УДК 519.253

Ю.Я. Водянников, А.Е. Нищенко, С.В. Кукин

ВЫБОРКА ОБЪЕКТОВ ИЗ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ПО ЗАДАННОМУ ЗНАЧЕНИЮ ПРЕДЕЛЬНОЙ ОШИБКИ

Изложена методика оценки предельной ошибки при выборке объектов из генеральной совокупности, базирующаяся на генерации случайных чисел, моделирующих генеральную совокупность. Приведены формулы для определения объема выборки при заданной величине предельной ошибки. Построены номограммы.

Для вычисления количества элементов, отбор которых обеспечил бы репрезентативность выборки, необходимо иметь представление о содержащихся в генеральной совокупности ошибках, однако до исследования аудитор может лишь предполагать их наличие (или отсутствие) и размер.

При определении объема (размера) выборки возникает необходимость установить риск выборки, допустимую и ожидаемую ошибки. Большинство используемых в мировой практике методов расчета объема выборки строятся именно на этих трех показателях.

Основу статистического исследования составляет множество данных, полученных в результате измерения одного или нескольких признаков. Реально наблюдаемая совокупность объектов статистического распределения $x_1, x_2, x_3, x_4 \cdots x_n$, случайной величины X, является выборкой, а гипотетически существующая (домысленная) – генеральной совокупностью. Генеральная совокупность может быть конечной (число наблюдений $N = const$) или бесконечной ($N = \infty$) а выборка из генеральной совокупности – это результат ограниченного ряда n наблюдений.

Достоверность статистических выводов и содержательная интерпретация результатов зависит от репрезентативности выборки, т. е. полноты и адекватности представления свойств генеральной совокупности, по отношению к которой эту совокупность можно считать представительной. Изучение статистических свойств совокупности может быть организована двумя способами: с помощью сплошного и не сплошного наблюдений. Сплошное наблюдение предусматривает обследование всех единиц изучаемой совокупности, а не сплошное (выборочное) наблюдение – только его части.

Основа выборки должна быть достоверной, полной и соответствовать цели исследования, а единицы отбора и их характеристики должны соответствовать действительному их состоянию на момент подготовки выборочного наблюдения.

Существует пять основных способов организации выборочного наблюдения:

1. **Простой случайный отбор**, при котором n объектов случайно извлекаются из генеральной совокупности N объектов (например с помощью таблиц или датчика случайных чисел), причем каждая из возможных выборок имеют равную вероятность. Такие выборки называются собственно случайными;

2. **Простой отбор с помощью регулярной процедуры** осуществляется с помощью механической составляющей (например даты, дня недели, месяца, буквы алфавита и др.), полученные таким способом выборки называются механическими;

3. **Стратифицированный отбор**. Стратифицированный отбор заключается в том, что генеральная совокупность объема N подразделяется на подсовокупности или слои (страты) N_1, N_2, N_3, $N_4, \cdots, N_r$, причем $N_1 + N_2 + N_3 + N_4 + \cdots + N_r = N$. В этом случае выборки называется стратифицированными;

4. **Серийный отбор**. Методы серийного отбора используются для формирования серийных или гнездовых выборок. Они удобны для обследования «блока» или серии объектов;

5. **Комбинированный отбор**. Комбинированный (ступенчатый) отбор может сочетать в себе сразу несколько способов отбора (например стратифицированный и случайный или случайный и механический).

По методу отбора различают повторную и бесповторную выборку.

Задачей всякого исследования является оценка величины ошибки выборки. При любом статистическом наблюдении могут встретиться ошибки двух видов: регистрации и репрезентативности. Ошибки регистрации могут иметь случайный и систематический характер. Случайные ошибки складываются из множества различных неконтролируемых причин. Систематические ошибки обусловлены в основном погрешностью средств измерений. Ошибки репрезентативности присущи только выборочному наблюдению, их невозможно избежать и они возникают в результате того, что выборочная совокупность не полностью воспроизводит генеральную.

Ошибка выборочного наблюдения ξ определяется как разность между значением параметра в генеральной совокупности и ее выборочным значением.

