

- Учебник
- для училищ и колледжей

А.Г. Луканкин

МАТЕМАТИКА

- **АЛГЕБРА И НАЧАЛА
МАТЕМАТИЧЕСКОГО АНАЛИЗА**
- **ГЕОМЕТРИЯ**

2-е издание, переработанное и дополненное

Министерство образования и науки РФ

Рекомендовано ФГАУ «Федеральный институт развития образования» в качестве учебника для использования в учебном процессе образовательных учреждений, реализующих программы среднего профессионального образования на базе основного общего образования с получением среднего общего образования по ЕН.02 «Математика»

Регистрационный номер рецензии 175 от 03 июня 2016 года



Москва
ИЗДАТЕЛЬСКАЯ ГРУППА
«ГЭОТАР-Медиа»
2018

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

7.1. ОСНОВНЫЕ ПОНЯТИЯ И ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

7.1.1. ПРЕДМЕТ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

В математической статистике разрабатываются теории и методы обработки информации о массовых явлениях. Исходным материалом для всякого статистического исследования служат *статистические данные*. Под статистическими данными понимаются сведения о числе объектов какой-либо более или менее обширной совокупности, обладающих теми или иными признаками. Статистическими данными, например, могут являться сведения о: а) числе студентов в каждой группе вуза; б) числе жителей города, имеющих рост 2 м; в) числе дорожно-транспортных происшествий за сутки; г) количестве осадков за каждый из дней месяца и т.д.

На основании статистических данных часто можно делать вполне определенные научно обоснованные выводы, представляющие интерес для науки и практики. Для этого статистические данные должны быть предварительно определенным образом систематизированы и обработаны. Математическая статистика изучает и разрабатывает математические методы систематизации, обработки и использования статистических данных для научных и производственных целей.

Методы математической статистики широко применяются в различных областях знаний — в физике, экономике, педагогике, психологии, геологии, биологии, медицине, социологии и др. Также широко используется математическая статистика и в промышленности.

Например, для организации надежного контроля за качеством выпускаемой продукции.

Основой математической статистики служит теория вероятностей, в которой изучаются математические модели реальных случайных явлений. Методы математической статистики позволяют на основе экспериментальных данных определять вероятностные характеристики этих моделей: математические ожидания, дисперсии, законы распределения и многие другие характеристики. Можно сказать, что математическая статистика связывает реальные случайные явления с их математическими вероятностными моделями.

Среди многих задач, решаемых математической статистикой, выделим следующие. В прикладных задачах вероятность исследуемого события обычно неизвестна. Она определяется приближенно по статистическим данным. Одна из основных задач математической статистики — дать оценку вероятности события, полученной на основе опытных данных. В практических задачах нередко удается в результате анализа исследуемой случайной величины определить ее распределение с точностью до некоторого случайного параметра. Дать оценку этого параметра в виде числа или интервала, в котором с заданной вероятностью заключено значение неизвестного параметра, — еще одна основная задача математической статистики.

Математическая статистика возникла в XVII веке одновременно с теорией вероятностей. Решения первых задач математической статистики содержатся в сочинениях создателей теории вероятностей — Я. Бернулли, П. Лапласа, С. Пуассона. В России методы математической статистики в применении к демографии и страховому делу развивал В.Я. Буняковский. Решающее значение для всего дальнейшего развития математической статистики имели работы русской классической школы теории вероятностей П.Л. Чебышева, А.А. Маркова, А.М. Ляпунова. Развитие методов математической статистики связано с именами К. Пирсона, Р. Фишера, Ю. Неймана и советских математиков А.Н. Колмогорова, Ю.В. Линника, Ю.В. Прохорова, Б.В. Гнеденского.

7.1.2. ВЫБОРКИ И ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ

В самых различных областях производственной и научной деятельности приходится проводить изучение (обследование, измерение, проверку) объектов, принадлежащих некоторой совокупности, по какому-либо признаку. При этом иногда приходится исследовать все объекты совокупности, то есть проводить *сплошное исследование*. Но во многих других случаях в силу различных причин исследовать все объекты

невыгодно или даже невозможно. Например, если рассматриваемая совокупность содержит слишком много объектов (перепись населения страны), если исследование требует очень больших затрат времени или является дорогостоящим (плебисцит, то есть опрос общественного мнения населения по тому или иному вопросу). Иногда исследование объекта связано с его уничтожением (например, исследование на прочность приводного ремня). Поэтому на практике гораздо чаще применяется *выборочное исследование*. При выборочном исследовании из всей совокупности отбирают некоторым образом определенное число объектов и только их подвергают исследованию. При этом совокупность всех исследуемых объектов называют *генеральной совокупностью*.

Выборочной совокупностью, или просто *выборкой*, называют совокупность случайно отобранных объектов из генеральной совокупности.

Под случайным отбором при образовании выборки понимают такой отбор, при котором все объекты генеральной совокупности имеют одинаковую вероятность попасть в выборку.

Выборка может производиться двумя основными способами. При первом объект извлекается из генеральной совокупности, исследуется, возвращается в генеральную совокупность; затем снова извлекается некоторый объект совокупности, исследуется и возвращается в генеральную совокупность и т.д. Полученную таким образом выборку называют *повторной*. При втором способе исследования объекты в генеральную совокупность не возвращаются, и выборку в этом случае называют *бесповторной*. Например, открытие банок консервов при исследовании на сертификат качества.

Таким образом, при бесповторной выборке каждый объект исследуется только один раз, при повторной выборке один и тот же объект может подвергаться исследованию несколько раз.

Число объектов выборочной (генеральной) совокупности называют *объемом выборки* (генеральной совокупности). Например, если из 5000 студентов института опрошено 100 студентов, то объем генеральной совокупности $N = 5000$, а объем выборки $n = 100$.

Математическая статистика занимается вопросом: можно ли, установив какое-либо свойство выборки, считать, что это свойство с определенной вероятностью присуще всей генеральной совокупности?

Для того чтобы по выборке можно было с определенной уверенностью судить о всей генеральной совокупности, выборка должна достаточно полно отражать изучаемое свойство объектов генеральной совокупности, быть достаточно *представительной* (*репрезентативной*). Для того чтобы выборка была представительной, необходимо, чтобы

отбор объектов в выборку осуществлялся действительно случайно. Необходимо также, чтобы изучаемому свойству была присуща статистическая устойчивость: при многократном повторении исследования должна иметь место статистическая устойчивость частот наблюдаемых событий.

Для статистической обработки результаты исследования объектов, составляющих выборку, представляют в виде последовательности чисел $x_1, x_2, \dots, x_n$. Это могут быть, например, результаты n измерений длин деталей (в см), попавших в выборку. Числовую выборку получают также при рассмотрении других случайных величин.

Пример 1. Выборка возрастов студентов одной академической группы (берется полное число лет) имеет следующий вид:

19, 18, 18, 20, 17, 18, 19, 18, 17, 18, 19, 20, 17, 18, 18, 19.

Разность наибольшего и наименьшего значений числовой выборки называют *размахом выборки*.

Выборку, представленную в виде неубывающей последовательности чисел, называют *вариационным рядом*. Очевидно, что любую выборку можно записать в виде вариационного ряда. Например, записав значения выборки

19, 18, 18, 20, 17, 18, 19, 18, 17, 18, 19, 20, 17, 18, 18, 19

в виде неубывающей последовательности, получим вариационный ряд

17, 17, 17, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 20, 20.

Размах выборки равен $20 - (17) = 3$.

Пусть при исследовании некоторой генеральной совокупности получена числовая выборка объема n , причем значение x_1 встретилось в выборке n_1 раз, значение x_2 — n_2 раз, ..., x_n — n_n раз. Заметим, что в примере 1 объем выборки $n = 16$, значение 17 представлено в выборке 3 раза, значение 18 — 7 раз, значение 19 — 4 раза, значение 20 — 2 раза. Числа $n_1, n_2, \dots, n_n$ называют *частотами*, а их отношения к объему выборки, то есть отношения $\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_n}{n}$ — *относительными частотами* соответствующих значений $x_1, x_2, \dots, x_n$ выборки. В рассмотренном примере 1 числа 3, 7, 4, 2 — частоты, а числа $\frac{3}{16}, \frac{7}{16}, \frac{4}{16}, \frac{2}{16}$ — относительные частоты значений 17, 18, 19, 20.

Очевидно, что сумма частот равна объему выборки, а сумма относительных частот — единице, то есть

$$n_1 + n_2 + \dots + n_n = n, \quad \frac{n_1}{n} + \frac{n_2}{n} + \dots + \frac{n_n}{n} = 1.$$

Отметим, что в примере 1 при объеме выборки $n = 16$ сумма относительных частот равна

$$\frac{3}{16} + \frac{7}{16} + \frac{4}{16} + \frac{2}{16} = 1.$$

Последовательность пар

$$(x_1; n_1), (x_2; n_2), \dots, (x_n; n_n) \quad (1)$$

называют статистическим рядом.

Обычно статистический ряд записывают в виде таблицы.

x_1	x_2	x_3	...	x_k	...	x_n
n_1	n_2	n_3	...	n_k	...	n_n

(1*)

В первой строке таблицы записывают значения выборки, во второй — их соответствующие частоты.

Рассмотрим таблицу, в которой, в отличие от таблицы (1*), во второй строке записаны не частоты, а относительные частоты соответствующих значений выборки.

x_1	x_2	x_3	...	x_k	...	x_n
$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\frac{n_3}{n}$	...	$\frac{n_k}{n}$	...	$\frac{n_n}{n}$

(2)

Этой таблицей задается так называемое *выборочное распределение*.

Пример. Для выборки

$$2, 4, -1, 4, 0, 5, 4, -1, 4, 5$$

определить объем и размах. Записать выборку в виде вариационного ряда и статистического ряда. Найти выборочное распределение.



Объем выборки $n = 10$, ее размах $5 - (-1) = 6$.

Записав значения выборки в виде неубывающей последовательности, получим вариационный ряд

$$-1, -1, 0, 2, 4, 4, 4, 4, 5, 5.$$

Статистический ряд можно записать в виде последовательности пар чисел:

$$(-1; 2), (0; 1), (2; 1), (4; 4), (5; 2)$$

или в виде таблицы

-1	0	2	4	5
2	1	1	4	2

Для контроля находим сумму частот

$$2 + 1 + 4 + 2 = 10$$

и убеждаемся в том, что она равна объему выборки.

Вычислив относительные частоты, найдем выборочное распределение

-1	0	2	4	5
2/10	1/10	1/10	4/10	2/10

Для контроля убеждаемся в том, что сумма относительных частот равна единице:

$$\frac{2}{10} + \frac{1}{10} + \frac{1}{10} + \frac{4}{10} + \frac{2}{10} = 1.$$

▲

7.1.3. ГРАФИЧЕСКИЕ ИЗОБРАЖЕНИЯ ВЫБОРКИ. ПОЛИГОН И ГИСТОГРАММА

Для наглядного представления о выборке часто используют различные графические изображения выборки. Простейшими изображениями выборки являются полигон и гистограмма. Пусть выборка задана статистическим рядом

$$(x_1; n_1), (x_2; n_2), \dots, (x_n; n_n). \quad (1)$$

Полигоном частот называют ломаную с вершинами в точках (1).

На рис. 7.1 построен полигон частот для выборки, рассмотренной в примере п. 7.1.2.

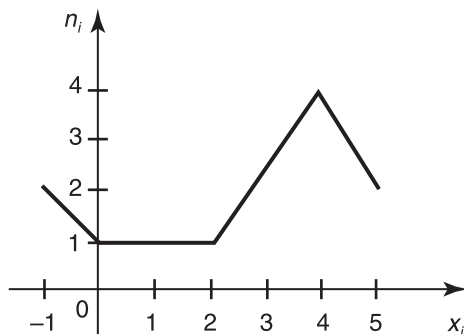


Рис. 7.1

Полигоном относительных частот называют ломаную с вершинами в точках

$$\left(x_1; \frac{n_1}{n}\right), \left(x_2; \frac{n_2}{n}\right), \dots, \left(x_n; \frac{n_n}{n}\right).$$

Очевидно, что полигон относительных частот получается из полигона частот сжатием вдоль оси ординат в n раз, где n — объем выборки.

При большом объеме выборки более наглядное представление о ней дает гистограмма. Для построения гистограммы частот промежутков от наименьшего значения выборки до наибольшего ее значения разбивают на несколько частичных промежутков. Для каждого частичного промежутка подсчитывают сумму s_i частот значений выборки, попавших в этот промежуток. Значение x_i выборки, совпавшее с правым концом промежутка, относят к следующему промежутку (если x_i — не наибольшее значение выборки).

Затем на каждом частичном промежутке, как на основании, строят прямоугольник с высотой $\frac{s_i}{h}$. Объединение всех построенных таким образом прямоугольников называют гистограммой частот.

Таким образом, *гистограммой частот* называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых являются частичные промежутки длины h , а высотами — отрезки длины $\frac{s_i}{h}$, где s_i — сумма частот значений выборки, попавших в i -й промежуток.

Из определения гистограммы ясно, что ее площадь равна объему выборки.

В практических задачах в зависимости от объема выборки в большинстве случаев целесообразно брать 10–20 частичных промежутков.

Аналогично определяют и строят гистограмму относительных частот.

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых являются частичные промежутки длины h , а высотами — отрезки длины $\frac{w_i}{h}$, где w_i — сумма частот значений выборки, попавших в i -й промежуток.

Площадь гистограммы относительных частот равна единице.

Пример. В результате измерения напряжения в электросети (в вольтах) получена выборка

218, 221, 215, 225, 225, 217,
224, 220, 220, 219, 221, 219,
222, 227, 218, 220, 223, 230,
223, 216, 224, 227, 220, 222.

Построить гистограмму частот, если число частичных промежутков равно 5.



Наименьшее значение выборки равно 215, наибольшее — 230. Находим длину частичных промежутков

$$h = \frac{230 - 215}{5} = 3.$$

Подсчитываем с учетом кратности число значений выборки, попавших в каждый промежуток. Для первого промежутка [215; 218) это число равно 3, для второго [218; 221) оно равно 8, для третьего [221; 224) — 6, для четвертого [224; 227) — 5, для пятого [227; 230] — 2. Следовательно, высоты прямоугольников (слева направо), образующих гистограмму, равны

$$1; \frac{8}{3}; 2; \frac{5}{3}; \frac{2}{3}.$$

По полученным данным строим гистограмму (рис. 7.2).

Для контроля убеждаемся в том, что площадь гистограммы равна объему выборки:

$$3 \left(1 + \frac{8}{3} + 2 + \frac{5}{3} + \frac{2}{3} \right) = 24.$$

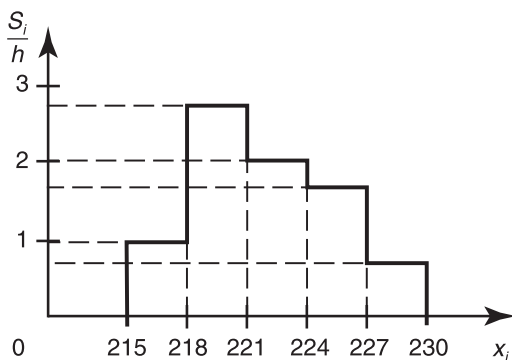


Рис. 7.2

7.1.4. ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ

В теории вероятностей изучались числовые характеристики случайных величин: математическое ожидание (среднее значение)

и дисперсия. В математической статистике аналогичные характеристики вводятся для выборки.

Пусть имеется некоторая выборка объема n :

$$x_1, x_2, \dots, x_n. \quad (1)$$

Выборочным математическим ожиданием, или *выборочным средним*, называют среднее арифметическое значений выборки.

Эту числовую характеристику принято обозначать $\bar{x}$. Таким образом, выборочное среднее для выборки (1) определяется формулой:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad \text{или} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (2)$$

Если выборка задана статистическим рядом или выборочным распределением, то формулу (2) естественно записать в следующем виде:

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_n x_n}{n}, \quad \text{или} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n n_i x_i. \quad (3)$$

В практических приложениях наряду со средним арифметическим также используются следующие средние:

— среднее геометрическое

$$g = \sqrt[n]{x_1 \times x_2 \times \dots \times x_n};$$

— среднее квадратическое

$$\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}};$$

— среднее гармоническое

$$h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}.$$

Выборочной дисперсией называют среднее арифметическое квадратов отклонений значений выборки от выборочного среднего.

Выборочную дисперсию будем обозначать S_0 . Таким образом, выборочная дисперсия для выборки (1) определяется формулой

$$S_0 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n},$$

или

$$S_0 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (4)$$

Если выборка задана статистическим рядом или выборочным распределением, то формулу (4) можно записать в виде:

$$S_0 = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \dots + n_n(x_n - \bar{x})^2}{n},$$

или

$$S_0 = \frac{1}{n} \sum_{i=1}^n n_i (x_i - \bar{x})^2. \quad (4)$$

Формулы (4) и (5) можно преобразовать к более удобному для вычислений виду. Преобразуем формулу (4) (для формулы (5) преобразование проводится аналогично):

$$\begin{aligned} S_0 &= \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + (\bar{x})^2) = \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2(\bar{x}) \left(\frac{1}{n} \sum_{i=1}^n x_i \right) + \frac{1}{n} (\bar{x})^2 n = \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2(\bar{x})^2 + (\bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2. \end{aligned}$$

Обозначим через $\overline{x^2}$ выборочное среднее квадратов значений выборки:

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2.$$

Тогда формула (4) примет вид

$$S_0 = \overline{x^2} - (\bar{x})^2, \quad (6)$$

то есть выборочная дисперсия равна среднему квадратов значений выборки без квадрата выборочного среднего.

В статистике наряду с выборочной дисперсией используется другая важная числовая характеристика выборки: несмещенная выборочная дисперсия.

Несмещенную выборочную дисперсию обозначим S и определим формулой

$$S = \frac{n}{n-1} S_0, \quad (7)$$

где S — выборочная дисперсия, n — объем выборки.

Используя формулу (4), получаем

$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (8)$$

Пример 1. Для выборки

$$1, 2, 0, 4, 5, 3, 7, 7, 3, 0$$

найти выборочное среднее $\bar{x}$, выборочное среднее S_0 , несмещенную выборочную дисперсию S .



Объем выборки $n = 10$. По формуле (2) находим выборочное среднее:

$$\bar{x} = \frac{1+2+0+4+5+3+7+7+3+0}{10} = 3,2.$$

Для вычисления выборочной дисперсии воспользуемся формулой (6). Для этого вычислим среднее квадратов значений выборки:

$$\overline{x^2} = \frac{1^2 + 2^2 + 0^2 + 4^2 + 5^2 + 3^2 + 7^2 + 7^2 + 3^2 + 0^2}{10} = 16,2.$$

Теперь по формуле (6) находим:

$$S_0 = 16,2 - 3,2^2 = 5,96.$$

По формуле (7) вычисляем несмещенную выборочную дисперсию:

$$S = \frac{10}{9} \times 5,96 = 6,62.$$



Пример 2. Для выборки

$$-1, 3, 8, 3, 0, 5, 3, -1, 3, 5$$

найти выборочное среднее $\bar{x}$, выборочное среднее S_0 , несмещенную выборочную дисперсию S .



Запишем выборку в виде статистического ряда:

$$(-1; 2), (0; 1), (3; 4), (5; 2), (8; 1).$$

Объем выборки $n = 10$. Выборочное среднее находим по формуле (3):

$$\bar{x} = \frac{2 \times (-1) + 4 \times 3 + 2 \times 5 + 1 \times 8}{10} = 2,8.$$

Вычислим среднее квадратов значений выборки:

$$\overline{x^2} = \frac{2 \times (-1)^2 + 4 \times 3^2 + 2 \times 5^2 + 1 \times 8^2}{10} = 15,2.$$

По формуле (6) вычислим выборочную дисперсию:

$$S_0 = 15,2 - 2,8^2 = 7,36.$$

Несмещенную выборочную дисперсию вычислим по формуле (7):

$$S = \frac{10}{9} \times 7,36 \approx 8,18.$$

▲

Пример 3. Путешественник, двигаясь из пункта А в пункт Б, первую половину пути проделал на велосипеде со скоростью $V_1 = 6 \frac{\text{км}}{\text{ч}}$, а вторую половину — на автомобиле со скоростью $V_2 = 60 \frac{\text{км}}{\text{ч}}$. С какой средней скоростью он двигался?

▼

По определению, средняя скорость равна отношению пути ко времени, за которое он пройден. Обозначим расстояние АБ через $2S$. Первая половина пути пройдена за время $t_1 = \frac{S}{V_1}$, вторая — за $t_2 = \frac{S}{V_2}$.

Тогда

$$V_{\text{ср}} = \frac{2S}{\frac{S}{V_1} + \frac{S}{V_2}} = \frac{2}{\frac{1}{V_1} + \frac{1}{V_2}} = \frac{2}{\frac{1}{6} + \frac{1}{60}} = \frac{120}{11} \approx 10,9 \left(\frac{\text{км}}{\text{ч}} \right).$$

Таким образом, средняя скорость есть среднее гармоническое, а не среднее арифметическое

$$\frac{V_1 + V_2}{2} = \frac{6 + 60}{2} = 33 \left(\frac{\text{км}}{\text{ч}} \right).$$

▲

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что называют: а) генеральной совокупностью; б) выборочной совокупностью; в) объемом выборки?

2. Дайте определение вариационного ряда. Что называют размахом выборки?

3. Что называют: а) статистическим рядом; б) выборочным распределением?

4. Какие графические изображения выборок вы знаете?

5. Дайте определения выборочных характеристик: а) выборочного среднего; б) выборочной дисперсии.

7.2. СТАТИСТИЧЕСКИЕ ОЦЕНКИ НЕИЗВЕСТНЫХ ПАРАМЕТРОВ

7.2.1. ТОЧЕЧНЫЕ ОЦЕНКИ. НЕСМЕЩЕННОСТЬ И СОСТОЯТЕЛЬНОСТЬ ОЦЕНОК

Одной из основных задач математической статистики является нахождение приближенного значения некоторого неизвестного параметра a случайной величины X по выборке ее значений:

$$x_1, x_2, \dots, x_n, \quad (1)$$

полученной в результате n измерений (наблюдений, опытов). Таким параметром могут быть, например, математическое ожидание случайной величины или ее дисперсия.

Приближенное значение параметра a , вычисленное каким-либо способом по значениям выборки (1), в статистике называют *точечной оценкой* этого параметра и обозначают $\bar{a}_n$.

Замечание. Следует иметь в виду, что значения выборки являются случайными величинами: только от случая зависит, что при k -м измерении получится число x_k , а не какое-нибудь другое значение случайной величины. Поэтому и оценка $\bar{a}_n$ — тоже случайная величина. На самом деле оценка $\bar{a}_n$ представляет собой числовую функцию от результатов n измерений.

Очевидно, что оценка $\bar{a}_n$ не обязательно будет давать удовлетворительное приближение для неизвестного параметра a . Чтобы эта оценка представляла интерес для практики, она должна удовлетворять определенным требованиям, которые обеспечивают ее близость к оцениваемому параметру. Такими требованиями являются несмещенность и состоятельность.

Точечная оценка $\bar{a}_n$ параметра a называется *несмещенной*, если математическое ожидание оценки равно a , то есть если

$$M\bar{a}_n = a. \quad (2)$$

В противном случае, то есть если $M\bar{a}_n \neq a$, оценка называется *смещенной*. Использование несмещенных оценок позволяет избежать систематических ошибок при замене неизвестного параметра a его оценкой $\bar{a}_n$.

При большом количестве измерений требуют также, чтобы с вероятностью, близкой к единице, оценка $\bar{a}_n$ мало отличалась от параметра a .

Оценка $\bar{a}_n$ параметра a называется *состоятельной*, если для любого числа $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|\bar{a}_n - a| < \varepsilon) = 1. \quad (3)$$

Использование состоятельных оценок обеспечивает сближение оценки с оцениваемым параметром при увеличении объема выборки.

Пусть для случайной величины X в результате n измерений получена выборка

$$x_1, x_2, \dots, x_n.$$

В качестве оценки математического ожидания $a = MX$ случайной величины X , как правило, берут выборочное математическое ожидание (выборочное среднее), то есть полагают

$$MX \approx \overline{a_n} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Для оценки дисперсии $a = DX$ используется несмещенная выборочная дисперсия S :

$$DX \approx \overline{a_n} = S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Можно доказать, что каждая из этих оценок является несмещенной и состоятельной.

Заметим, что выборочная дисперсия

$$S_0 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

является смещенной оценкой для DX .

Закон больших чисел в форме Бернулли означает, что частота $\frac{k}{n}$ является состоятельной оценкой для вероятности p .

7.2.2. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

Любая точечная оценка $\overline{a_n}$ неизвестного параметра a является случайной величиной. Принимая оценку $\overline{a_n}$ за значение параметра a , мы, как правило, делаем ошибку даже в том случае, когда оценка является несмещенной и состоятельной. Поэтому важно знать, каковы точность и надежность используемой оценки. Задача состоит в том, чтобы уметь находить границы, в которых с определенной вероятностью заключено неизвестное значение параметра. Эту задачу решают следующим образом.

Пусть по сделанной выборке удалось найти два числа a_1 и a_2 такие, что при любых значениях параметра a выполняется равенство:

$$P(a \in (a_1; a_2)) = \gamma. \quad (1)$$

В этом случае числа a_1 и a_2 называют *доверительными границами*, а интервал $(a_1; a_2)$ — *доверительным интервалом*, соответствующими *доверительной вероятности* γ . Число γ называют также *надежностью оценки*.

Нижняя и верхняя границы a_1 и a_2 доверительного интервала находятся по сделанной выборке и, следовательно, являются случайными величинами. Доверительный интервал $(a_1; a_2)$ поэтому также определяется случайным образом. Если случайное событие $a \in (a_1; a_2)$ происходит, то говорят, что доверительный интервал $(a_1; a_2)$ *накрывает* неизвестный параметр a . Равенство (1) означает, что интервал $(a_1; a_2)$ накрывает неизвестный параметр a с вероятностью γ .

Выбор доверительной вероятности (надежности) определяется конкретными условиями. Обычно берут γ , равной 0,90; 0,95; 0,99. Если конкретные условия позволяют считать все события, вероятность которых меньше соответственно 0,10; 0,05; 0,01, невозможными, то считают практически достоверным, что доверительный интервал накрывает неизвестный параметр, и за его значение принимают середину $\frac{a_1 + a_2}{2}$ достоверного интервала. При этом считают, что ошибка не будет превосходить числа $\frac{a_1 - a_2}{2}$.

В математической статистике разработаны разнообразные методы отыскания доверительных интервалов для неизвестных параметров случайных величин. В качестве примера рассмотрим, как можно найти доверительный интервал для неизвестного параметра p случайной величины, распределенной по биномиальному закону.

Рассмотрим случайную величину X — число наступлений события A в серии из n независимых опытов, в каждом из которых событие A наступает с вероятностью p . Как мы уже знаем, в случае, когда p неизвестно, в качестве оценки вероятности p берется частота, причем, как доказано ранее, эта оценка является несмещенной и состоятельной. Можно доказать, что при большом числе опытов границы a_1 и a_2 доверительного интервала, соответствующего надежности γ , могут быть приближенно найдены по формулам:

$$a_1 \approx \frac{k}{n} - \frac{x}{n} \sqrt{\frac{k(n-k)}{n}}, \quad a_2 \approx \frac{k}{n} + \frac{x}{n} \sqrt{\frac{k(n-k)}{n}}, \quad (2)$$

где k — число наступлений события A в серии из n опытов, а x — корень уравнения

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left\{-\frac{t^2}{2}\right\} dt = \frac{\gamma}{2}.$$

Корень этого уравнения находится по таблице значений функции Лапласа $\Phi_0(x)$. Формулами (2) можно пользоваться при

$$n > 50, \quad k > 5, \quad n - k > 5, \quad (3)$$

то есть во многих практически важных случаях. Если условия (3) не выполняются, для построения доверительного интервала пользуются другими формулами.

Пример. С автоматической линии было отобрано и проверено 400 деталей, 10 деталей оказались бракованными. