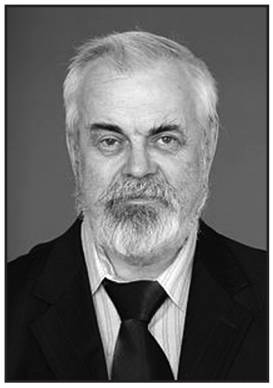

Орлов
Александр Иванович
доктор экон. наук, доктор техн. наук, канд. физ.-мат. наук, профессор,
зав. лаб. экономико-математических методов в контроллинге
МГТУ им. Н.Э. Баумана

УДК 303.5:519.2

О ВЫБОРЕ ОБЪЕМА ВЫБОРКИ

При планировании статистических исследований необходимо выбрать объем выборки. Автором получены правила расчета необходимого объема выборки при изучении значений вероятностей и математических ожиданий. Рассмотрены два подхода. В первом из них рассматриваются задачи оценивания с заданной точностью, т.е. полушириной доверительного интервала. Во втором речь идет о выборе между нулевой и альтернативной гипотезами, исходя из заданных значимости и мощности статистического критерия. При известной точности исходных измерений выбрать необходимый объем выборки позволяет статистика интервальных данных. Рассмотрен общий подход на основе принципа уравнивания статистических и метрологических погрешностей. Как пример проанализировано оценивание математического ожидания.

Ключевые слова: *статистические методы, необходимый объем выборки, доверительные интервалы, мощность критерия, статистика интервальных данных.*

Orlov Alexander, *Doctor of Economics, Doctor of Technical Sciences, PhD in Physical and Mathematical Sciences, Professor, Head of the laboratory of economic and mathematical methods in controlling, BMSTU*

ABOUT CHOOSING THE SAMPLE SIZE

When planning statistical research, it is necessary to select the sample size. The author obtained the rules for calculating the required sample size when studying the values of probabilities and mathematical expectations. Two approaches are considered. In the first, the problems of estimation with a given accuracy, i.e., the half-width of the confidence interval, are considered. The second one deals with the choice between the null and alternative hypotheses, based on the given significance and power of the statistical test. With a known accuracy of the initial measurements, the statistics of interval data allows you to select the required sample size. A general approach based on the principle of equalization of statistical and metrological errors is considered. As an example, the estimation of mathematical expectation is analyzed.

Keywords: *statistical methods, required sample size, confidence intervals, test power, interval data statistics.*

Введение

При проведении научных и практических работ, составной частью которых является статистический анализ данных, часто возникает вопрос: какой объем выборки выбрать? Для обоснования необходимого объема выборки разработан ряд подходов, рассмотрению которых посвящена настоящая статья. Ее необходимость вызвана тем, что затрагивающие эту тему электронные и бумажные источники часто неполны, в них встречаются неточности и ошибки.

С точки зрения классической математической статистики ответ на поставленный вопрос ясен: чем больше объем выборки, тем лучше. Однако на практике приходится учитывать объем ресурсов, необходимых для сбора данных, поскольку при увеличении объема выборки точность оценивания улучшается все меньше. А именно, при увеличении объема выборки в k раз среднее отклонение от оцениваемого параметра уменьшается лишь как $1/\sqrt{k}$. Выбор объема выборки приходится делать на основе компромисса между точностью и стоимостью исследования. Например, в социологии число опрошенных обычно не превышает несколько тысяч.

Необходимый объем выборки можно найти, задав точность статистических выводов. Например, найти ее, задав ширину доверительного интервала при оценивании параметра или исходя из мощности критерия при выборе между двумя гипотезами. Если известна точность проводимых измерений, то ответить на вопрос о необходимом объеме выборки позволяет статистика интервальных данных. Она исходит из принципа уравнивания погрешностей: статистическая погрешность должна равняться метрологической.

Рассмотрим подробнее перечисленные подходы к выбору объема выборки, разобрав типовые примеры постановок статистических задач оценивания и проверки гипотез.

Согласно [1], начать надо с выбора вероятностно-статистической модели.

Пусть анализируемые данные X_1, X_2, ..., X_n рассматриваются как выборка, т.е. набор независимых (в совокупности) одинаково распределенных случайных величин. Будем использовать термины, определения и факты, приведенные в справочнике [2].

Необходимый объем выборки при оценивании вероятности

Пусть элементы выборки принимают только два значения 0 и 1 (например, 0 соответствует тому, что деталь годная, а 1 – что 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). Как установлено в прикладной статистике [3], доверительный интервал $(p_н, p_в)$ для неизвестного параметра p строится по объему выборки n и числу ответов 1 в выборке, равному $X_1 + X_2 + ... + X_n = X$, следующим образом:

$$p_н = p^* - C(\gamma)\sqrt{\frac{p^*(1-p^*)}{n}}, p_в = p^* + C(\gamma)\sqrt{\frac{p^*(1-p^*)}{n}}, (1)$$

где $p_н$ – нижняя доверительная граница, $p_в$ – верхняя доверительная граница, $p^* = X/n$ – выборочная доля (доля ответов 1 в выборке), γ – доверительная вероятность, $C(\gamma)$ – коэффициент, соответствующий доверительной вероятности:

$$C(\gamma) = \Phi^{-1}\left(\frac{1+\gamma}{2}\right),$$

где $\Phi^{-1}(x)$ – функция, обратная к функции стандартного нормального распределения 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 [3] на основе теоремы Муавра-Лапласа в предположении безграничного роста объема выборки. При решении практических задач их можно использовать при объеме выборки $n \geqslant 10$.

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:

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

Из формулы (2) следует, что:

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

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

Эту формулу нельзя непосредственно применить для определения необходимого объема выборки, поскольку $p^* = X/n$, т.е. знание объема выборки необходимо для применения формулы (3). Есть два пути для преодоления этой сложности.

Во-первых, величина p^* может быть оценена по результатам предыдущих исследований или в результате анализа данных предварительной выборки небольшого объема.

Во-вторых, можно воспользоваться тем, что:

$$p^*(1 - p^*) \leqslant 1/4,$$

причем равенство достигается при $p^* = 1/2$. Следовательно, вместо (3) можно использовать формулу:

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

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 в этом случае можно заменить формулу (4) на более простую:

$$n \geqslant \frac{1}{\Delta^2}. \quad (5)$$

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, соответственно 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

В этом случае элементы выборки X_1, X_2, ..., X_n могут принимать любые числовые значения. Согласно принятой модели элементы выборки – независимые одинаково распределенные случайные величины. Поскольку все они одинаково распределены, то математические ожидания и дисперсии элементов выборки совпадают:

$$M(X_i) = a,\ M(X_i) = \sigma^2,\ i = 1, 2, ..., n.$$

Как известно [3], точечной оценкой математического ожидания является выборочное среднее арифметическое:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n},$$

а доверительный интервал $(a_н, a_в)$ для математического ожидания a имеет вид:

$$a_н = \overline{X} - C(\gamma)\frac{s}{\sqrt{n}},\ a_в = \overline{X} + C(\gamma)\frac{s}{\sqrt{n}},$$

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

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

Аналогом формулы (2) является соотношение:

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

Следовательно, необходимый объем выборки – это минимальное n такое, что:

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

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 в результате 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 использовать формулу:

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

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 $\overline{X} \pm s$. В этом случае $C(\gamma) = 1{,}00$, т.е. $\gamma = 0{,}68$.

Выше рассмотрены две задачи оценивания. Используются и другие задачи оценивания. Во всех из них необходимый объем выборки находят путем приравнивания полуширины симметричного доверительного интервала и заданной точности.

Необходимый объем выборки при проверке статистической гипотезы о вероятности

Построение вероятностно-статистической модели [1] начинается с задания двух гипотез –

нулевой H_0 и альтернативной H_1. Рассмотрим две типовые постановки задачи проверки статистических гипотез.

Задача 1. Если случайные величины принимают два значения, то задают 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). Как известно (см., например, [2]):

$$P(X = k) = C_n^k p^k (1 - p)^{n-k},\ k = 1, 2, ..., n,$$

где C_n^k – число сочетаний из n элементов по k. Исходя из X и n, необходимо принять решение о том, какая из двух гипотез верна. Например, по объему выборки и числу дефектных единиц в ней надо установить, какова доля дефектных единиц в партии – p_0 или p_1.

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

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

По выборке 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 [4, гл. 10].

Для проверки статистических гипотез используют статистические критерии [2-3], т.е. функции от статистических данных (от объема выборки и числа дефектных единиц в задаче 1 и от X_1, X_2, ..., 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 задается исследователем, обычно принимают, что $\alpha = 0{,}05$. Необходимый объем выборки находят, задав требуемую мощность критерия.

В задаче 1 для проверки нулевой гипотезы в [3] рекомендуется статистический критерий, основанный на статистике:

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

Как вытекает 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:

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

Правило принятия решения на основе значения Y имеет следующий вид:

– если $|Y| \leqslant k(\alpha)$, то принимают нулевую гипотезу, если же $|Y| > k(\alpha)$, то отклоняют нулевую гипотезу и принимают альтернативную.

Здесь $k(\alpha)$ – коэффициент, зависящий 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). Из двух последних равенств получаем, что:

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

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

Перейдем к изучению мощности критерия. Она равна вероятности того, что при справедливости альтернативной гипотезы нулевая гипотеза отклоняется. Таким образом, необходимо найти вероятность события $|Y| > k(\alpha)$ в предположении, что случайная величина Y имеет биномиальное распределение с параметром $p = p_1$ и объемом выборки n.

Справедливо тождество:

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

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 $p_1 > p_0$. Введем обозначения:

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

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

Тогда:

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

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

Поскольку в предположении, что случайная величина Y имеет биномиальное распределение с параметром $p = p_1$, распределение центрированной 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, то распределение Y сближается с распределением $W = AZ + B\sqrt{n}$, что позволяет рассчитать мощность критерия.

Поскольку:

$$P(|Y| > k(\alpha)) = P(|AZ + B\sqrt{n}| > k(\alpha)) =$$
$$= P(AZ + B\sqrt{n} < -k(\alpha)) + P(AZ + B\sqrt{n} > k(\alpha)),$$

то мощность рассматриваемого критерия проверки статистических гипотез равна:

$$P(Z < \frac{-k(\alpha) - B\sqrt{n}}{A}) + P(Z > \frac{k(\alpha) - B\sqrt{n}}{A}).$$

Поскольку Z, имеет стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, то мощность выражается через функцию Лапласа $\Phi(u)$ следующим образом:

$$f(A, B, n) = 1 + \Phi\left(\frac{-k(\alpha) - B\sqrt{n}}{A}\right) - \Phi\left(\frac{k(\alpha) - B\sqrt{n}}{A}\right). \quad (10)$$

Необходимый объем выборки находят из уравнения:

$$f(A, B, n) = 1 - \beta. \quad (11)$$

Поскольку, как легко видеть:

$$\lim_{n\to\infty} f(A, B, n) = 1,$$

то для сколь угодно малой вероятности ошибки второго рода β (ее задают при постановке задачи определения необходимого объема выборки) решение задачи (11) существует. К сожалению, найти его можно лишь численными методами. Оценку сверху можно найти, отбросив второй член в (10), т.е. из уравнения:

$$1 - \Phi\left(\frac{k(\alpha) - B\sqrt{n}}{A}\right) = 1 - \beta. \quad (12)$$

Из (12) следует, что:

$$\frac{k(\alpha) - B\sqrt{n}}{A} = \Phi^{-1}(\beta),\ n = \left(\frac{k(\alpha) - A\Phi^{-1}(\beta)}{B}\right)^2. \quad (13)$$

Как уже говорилось, уровень значимости α обычно принимают равным 0,05, тогда $k(\alpha) = 1{,}96$.

Рассмотрим пример. Пусть $p_0 = 0{,}1$ и $p_1 = 0{,}5$. Тогда:

$$A = \frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_0(1-p_0)}} = \frac{\sqrt{0{,}5 \cdot 0{,}5}}{\sqrt{0{,}1 \cdot 0{,}9}} = 1{,}67,$$

$$B = \frac{p_1 - p_0}{\sqrt{p_0(1-p_0)}} = 1{,}33$$

и объем выборки (оценка сверху, т.е. заведомо достаточный объем) равен:

$$n = \left(\frac{1{,}96 - 1{,}67\Phi^{-1}(\beta)}{1{,}33}\right)^2.$$

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

$$n = \left(\frac{1{,}96 + 1{,}67 \cdot 2{,}33}{1{,}33}\right)^2 = 19{,}36.$$

Поскольку объем выборки – натуральное число, то, округляя до ближайшего натурального числа сверху, получаем, что необходимый объем выборки – это 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 $X_1, X_2, ..., X_n$ необходимо принять решение о том, какая из двух гипотез верна. Обычно принимают вероятностно-статистическую модель [1], согласно которой при справедливости нулевой гипотезы H_0 элементы выборки имеют функцию распределения $F(x)$ с дисперсией σ_0, а при справедливости альтернативной гипотезы H_1 – некоторую другую функцию распределения $G(x)$ с дисперсией σ_1. 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 предполагается, что существуют первые три центральные момента, а потому при справедливости обеих гипотез спра-

ведлива Центральная предельная теорема теории вероятностей (как следствие теоремы Ляпунова).