Если объем выборки достаточно большой (n > 20-30), то распределение выборочной средней $\bar{X}^*$, согласно центральной предельной теореме, независимо от характера генерального распределения приближается к нормальному распределению с параметрами [1]:

$$\sigma(\bar{X}^*) = \frac{\sigma}{\sqrt{n}} \qquad M(\bar{X}^*) = \bar{X}\ , \tag{1}$$

где $\bar{X}^*$ -генеральная средняя;

σ - генеральное среднее квадратичное отклонение;

n - объем выборки.

Предельная ошибка выборки (ξ_α) определяется по формуле [1]:

$$\xi_\alpha = z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\ , \tag{2}$$

где z_α- коэффициент, определяемый по вероятности появления случайной ошибки выборки.

При конечной генеральной совокупности объектов может быть использован метод статистического моделирования, состоящий в предварительной генерации матрицы массива случайных чисел с размерностью равной числу объектов генеральной совокупности. Для генерации случайных чисел может быть использована электронная таблица Excel.

Электронная таблица Excel содержит набор встроенных функций категории Статистические, а также предоставляет специальные информационные технологии, выполняемые в среде Пакета анализа.

Для загрузки пакета анализа выполняются следующие действия:

Выполните команду Сервис\Надстройки. На экране появится окно диалога «Надстройки».

Выберите Пакет анализа, а затем нажмите кнопку ОК.

После окончания загрузки в списке опций пункта Сервис основного меню появится строка Анализ данных. При выборе этой строки появляется окно диалога «Анализ данных».

В окне диалога «Анализ данных» отображается список инструментов.

При статистическом моделировании и первичной обработке данных используются следующие инструменты: Генерация случайных чисел, Гистограмма.

Инструмент Генерация случайных чисел заполняет интервал независимыми случайными числами.

При помощи параметра Число переменных можно получить многомерную выборку. Для этого вводится число столбцов в выходной таблице.

Параметром Число случайных чисел определяется число точек данных, которое генерируется для каждой переменной.

Выбор закона распределения случайных чисел задаётся параметром Распределение:

1. Равномерное распределение характеризуется верхней и нижней границами. Вероятность попадания переменной в отрезок фиксированной длины зависит только от длины отрезка и не зависит от его расположения на интервале. Как правило, в приложениях используют равномерное распределение в интервале [0,1].

2. Нормальное распределение характеризуется средним значением и стандартным отклонением. Обычно приложения для этого распределения используют среднее значение 0 и стандартное отклонение 1.

3. Распределение Бернулли характеризуется вероятностью успеха в данном испытании. Случайная величина принимает значение 0 или 1.

4. Биноминальное распределение характеризуется вероятностью успеха для некоторого числа испытаний. Например, вы можете генерировать случайные числа, моделирующие процесс бросания монеты с вероятностью успеха ровно в “k” случаях из “n” испытаний.

5. Распределение Пуассона характеризуется значением Лямбда, равным 1/среднее. Распределение Пуассона часто используется для характеристики числа событий, случающихся в единицу времени, например, число телефонных соединений в минуту.

6. Модельное распределение характеризуется нижней и верхней границей, шагом, числом повторений значений и числом повторений последовательности.

7. Дискретное распределение характеризуется значением и связанным с ним интервалом вероятности. Интервал должен содержать два столбца: левый содержит значения, правый – вероятности, связанные со значением в данной строке. Сумма вероятностей должна быть равна 1.

При помощи параметра Случайное рассеивание фиксируется последовательность выводимых случайных чисел. При повторных запусках генератора можно использовать это значение для получения тех же самых случайных чисел.

В качестве примера определим зависимость ошибки от величины выборки из генеральной совокупности при условии, что исследуемый параметр объекта (появление трещины, отклонение от нормированного значения и др.) подчиняется нормальному закону распределения. Выборка производится простым случайным отбором.

Генеральная совокупность состоит из 10000 объектов.

Среднее квадратичное отклонение случайных чисел, моделирующих генеральную совокупность, составит:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{10000}\left(x_i - \frac{\sum_{i=1}^{10000} x_i}{N}\right)^2}{N}} = 0,9986 \cdot \qquad (3)$$

Расчет предельной ошибки при различных значениях доверительной вероятности и объемов выборки с использованием формулы (2) представлен в таблице 1.

Таблица 1. **Дискретные значения предельных ошибок**

Величина выборки	Доверительная вероятность				
	0,95	0,96	0,97	0,98	0,99
50	0,283664	0,29781	0,315473	0,339409	0,378179
100	0,200581	0,207803	0,219848	0,236098	0,262229
150	0,163774	0,168927	0,178644	0,191733	0,212732
200	0,141832	0,145975	0,15434	0,165599	0,183641
250	0,126858	0,130393	0,137847	0,147878	0,163938
300	0,115805	0,118928	0,125717	0,134848	0,149463
350	0,107215	0,110037	0,116312	0,124749	0,138249
400	0,10029	0,102882	0,108744	0,116625	0,129232
450	0,094555	0,096963	0,102484	0,109906	0,121776
500	0,089702	0,091961	0,097194	0,104229	0,115478
1000	0,063429	0,064941	0,068628	0,073583	0,081499
2000	0,044851	0,04589	0,048493	0,051989	0,057573

Зависимость величины выборки n от значения предельной ошибки ξ_α может быть представлена степенной функцией:

$$n = c \cdot \xi_\alpha^d. \qquad (4)$$

Коэффициенты a и b уравнения (4) определяются методом наименьших квадратов, для этого формула (3) после логарифмирования:

$$ln(n) - ln(c) - d \cdot ln\,(\xi_\alpha) = 0. \qquad (5)$$

и ввода обозначений $y=ln(n)$; $a=ln(c)$; $b=d$; $z= ln(\xi_\alpha)$, преобразовыватся к виду: $y-a+bz$, при этом разрешающее уравнение метода наименьших квадратов принимает вид [2]:

$$U = \sum_{i=1}^{m} (y_i - (a + bz))^2 \quad , \tag{6}$$

где m- число интервалов.

После дифференцирования уравнения (6) по неизвестным коэффициентам a и b система уравнений примет вид:

$$\begin{cases} a \cdot m + \sum_{i=1}^{m} z_i = \sum_{i=1}^{m} y_i \\ a\sum_{i=1}^{m} z_i + b\sum_{i=1}^{m} z_i^2 = \sum_{i=1}^{m} y_i z_i \end{cases} , \tag{7}$$

решение которой определяется выражениями:

$$a = \frac{\sum_{i=1}^{m} y_i \sum_{i=1}^{m} z_i^2 - \sum_{i=1}^{m} z_i y_i \sum_{i=1}^{m} z_i}{m\sum_{i=1}^{m} z_i^2 - (\sum_{i=1}^{m} z_i)^2} , \tag{8}$$

$$b = \frac{m\sum_{i=1}^{m} z_i y_i - \sum_{i=1}^{m} z_i \sum_{i=1}^{m} y_i}{m\sum_{i=1}^{m} z_i^2 - (\sum_{i=1}^{m} z_i)^2} , \tag{9}$$

Используя введенные обозначения, коэффициенты уравнения (3) определятся по формулам:

$$c = \exp\left(\frac{\sum_{i=1}^{m} y_i \sum_{i=1}^{m} z_i^2 - \sum_{i=1}^{m} z_i y_i \sum_{i=1}^{m} z_i}{m\sum_{i=1}^{m} z_i^2 - (\sum_{i=1}^{m} z_i)^2}\right) ; \tag{10}$$

$$d = \frac{m\sum_{i=1}^{m} z_i y_i - \sum_{i=1}^{m} z_i \sum_{i=1}^{m} y_i}{n\sum_{i=1}^{m} z_i^2 - (\sum_{i=1}^{m} z_i)^2} \tag{11}$$

Коэффициент детерминации R^2, определяется по формуле:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (a_i - \cdot \xi_{ai}^{d}))^2}{\sum_{i=1}^{m} n_i^2 - \frac{(\sum_{i}^{m} n_i)^2}{m}} . \tag{12}$$

Аналитические выражения для определения объема выборки по заданному значению предельной ошибки представлены в таблице 2.

