ — доверительная вероятность;

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

коэффициент, соответствующий доверительной вероятности ($\Phi^{-1}(x)$ — функция, обратная к функции стандартного нормального распреде-

ления с математическим ожиданием 0 и дисперсией 1). Обычно используют доверительную вероятность $\gamma = 0,95$, тогда $C(\gamma) = C(0,95) = 1,96$.

Замечание. Формулы (1) получены в [23] на основе теоремы Муавра – Лапласа в предположении безграничного роста объема выборки. При решении практических задач обычно полагают, что их можно использовать при объеме выборки $n \geq 10$.

Пусть задана необходимая точность оценивания Δ вероятности p . Это значит, что доверительный интервал должен полностью входить в интервал $(p^* - \Delta; p^* + \Delta)$. Согласно формулам (1) это означает, что должно быть выполнено неравенство

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

Из формулы (2) следует, что

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

(поскольку объем выборки — натуральное число, то правую часть неравенства в формуле (3) необходимо увеличить до ближайшего целого).

Формулу (3) нельзя непосредственно применить для определения объема выборки, поскольку $p^* = X/n$, т.е. знание объема выборки необходимо для применения этой формулы. Есть два пути для преодоления данной сложности. Во-первых, величина p^* может быть оценена по результатам предыдущих исследований или анализа данных предварительной выборки небольшого объема. Во-вторых, можно воспользоваться тем, что $p^*(1-p^*) \leq 1/4$, причем равенство достигается при $p^* = 1/2$. Следовательно, вместо (3) можно использовать формулу

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

Поскольку, как уже сказано, для наиболее часто применяемого значения доверительной вероятности $\gamma = 0,95$ имеем $C(\gamma) = 1,96$, а это значение с достаточной для практики точностью можно заменить на 2, то получим формулу (4) в виде

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

Согласно формуле (5), для оценки вероятности с точностью $\pm 10\%$ (т.е. $\Delta = 0,1$) необходимый объем выборки равен 100, а для оценки с точностью $\pm 5\%$ (т.е. $\Delta = 0,05$) требуется уже 400 наблюдений.

Ясно, что формула (4) дает завышенные объемы для вероятностей, отличающихся от 1/2. Например, если оценка вероятности около 0,1 (или 0,9), то $p^*(1 - p^*)$ равно не 0,25, а 0,09, следовательно, необходимый объем выборки по формуле (3) сокращается в $0,25/0,09 = 2,78$ раза по сравнению с правой частью формулы (4).

Необходимый объем выборки при оценивании математического ожидания

Согласно принятой модели, элементы выборки $X_1, X_2, \dots, X_n$ — независимые одинаково распределенные случайные величины. Поскольку они одинаково распределены, то их математические ожидания совпадают, как и дисперсии: $M(X_i) = a$, $D(X_i) = \sigma^2$, $i = 1, 2, \dots, n$. Точечной оценкой математического ожидания является выборочное среднее арифметическое

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

а доверительный интервал (a_n, a_b) для математического ожидания a имеет следующий вид [23]:

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

где

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

выборочное среднее квадратическое отклонение (т.е. статистическая оценка теоретического среднего квадратичного отклонения σ).

Аналогом формулы (2) является соотношение

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

Следовательно, необходимый объем выборки — это минимальное n , такое, что

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

Как и в случае оценивания вероятности, возникает проблема из-за того, что до проведения измерений неизвестна величина выборочной дисперсии s^2 . Эта величина может быть оценена по результатам предыдущих исследований или анализа данных предварительной выборки небольшого объема.

В техническом паспорте средства измерения обычно указывают такую характеристику случайной ошибки, как ее среднее квадратическое

отклонение σ . В таком случае вместо (5) можно применить формулу

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

Как и при оценивании вероятности, обычно используют доверительную вероятность $\gamma = 0,95$, тогда $C(\gamma) = 1,96$. Однако иногда теоретическую точность указывают в виде $a \pm \sigma$, что соответствует выборочной оценке $\bar{X} \pm s$. В этом случае $C(\gamma) = 1,00$, т.е. $\gamma = 0,68$.

Выше рассмотрены две задачи оценивания. В прикладной статистике используют и многие другие задачи оценивания. Во всех из них необходимый объем выборки находят путем приравнивания полуширины симметричного доверительного интервала к заданной точности.