Статистика критерия для проверки нулевой гипотезы согласно [3] имеет вид:

$$Y = \sqrt{n}\frac{\overline{X} - a_0}{s}.$$

Согласно Центральной предельной теореме теории вероятностей, распределение статистики Y сходится при росте объема выборки к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1, а потому решающее правило, соответствующее уровню значимости α, таково:

– если $|Y| \leqslant k(\alpha)$, то принимают нулевую гипотезу, если же $|Y| > k(\alpha)$, то отклоняют нулевую гипотезу и принимают альтернативную.

Здесь, как и в предыдущем разделе:

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

С целью изучения мощности рассмотрим распределение Y при справедливости альтернативной гипотезы H_1: $M(X) = a_1$. Имеем:

$$Y = \sqrt{n}\frac{\overline{X} - a_1 + (a_1 - a_0)}{s} =$$
$$= \sqrt{n}\frac{\overline{X} - a_1}{s} + \sqrt{n}\frac{(a_1 - a_0)}{s}. \quad (14)$$

Распределение первого слагаемого в (14) сходится к стандартному нормальному, для второго справедливо предельное соотношение:

$$\lim_{n\to\infty}\frac{a_1 - a_0}{s} = \frac{a_1 - a_0}{\sigma_1} = B. \quad (15)$$

Следовательно, распределение Y сближается с распределением случайной величины $W = Z + B\sqrt{n}$, где Z имеет стандартное нормальное распределение. Это позволяет рассчитать мощность критерия. Она равна (в асимптотике):

$$P(|Y| > k(\alpha)) = P(|Z + B\sqrt{n}| > k(\alpha)) =$$
$$= P(Z > k(\alpha) - B\sqrt{n}) + P(Z < -k(\alpha) - B\sqrt{n}) =$$
$$= 1 - \Phi(k(\alpha) - B\sqrt{n}) + \Phi(-k(\alpha) - B\sqrt{n}) = g(B, n).$$

Аналогично предыдущему разделу необходимый объем выборки по заданной мощности критерия находят из уравнения:

$$g(B, n) = 1 - \beta.$$

Справедливы и рассуждения, приведенные в конце предыдущего раздела. Если математические ожидания или дисперсии, соответствующие рассматриваемым гипотезам, неизвестны, то при достаточно большом объеме выборки их можно заменить соответствующими состоятельными оценками [3].

Рассмотренное выше решающее правило соответствует двусторонней альтернативе – отклонение значения математического ожидания может быть как в большую, так и в меньшую сторону. Рассмотрим другую постановку – пусть заранее известно, что a_0 меньше, чем a_1, т.е. отклонение может быть только в большую сторону. Внесем соответствующие изменения в цепь рассуждений.

Решающее правило, соответствующее уровню значимости α, таково: если $Y \leqslant q(\alpha)$, то принимают нулевую гипотезу, если же $Y > q(\alpha)$, то отклоняют нулевую гипотезу и принимают альтернативную. Здесь при справедливости нулевой гипотезы:

$$P(Y > q(\alpha)) = 1 - P(Y \leqslant q(\alpha)) = \alpha.$$

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 $q(0{,}05) = \Phi^{-1}(0{,}95) = 1{,}64$.

В соответствии с (14) и (15) найдем мощность критерия. При справедливости альтернативной гипотезы и достаточно большом объеме выборки:

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

Необходимый объем выборки по заданной мощности критерия находят из уравнения:

$$1 - \Phi(q(\alpha) - B\sqrt{n}) = 1 - \beta. \quad (16)$$

Из (16) следует, что:

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

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

$$n = \left(\frac{1{,}96 + 2{,}33}{B}\right)^2 = \frac{18{,}4}{B^2} = \frac{18{,}4\sigma_1^2}{(a_1 - a_0)^2}. \quad (17)$$

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 (17) находим n = 41,4. Поскольку объем выборки – натуральное

число, то, округляя до ближайшего натурального числа сверху, получаем, что необходимый объем выборки – это 42.

Необходимый объем выборки на основе статистики интервальных данных

В предыдущих разделах 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. Поэтому интересен и полезен метод на основе статистики интервальных данных, в котором привлекаются соображения метрологии.

Если известна точность проводимых измерений, то ответить на вопрос о необходимом объеме выборки позволяет статистика интервальных данных [3, гл. 12]. Она исходит из принципа уравнивания погрешностей: статистическая погрешность должна равняться метрологической. Статистика интервальных данных разработана в рамках новой парадигмы математических методов исследования [5].

В различных задачах математической статистики обычно исходят из базового понятия выборки $X_1, X_2, ..., X_n$ и рассматривают значения тех или иных функций $f(X_1, X_2, ..., X_n)$ 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 приходится делать не на основе $f(X_1, X_2, ..., X_n)$, а на основе искаженного значения $f(Y_1, Y_2, ..., Y_n)$. Здесь $Y_i = X_i + \varepsilon_i$, $i = 1, 2, ..., n$, где $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ – вектор погрешностей.

Важен частный случай, когда $|\varepsilon_i| \leqslant \Delta$. 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, $i = 1, 2, ..., n$, а интервалы $[a_i, b_i]$, $i = 1, 2, ..., n$, где $a_i = X_i - \Delta$, $b_i = X_i + \Delta$. Методам обработки подобных интервальных данных посвящена статистика интервальных данных [3, гл.12] – сравнительно новая область современной прикладной математической статистики (она 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:

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

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

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

$$N_f(X_1, X_2, ..., X_n) = \left(\sum_{i=1}^{n} \left|\frac{\delta f(X_1, X_2, ..., X_n)}{\delta X_i}\right|\right)\Delta. \quad (18)$$

В терминах метрологии $f(X_1, X_2, ..., X_n)$ – это косвенное измерение на основе прямых измерений X_i, $i = 1, 2, ..., n$. Погрешность косвенного измерения равна $N_f(X_1, X_2, ..., X_n)$, тогда результат косвенного измерения есть $f(X_1, X_2, ..., X_n) \pm N_f(X_1, X_2, ..., X_n)$.

Формула (18) позволяет находить погрешность косвенного измерения. При решении задач статистического анализа данных обычно можно доказать, что существует предел:

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

Тогда при достаточно больших объемах выборки n нотна равна $C\Delta$. Следовательно, погрешность вычисления статистики $f(X_1, X_2, ..., X_n)$ 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 статистика $f(X_1, X_2, ..., X_n)$ в большинстве случаев является асимптотически нормальной, т.е. существуют константы a и σ такие, что:

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

где, $\Phi(x)$, как и в предыдущих рассуждениях настоящей статьи, является функцией стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. В рассматриваемых случаях, как правило, выполнены предельные соотношения для моментов:

$$\lim_{n\to\infty}\sqrt{n}(Mf(X_1, X_2, ..., X_n) - a) = 0.$$

$$\lim_{n\to\infty} nDf(X_1, X_2, ..., X_n) = \sigma^2.$$

В соответствии с этими соотношениями в классической математической статистике средний квадрат ошибки рассматриваемой статистической оценки таков:

$$M(f(X_1, X_2, ..., X_n) - a)^2 = (Mf(X_1, X_2, ..., X_n) - a)^2 + Df(X_1, X_2, ..., X_n) = \frac{\sigma^2}{n}. \quad (19)$$

с точностью до бесконечно малых более высокого порядка.

В статистике интервальных данных, в которой мы учитываем наличие погрешностей наблюдений, вместо формулы (19) обычно получаем другое выражение для квадрата ошибки:

$$\sup_{\{\varepsilon\}} M(f(X_1, X_2, ..., X_n) - a)^2 = \frac{\sigma^2}{n} + N_f^2(X_1, X_2, ..., X_n) + o\left(\Delta^2 + \frac{1}{n}\right), \quad (20)$$

где супремум берется по множеству всех возможных погрешностей $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$. Первое слагаемое в правой части (20) соответствует статистической погрешности, второе –метрологической.

Согласно (12) при росте объема выборки средний квадрат ошибки не стремится к 0, как в классической математической статистике, но всегда остается больше некоторой положительной константы (квадрата нотны). Следовательно, нерационально безгранично увеличивать объем выборки. В соответствии с теорией устойчивости экономико-математических методов и моделей [6] используем здесь принцип уравнивания погрешностей, согласно которому целесообразно уравнять величины погрешностей различной природы. Применительно к (20) принцип уравнивания погрешностей приводит к соотношению:

$$\frac{\sigma^2}{n} = N_f^2(X_1, X_2, ..., X_n). \quad (21)$$

Формула (21) позволяет найти необходимый объем выборки. Он равен:

$$n = \frac{\sigma^2}{N_f^2(X_1, X_2, ..., X_n)}. \quad (22)$$

Для практического применения формулы (22) надо заменить теоретические значения дисперсии σ^2 рассматриваемой статистики $f(X_1, X_2, ..., X_n)$ и квадрата ее нотны $N_f^2(X_1, X_2, ..., X_n)$ 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 [3]. В статистике интервальных данных формула (22) 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 $X_1, X_2, ..., X_n$ математического ожидания ее элементов с помощью выборочного среднего арифметического. В этом случае:

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

Поскольку:

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

то согласно формуле (18) нотна выборочного среднего арифметического равна:

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

Согласно Центральной предельной теореме теории вероятностей:

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

Следовательно, необходимый объем выборки равен:

$$n = \frac{\sigma^2}{N_f^2(X_1, X_2, ..., X_n)} = \frac{\sigma^2}{\Delta^2}. \quad (23)$$

Например, если $\sigma = 1$, $\Delta = \frac{1}{6}$ (т.е. среднее квадратическое отклонение результата измерения в 6 раз превосходит точность измерения), то необходимый объем выборки равен 36.

Если точность измерения увеличивается, т.е. Δ уменьшается, то согласно (23) необходимый объем выборки увеличивается. И, наоборот, при применении средства измерения с малой точно-

стью (т.е. с большим Δ) достаточно использовать небольшой объем выборки.

Если используемая в формуле (23) теоретическая дисперсия 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 при изучении значений вероятностей и математических ожиданий случайной величины.

Для каждой из этих статистических задач в рамках классической математической статистики рассмотрены 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. Автором рассмотрен общий подход на основе принципа уравнивания показателей статистических и метрологических погрешностей. В качестве примера проанализировано оценивание математического ожидания.

Литература:

1. Орлов А. И. Контроллинг статистических методов // Журнал «Контроллинг». 2022. № 4 (86). С. 2-11.
2. Орлов А.И. Вероятность и прикладная статистика: основные факты: справочник. – М.: КноРус, 2023. – 190 с.
3. Орлов А.И. Прикладной статистический анализ. – М.: Ай Пи Ар Медиа, 2022. – 812 с.
4. Орлов А.И. Искусственный интеллект: статистические методы анализа данных. – М.: Ай Пи Ар Медиа, 2022. – 843 с.
5. Орлов А.И. О новой парадигме математических методов исследования // Научный журнал КубГАУ. 2016. № 122. С. 807-832.
6. Орлов А.И. Устойчивые экономико-математические методы и модели. – М.: Ай Пи Ар Медиа, 2022. – 337 с.

References:

1. Orlov A. I. Kontrolling statisticheskih metodov // Zhurnal «Kontrolling». 2022. № 4 (86). S. 2-11.
2. Orlov A.I. Verojatnost' i prikladnaja statistika: osnovnye fakty: spravochnik. – M.: KnoRus, 2023. – 190 s.
3. Orlov A.I. Prikladnoj statisticheskij analiz. – M.: Aj Pi Ar Media, 2022. – 812 c.
4. Orlov A.I. Iskusstvennyj intellekt: statisticheskie metody analiza dannyh. – M.: Aj Pi Ar Media, 2022. – 843 c.
5. Orlov A.I. O novoj paradigme matematicheskih metodov issledovanija // Nauchnyj zhurnal KubGAU. 2016. № 122. S. 807-832.
6. Orlov A.I. Ustojchivye jekonomiko-matematicheskie metody i modeli. – M.: Aj Pi Ar Media, 2022. – 337 c.