Таблица 2. **Аналитические выражения для определения объема выборки**

Доверительная вероятность	Математическое выражение	Коэффициент детерминации
0,95	$n = 4{,}0233 \cdot \xi_\alpha^{-2}$	$R^2 = 1$
0,96	$n = 4{,}4934 \cdot \xi_\alpha^{-1.9753}$	$R^2 = 9999$
0,97	$n = 5{,}0406 \cdot \xi_\alpha^{-1.973}$	$R^2 = 9999$
0,98	$n = 5{,}8309 \cdot \xi_\alpha^{-1.9696}$	$R^2 = 9999$
0,99	$n = 7{,}2279 \cdot \xi_\alpha^{-1.9638}$	$R^2 = 9999$

Полученные математические зависимости табл. 2, представляющие непрерывные функции, позволяют определять величины выборки по заданному значению предельной ошибки.

Так, например, для заданной предельной ошибки 0,1 выборка составит:

$n = 4{,}0233 \cdot 0{,}1^{-2} = 402{,}33 \approx 403$ при p=0,95;

$n = 4{,}4934 \cdot 0{,}1^{-1.9753} = 424{,}49 \approx 425$ при p=0,96;

$n = 5{,}0406 \cdot 0{,}1^{-1.973} = 473{,}68 \approx 474$ при p=0,97;

$n = 5{,}8309 \cdot 0{,}1^{-1.9696} = 543{,}67 \approx 544$ при p=0,98;

$n = 7{,}2279 \cdot 0{,}1^{-1.9638} = 671{,}29 \approx 672$ при p=0,99.

На рис. 1 представлена номограмма для определения величины выборки.

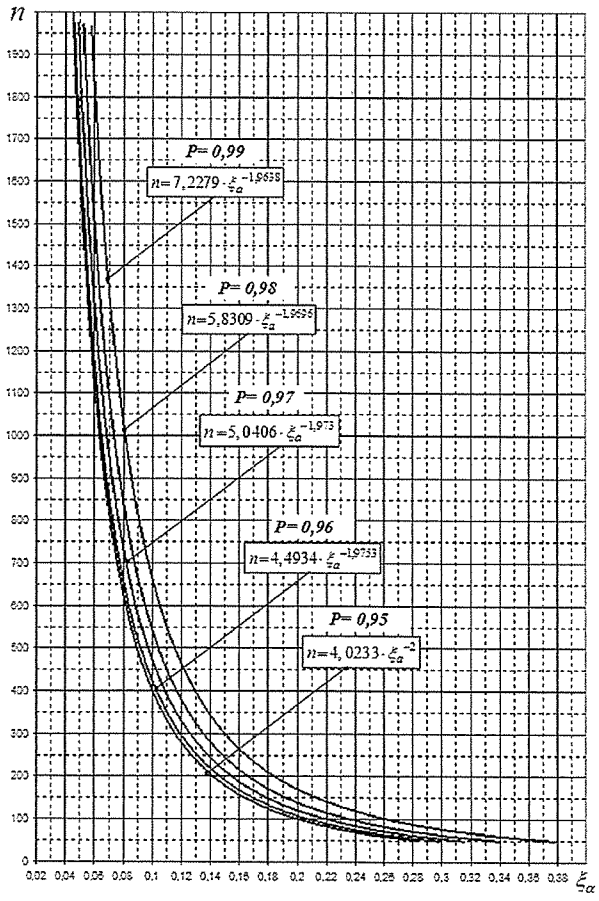

Рис. 1. **Номограммы для определения величины выборки.**

Вывод.

Изложенные расчетные исследования позволяют оценить предельную ошибку в зависимости от величины выборочной совокупности и могут быть использованы для целей исследования статистических закономерностей различных по физической природе объектов.

ЛИТЕРАТУРА

1. Михок Г. Выборочный метод и статистическое оценивание / Г. Михок,В. Урсяну. – М.: Финансы и статистика, 1982.-246 с.

2. Е.Н.Львовский. Статистические методы построения эмпирических формул. - М.: «Высшая школа», 1988 г.