Необходимый объем выборки при проверке статистической гипотезы о вероятности

Построение вероятностно-статистической модели [23, 24] начинается с задания двух гипотез — нулевой H_0 и альтернативной H_1 . Рассмотрим две типовые постановки.

Если случайные величины принимают два значения, то вероятностно-статистическая модель описывается двумя параметрами — объемом выборки n и вероятностью определенного значения p .

Задача 1. Пусть заданы гипотезы

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

Рассматривают случайную величину X , имеющую биномиальное распределение с параметрами n (объем выборки) и p (вероятность определенного значения). Исходя из X и n , необходимо принять решение о том, какая из двух гипотез верна. Например, при статистическом контроле качества [25, гл. 10] по объему выборки и числу дефектных единиц в ней надо установить, какова доля дефектных единиц в партии — p_0 или $p = p_1$.

Задача 2. Во второй постановке речь идет о математическом ожидании $M(X)$ случайной величины X . Задают гипотезы

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

По выборке $X_1, X_2, \dots, X_n$ необходимо принять решение о том, какая из двух гипотез верна. Например, при статистическом регулировании технологических процессов методом контрольных карт Шухарта гипотеза H_0 соответствует налаженному состоянию процесса, а гипотеза H_1 — разлаженному [25, гл. 10].

Для проверки статистических гипотез используют статистические критерии [23], т.е. функции от статистических данных (а именно: от объема выборки и числа дефектных единиц — в задаче 1 и от $X_1, X_2, \dots, X_n$ — в задаче 2). С каждым статистическим критерием связаны две ошибки — ошибка первого рода (вероятность того, что нулевая гипотеза отклоняется, хотя она верна) и ошибка второго рода (вероятность того, что нулевая гипотеза принимается, хотя верна альтернативная). Вероятность ошибки первого рода обычно обозначают α , вероятность ошибки второго рода — β . Величину α называют уровнем значимости статистического критерия, а величину $(1 - \beta)$ — мощностью критерия. Уровень значимости задает исследователь, обычно принимают $\alpha = 0,05$. Необходимый объем выборки находят, задав требуемую мощность критерия.

В задаче 1 для проверки нулевой гипотезы (8) рекомендуется [23] статистический критерий, основанный на статистике

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

Как следует из теоремы Муавра – Лапласа, если верна нулевая гипотеза, то при росте объема выборки распределение этой статистики сходится к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1:

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

Правило принятия решения на основе значения Y зависит от предположений об альтернативных гипотезах.

Если неизвестно, в какую сторону может быть отклонение от нулевой гипотезы (т.е. если неизвестно, какое значение больше — p_0 или p_1 , рассматривают двухстороннюю альтернативу), то правило принятия решения имеет следующий вид [23]: если $|Y| \leq k(\alpha)$, то принимают нулевую гипотезу, если же $|Y| > k(\alpha)$, то отклоняют нулевую гипотезу и принимают альтернативную. Здесь коэффициент $k(\alpha)$ зависит от уровня значимости α . Его определяют из условия $P(|Y| > k(\alpha)) = \alpha$. Если верна нулевая гипотеза, то в случае $|Y| > k(\alpha)$ совершаем ошибку первого рода. Из (9) следует, что

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

$$= \Phi(-k(\alpha)) + 1 - \Phi(k(\alpha)) = 2 - 2\Phi(k(\alpha))$$

(при достаточно большом объеме выборки). Следовательно,

$$\alpha = 2 - 2\Phi(k(\alpha)), \quad \Phi(k(\alpha)) = 1 - \alpha/2,$$

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

Если $\alpha = 0,05$, то $k(\alpha) = 1,96$.

Однако на практике задачу проверки гипотез (7) достаточно часто рассматривают, когда вероятности p_0 и p_1 полностью известны. Пусть для определенности $p_1 > p_0$, т.е. рассматривают одностороннюю альтернативу. Тогда естественно использовать другое правило принятия решения: если $Y \leq m(\alpha)$, то принимают нулевую гипотезу, если же $Y > m(\alpha)$, то отклоняют нулевую гипотезу и принимают альтернативную. Здесь $m(\alpha)$ — коэффициент, зависящий от уровня значимости α . Его определяют из условия $P(Y > m(\alpha)) = \alpha$. Если верна нулевая гипотеза, то в случае $Y > m(\alpha)$ совершаем ошибку первого рода. Из (9) следует, что $P(Y > m(\alpha)) = 1 - \Phi(m(\alpha))$ (при достаточно большом объеме выборки). Итак,

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

Если $\alpha = 0,05$, то $m(\alpha) = 1,64$.

Перейдем к изучению мощности критерия. Она равна вероятности того, что при справедливости альтернативной гипотезы нулевая гипотеза отклоняется. Справедливо тождество

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

из которого следует

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

Рассмотрим случай односторонней альтернативы $p_1 > p_0$. Введем обозначения:

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

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

Тогда

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

Поскольку случайная величина Y имеет биномиальное распределение с параметром $p = p_1$, то распределение централизованной и нормированной случайной величины

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

приближается при росте объема выборки к распределению случайной величины Z , имеющей стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, распределение Y сближается с распределением $W = AZ + B\sqrt{n}$, что позволяет рассчитать мощность рассматриваемого статистического критерия. Поскольку $P(Y > m(\alpha)) = P(AZ + B\sqrt{n} > m(\alpha))$, то (асимптотическая) мощность

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

Следовательно,

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

Рассуждения для двухсторонней альтернативной гипотезы аналогичны.

Рассмотрим пример. Пусть $p_0 = 0,1$ и $p_1 = 0,5$. Необходимый объем выборки

$$n = \left(\frac{1,64 + 1,67 \cdot 2,33}{1,33}\right)^2 = 19,14.$$

Поскольку объем выборки — натуральное число, то, округляя 19,14 до ближайшего натурального числа сверху, получаем, что необходимый объем выборки — 20.

Необходимый объем выборки при проверке статистической гипотезы о математическом ожидании

В сформулированной выше задаче 2 заданы гипотезы $H_0: M(X) = a_0$, $H_1: M(X) = a_1$. По выборке $X_1, X_2, \dots, X_n$ необходимо принять решение о том, какая из двух гипотез верна. Обычно принимают вероятностно-статистическую модель [23, 24], согласно которой при справедливости нулевой гипотезы H_0 элементы выборки имеют функцию распределения $F(x)$ с дисперсией σ_0^2 , а при справедливости альтернативной гипотезы H_1 — некоторую другую функцию распределения $G(x)$ с дисперсией σ_1^2 . Если же для определения значения контролируемого параметра в обоих случаях используют одно и то же средство измерения, то полагают, что $\sigma_0 = \sigma_1$, а функции

распределения $F(x)$ и $G(x)$ отличаются только сдвигом. Для математической строгости предполагаем, что существуют первые три центральных момента, а потому для обеих гипотез справедлива Центральная предельная теорема теории вероятностей (ЦПТ).

Статистика критерия для проверки нулевой гипотезы согласно [23] имеет вид

$$Y = \sqrt{n} \frac{\bar{X} - a_0}{s},$$

где s — выборочное среднее квадратическое отклонение. Согласно ЦПТ распределение статистики Y при росте объема выборки сходится к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Как часто бывает на практике [25], пусть известно, что a_0 меньше, чем a_1 , т.е. отклонение может быть только в большую сторону. В этом случае соответствующее уровню значимости α решающее правило таково: если $Y \leq q(\alpha)$, то принимают нулевую гипотезу, если же $Y > q(\alpha)$, то отклоняют нулевую гипотезу и принимают альтернативную. Здесь при справедливости нулевой гипотезы $P(Y > q(\alpha)) = 1 - P(Y \leq q(\alpha)) = \alpha$. В соответствии с Центральной предельной теоремой $1 - \Phi(q(\alpha)) = \alpha$, $\Phi(q(\alpha)) = 1 - \alpha$, $q(\alpha) = \Phi^{-1}(1 - \alpha)$. Для наиболее распространенного значения уровня значимости $\alpha = 0,05$ имеем $q(0,05) = \Phi^{-1}(0,95) = 1,64$.

Справедливо равенство

$$\begin{aligned} Y &= \sqrt{n} \frac{\bar{X} - a_1 + (a_1 - a_0)}{s} = \\ &= \sqrt{n} \frac{\bar{X} - a_1}{s} + \sqrt{n} \frac{a_1 - a_0}{s}. \end{aligned} \quad (10)$$

Распределение первого слагаемого в (10) сходится к распределению стандартной нормальной случайной величины W . Для второго справедливо предельное соотношение

$$\lim_{n \rightarrow +\infty} \frac{a_1 - a_0}{s} = \frac{a_1 - a_0}{\sigma_1} = B. \quad (11)$$

Следовательно, распределение Y сближается с распределением случайной величины $W = Z + B\sqrt{n}$.

В соответствии с (10) и (11) найдем мощность критерия. При справедливости альтернативной гипотезы и достаточно большом объеме выборки

$$\begin{aligned} P(Y > q(\alpha)) &= P(Z + B\sqrt{n} > q(\alpha)) = \\ &= P(Z > q(\alpha) - B\sqrt{n}) = 1 - \Phi(q(\alpha) - B\sqrt{n}). \end{aligned}$$

Необходимый объем выборки по заданной мощности критерия находят из уравнения $1 - \Phi(q(\alpha) - B\sqrt{n}) = 1 - \beta$. Из него следует, что

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

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

Если принять уровень значимости равным 0,05, а вероятность ошибки второго рода равной 0,01, то $\Phi^{-1}(0,01) = -2,33$ и

$$n = \left(\frac{1,96 + 2,33}{B} \right)^2 = \frac{18,4}{B^2} = \frac{18,4\sigma_1^2}{(a_1 - a_0)^2}. \quad (12)$$

Рассмотрим численный пример. Пусть разность математических ожиданий $a_1 - a_0$ равна 1,0, а среднее квадратическое отклонение при справедливости альтернативной гипотезы σ_1 равно 1,5. Тогда $B = 1,0/1,5 = 0,67$ и по формуле (12) находим $n = 41,4$. Поскольку объем выборки — натуральное число, то, округляя 41,4 до ближайшего натурального числа сверху, получаем, что необходимый объем выборки — 42.

Рассуждения для двухсторонней альтернативной гипотезы аналогичны.

Необходимый объем выборки на основе статистики интервальных данных

Выше приведен ряд методов расчета необходимого объема выборки на основе задания ширины доверительного интервала, уровня значимости, мощности критерия. Выбор конкретных значений этих характеристик проводится при постановке задачи и потому во многом субъективен. Интересен и полезен метод на основе статистики интервальных данных с привлечением соображений метрологии.

Если известна точность проводимых измерений, то ответить на вопрос о необходимом объеме выборки позволяет статистика интервальных данных [23, гл. 12]. Она исходит из принципа уравнивания погрешностей, согласно которому статистическая погрешность должна равняться метрологической. Статистика интервальных данных соответствует новой парадигме прикладной статистики [26].

В различных задачах математической статистики обычно исходят из базового понятия выборки $X_1, X_2, \dots, X_n$ и рассматривают значения тех или иных функций $f(X_1, X_2, \dots, X_n)$ от элементов выборки. В статистике интервальных данных в явной форме учитываются неизбежные погрешности измерений. Их наличие приводит к тому, что обрабатывать приходится не элементы выборки X_i , а их искаженные значения Y_i , $i = 1,$

2, ..., n , а потому статистические выводы делаются не на основе $f(X_1, X_2, \dots, X_n)$, а на основе искаженного значения $f(Y_1, Y_2, \dots, Y_n)$. Здесь $Y_i = X_i + \varepsilon_i$, $i = 1, 2, \dots, n$, где $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ — вектор погрешностей.

Важен частный случай, когда $|\varepsilon_i| \leq \Delta$, $i = 1, 2, \dots, n$. Тогда используют запись $X_i \pm \Delta$. Она означает, что неизвестное нам реальное значение элемента выборки лежит в (замкнутом) интервале $[X_i - \Delta, X_i + \Delta]$. Элементами выборки являются не числа X_i , $i = 1, 2, \dots, n$, а интервалы $[a_i, b_i]$, $i = 1, 2, \dots, n$, где $a_i = X_i - \Delta$, $b_i = X_i + \Delta$. Методам обработки подобных интервальных данных посвящена статистика интервальных данных [23, гл. 12] — сравнительно новая область современной прикладной математической статистики (развивается с 1980-х годов, в то время как классическая математическая статистика сформировалась к середине XX в.).

Для изучения свойств методов анализа статистических данных большое значение имеет базовое для статистики интервальных данных понятие нотны

$$N_f(X_1, X_2, \dots, X_n) =$$

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

где супремум берется по множеству всех возможных погрешностей $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$.

Если все погрешности удовлетворяют ограничению $|\varepsilon_i| \leq \Delta$ и максимально возможная погрешность Δ мала, то для достаточно гладкой функции $f(X_1, X_2, \dots, X_n)$ (достаточно наличие у этой функции вторых непрерывных частных производных) с точностью до бесконечно малых более высокого порядка справедливо равенство

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

В терминах метрологии $f(X_1, X_2, \dots, X_n)$ — это косвенное измерение на основе прямых измерений X_i , $i = 1, 2, \dots, n$. Погрешность косвенного измерения равна $N_f(X_1, X_2, \dots, X_n)$, другими словами, результат косвенного измерения имеет вид $f(X_1, X_2, \dots, X_n) \pm N_f(X_1, X_2, \dots, X_n)$.

Формула (13) позволяет находить погрешность косвенного измерения. При решении задач анализа данных обычно можно доказать, что существует предел

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

где C — некоторая положительная константа, своя для каждой рассматриваемой задачи. Тогда

при достаточно больших объемах выборки n нотна равна $СД$. Следовательно, погрешность вычисления статистики $f(X_1, X_2, \dots, X_n)$ не стремится к 0. Отсюда следует, что в статистике интервальных данных не существует состоятельных оценок (в терминах классической математической статистики).

При использовании распространенных статистических методов рассматриваемая статистика $f(X_1, X_2, \dots, X_n)$ в большинстве случаев является асимптотически нормальной, т.е. существуют константы a и σ такие, что

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

где $\Phi(x)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Обычно выполнены предельные соотношения

$$\begin{aligned} \lim_{n \rightarrow +\infty} \sqrt{n}(Mf(X_1, X_2, \dots, X_n) - a) &= 0, \\ \lim_{n \rightarrow +\infty} nDf(X_1, X_2, \dots, X_n) &= \sigma^2. \end{aligned}$$

В соответствии с ними в классической математической статистике средний квадрат ошибки рассматриваемой статистической оценки составляет

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

с точностью до бесконечно малых более высокого порядка.

В статистике интервальных данных, в которой учитывается наличие погрешностей наблюдений, вместо формулы (14) для квадрата ошибки обычно получаем выражение

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

где супремум берется по множеству всех возможных погрешностей $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$. Первое слагаемое в правой части (15) соответствует статистической погрешности, второе — метрологической.

Согласно (15) при росте объема выборки средний квадрат ошибки не стремится к 0, как в классической математической статистике, но всегда остается больше некоторой положительной константы (а именно — квадрата нотны). Следовательно, нерационально безгранично увеличивать объем выборки. В соответствии с теорией устойчивости экономико-математических методов и моделей [27] используем здесь принцип уравнивания погрешностей, согласно которому целесообразно уравнивать величины погрешностей различной природы. Применительно к (15) принцип

уравнивания погрешностей приводит к соотношению

$$\sigma^2/n = N_f^2(X_1, X_2, \dots, X_n). \quad (16)$$

Формула (16) позволяет найти необходимый объем выборки:

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

Для практического применения формулы (17) надо заменить теоретические значения дисперсии σ^2 рассматриваемой статистики $f(X_1, X_2, \dots, X_n)$ и квадрата ее нотны $N_f^2(X_1, X_2, \dots, X_n)$ на их оценки по выборочным данным. Алгоритмы расчетов разработаны для решения различных задач прикладной статистики: для оценивания характеристик и параметров распределений, проверки гипотез, линейного регрессионного анализа, дискриминантного анализа, кластер-анализа, а также для оценки погрешностей характеристик финансовых потоков инвестиционных проектов [23]. В статистике интервальных данных формула (17) задает «рациональный объем выборки». Этот термин используется в том же смысле, что и термин «необходимый объем выборки» в настоящей статье.

В качестве примера рассчитаем необходимый объем выборки для оценивания по выборке $X_1, X_2, \dots, X_n$ математического ожидания ее элементов с помощью выборочного среднего арифметического. В этом случае

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

Поскольку

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

то согласно формуле (13) нотна выборочного среднего арифметического составит

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

Согласно Центральной предельной теореме теории вероятностей

$$\lim_{n \rightarrow +\infty} P\left(\sqrt{n} \frac{\bar{X} - a}{\sigma} < x\right) = \Phi(x),$$

$$a = M(X_i), \quad \sigma = \sqrt{D(X_i)}, \quad i = 1, 2, \dots, n.$$

Следовательно, необходимый объем выборки

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

Например, если $\sigma = 1$, $\Delta = 1/6$ (т.е. среднее квадратическое отклонение результата измерения в шесть раз превосходит точность измерения), то необходимый объем выборки равен 36. Если точность измерения повышается, т.е. Δ уменьшается, то согласно (18) необходимый объем выборки увеличивается. И, наоборот, при применении средства измерения с малой точностью (т.е. с большим Δ) достаточно использовать небольшой объем выборки. Если используемая в формуле (18) теоретическая дисперсия σ^2 неизвестна, то ее следует оценить либо по результатам предыдущих исследований того же явления или процесса, либо с помощью выборочной дисперсии, найденной по результатам анализа пробной выборки небольшого объема.

Заключение

При планировании исследований, связанных со сбором и анализом статистических данных, часто возникает вопрос о том, какой объем выборки следует использовать. В настоящей статье получены правила расчета необходимого объема выборки на примере изучения значений вероятностей и математических ожиданий случайной величины. Для каждой из этих статистических задач в рамках классической математической статистики предложены два подхода. В первом из них рассматриваются задачи статистического оценивания с заданной точностью, под которой понимается полуширина доверительного интервала. Во втором речь идет о проверке статистических гипотез, о выборе между нулевой и альтернативной гипотезами исходя из заданных значимости и мощности статистического критерия.

Если известна точность проводимых измерений, то ответить на вопрос о необходимом объеме выборки позволяет статистика интервальных данных. Рассмотрен общий подход на основе принципа уравнивания показателей статистических и метрологических погрешностей. В качестве примера применения этого подхода проанализировано оценивание математического ожидания.

В дальнейших исследованиях целесообразно изучить скорость сходимости в полученных асимптотических выражениях, другими словами, рассмотреть конечные объемы выборок. Для оценки погрешностей полученных в статье предельных теорем может быть применен метод статистических испытаний согласно [28].

Дополнительные сведения

При работе со статьей не использованы инструменты искусственного интеллекта.

ЛИТЕРАТУРА

1. Алесандровская Л. Н., Солонников Ю. И. Планирование объема статистических испытаний / Труды ФГУП «НПЦАП». Системы и приборы управления. 2009. № 4. С. 33 – 36.
2. Фейгенбаум Ю. М. Минимально необходимый объем обработки полетной информации для оценки типовых условий эксплуатации и нагруженности экземпляра самолета / Научный вестник Московского государственного технического университета гражданской авиации. 2012. № 175. С. 25 – 29.
3. Агамиров Л. В., Агамиров В. Л., Вестяк В. А. Алгоритмы планирования усталостных испытаний авиационных материалов / Программные продукты и системы. 2014. № 4. С. 205 – 210.
4. Шлеенко А. В., Волкова С. Н., Сивак Е. Е. Оптимизация выборки для постановки научного эксперимента технологического процесса строительства / БСТ: Бюллетень строительной техники. 2018. № 11(1011). С. 46 – 48.
5. Ларина Т. Н. Обоснование необходимого объема выборки при исследовании качества жизни сельского населения России / Интернет-журнал «Науковедение». 2015. Т. 7. № 1(26). С. 21.
6. Попов А. М. Устойчивость метода проверки гипотезы о равенстве средних двух выборок / Наука и бизнес: пути развития. 2017. № 6(72). С. 22 – 24.
7. Прокопчина С. В. Определение необходимого объема выборочных данных для создания шкалы с динамическими ограничениями / Экономика и управление: проблемы, решения. 2018. Т. 4. № 6. С. 109 – 114.
8. Левин Д. М., Стефан Д., Кребиль Т. С., Беренсон М. Л. Статистика для менеджеров с использованием Microsoft Excel. Изд. 4-е. — М.: ИД «Вильямс», 2004. — 1312 с.
9. Макарова М. Н., Шекунова Е. В., Рыбакова А. В., Макаров В. Г. Объем выборки лабораторных животных для экспериментальных исследований / Фармация. 2018. Т. 67. № 2. С. 3 – 8.
10. Кошевой О. С., Карпова М. К. Определение объема выборочной совокупности при проведении региональных социологических исследований / Известия вузов. Поволжский регион. Общественные науки. 2011. № 2. С. 98 – 104.
11. Савченко В. В. Определение объема контрольной выборки в условиях априорной неопределенности по принципу гарантированного результата / Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 2015. № 1(198). С. 74 – 78.
12. Мотренко А. П. Оценка необходимого объема выборки пациентов при прогнозировании сердечно-сосудистых заболеваний / Машинное обучение и анализ данных. 2012. Т. 1. № 3. С. 354 – 366.
13. Реброва О. Ю., Гусев А. В. Расчет объема выборки для клинических испытаний систем поддержки принятия врачебных решений с бинарным откликом / Современные технологии в медицине. 2022. Т. 14. № 3. С. 6 – 14.
14. Наркевич А. Н., Виноградов К. А. Методы определения минимально необходимого объема выборки в медицинских исследованиях / Социальные аспекты здоровья населения. 2019. Т. 65. № 6. С. 10.
15. Гржибовский А. М., Горбатова М. А., Наркевич А. Н., Виноградов К. А. Объем выборки для корреляционного анализа / Морская медицина. 2020. Т. 6. № 1. С. 101 – 106.
16. Гржибовский А. М., Горбатова М. А., Наркевич А. Н., Виноградов К. А. Необходимый объем выборки для сравнения средних арифметических в двух независимых группах / Морская медицина. 2020. Т. 6. № 2. С. 106 – 113.
17. Гржибовский А. М., Горбатова М. А., Наркевич А. Н., Виноградов К. А. Необходимый объем выборки для сравне-

- ния долей в двух независимых группах / Морская медицина. 2020. Т. 6. № 3. С. 76 – 83.
18. **Тихова Г. П.** Планируем клиническое исследование. Вопрос № 1: Как определить необходимый объем выборки? / Региональная анестезия и лечение острой боли. 2014. Т. 8. № 3. С. 57 – 63.
 19. **Тихова Г. П.** Планируем клиническое исследование. Вопрос № 1: как определить необходимый объем выборки? / Здоровоохранение (Минск). 2016. № 9. С. 47 – 53.
 20. **Шпитонков М. И.** Моделирование размера выборки в клинических испытаниях для бинарных данных / Исследование операций (модели, системы, решения). 2018. № 4(13). С. 28 – 35.
 21. **Шпитонков М. И.** Моделирование размера выборки в клинических испытаниях в случае типа «не хуже» / Исследование операций (модели, системы, решения). 2020. Т. 6. С. 39 – 45.
 22. **Ветров Л. Г., Кузнецова А. А., Сунчалина А. Л.** Прикладная статистика: методические указания к выполнению лабораторных работ. — М.: Московский государственный технический университет им. Н. Э. Баумана, 2017. — 52 с.
 23. **Орлов А. И.** Прикладной статистический анализ. — М.: Ай Пи Ар Медиа, 2022. — 812 с.
 24. **Орлов А. И.** О требованиях к статистическим методам анализа данных (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2023. Т. 89. № 11. С. 98 – 106. DOI: 10.26896/1028-6861-2023-89-11-98-106
 25. **Орлов А. И.** Искусственный интеллект: статистические методы анализа данных. — М.: Ай Пи Ар Медиа, 2022. — 843 с.
 26. **Орлов А. И.** Новая парадигма прикладной статистики / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 1. Ч. I. С. 87 – 93.
 27. **Орлов А. И.** Устойчивые экономико-математические методы и модели. — М.: Ай Пи Ар Медиа, 2022. — 337 с.
 28. **Орлов А. И.** Метод статистических испытаний в прикладной статистике / Заводская лаборатория. Диагностика материалов. 2019. Т. 85. № 5. С. 67 – 79. DOI: 10.26896/1028-6861-2019-85-5-67-79
 29. **Levin D. M., Stefan D., Krebil' T. S., Berenson M. L.** Statistics for managers using Microsoft Excel. 4th edition. — Moscow: Vil'yams, 2004. — 1312 p. [in Russian].
 30. **Makarova M. N., Shekunova E. V., Rybakova A. V., Makarov V. G.** Sample size of laboratory animals for experimental studies / Farmatsiya. 2018. Vol. 67. No. 2. P. 3 – 8 [in Russian].
 31. **Koshevoi O. S., Karpova M. K.** Determining the size of the sample population when conducting regional sociological research / Izv. Vuzov. Povolzh. Reg. Obshch. Nauki. 2011. No. 2. P. 98 – 104 [in Russian].
 32. **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. Belgorod. Gos. Univ. Ser. Ekon. Inform. 2015. No. 1(198). P. 74 – 78 [in Russian].
 33. **Motrenko A. P.** Estimation of the required sample size of patients when predicting cardiovascular diseases / Mash. Obuch. Anal. Dannyykh. 2012. Vol. 1. No. 3. P. 354 – 366 [in Russian].
 34. **Rebrova O. Yu., Gusev A. V.** Sample size calculation for clinical trials of clinical decision support systems with binary response / Sovr. Tekhnol. Med. 2022. Vol. 14. No. 3. P. 6 – 14 [in Russian].
 35. **Narkevich A. N., Vinogradov K. A.** Methods for determining the minimum required sample size in medical research / Sots. Aspekty Zdor. Naseleniya. 2019. Vol. 65. No. 6. P. 10 [in Russian].
 36. **Grzhibovskii A. M., Gorbatova M. A., Narkevich A. N., Vinogradov K. A.** Sample size for correlation analysis / Mor. Meditsina. 2020. Vol. 6. No. 1. P. 101 – 106 [in Russian].
 37. **Grzhibovskii A. M., Gorbatova M. A., Narkevich A. N., Vinogradov K. A.** Required sample size to compare arithmetic means in two independent groups / Mor. Meditsina. 2020. Vol. 6. No. 2. P. 106 – 113 [in Russian].
 38. **Grzhibovskii A. M., Gorbatova M. A., Narkevich A. N., Vinogradov K. A.** Required sample size to compare proportions in two independent groups / Mor. Meditsina. 2020. Vol. 6. No. 3. P. 76 – 83 [in Russian].
 39. **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. P. 57 – 63 [in Russian].
 40. **Tikhova G. P.** We are planning a clinical trial. Question No. 1: how to determine the required sample size? / Zdravookhranenie. 2016. No. 9. P. 47 – 53 [in Russian].
 41. **Shpitonkov M. I.** Sample size modeling in clinical trials for binary data / Issl. Oper. (Mod., Sist., Resh.). 2018. No. 4(13). P. 28 – 35 [in Russian].
 42. **Shpitonkov M. I.** Modeling sample size in non-inferiority clinical trials / Issl. Oper. (Mod., Sist., Resh.). 2020. Vol. 6. P. 39 – 45 [in Russian].
 43. **Vetrov L. G., Kuznecova A. A., Sunchalina A. L.** Applied statistics: guidelines for performing laboratory work. — Moscow: MGTU im. N. É. Bauman, 2017. — 52 p. [in Russian].
 44. **Orlov A. I.** Applied statistical analysis. — Moscow: IPR MEDIA, 2022. — 812 p. [in Russian].
 45. **Orlov A. I.** On the requirements for statistical methods of data analysis (generalising article) / Industr. Lab. Mater. Diagn. 2023. Vol. 89. No. 11. P. 98 – 106 [in Russian]. DOI: 10.26896/1028-6861-2023-89-11-98-106
 46. **Orlov A. I.** Artificial intelligence: statistical methods for data analysis. — Moscow: IPR MEDIA, 2022. — 843 p. [in Russian].
 47. **Orlov A. I.** The New Paradigm for Applied Statistics / Industr. Lab. Mater. Diagn. 2012. Vol. 78. No. 1. Pt. I. P. 87 – 93.
 48. **Orlov A. I.** Stability economic and mathematical methods and models. — Moscow: IPR MEDIA, 2022. — 337 p.
 49. **Orlov A. I.** Statistical simulations method in applied statistics / Industr. Lab. Mater. Diagn. 2019. Vol. 85. No. 5. P. 67 – 79 [in Russian]. DOI: 10.26896/1028-6861-2019-85-5-67-79

REFERENCES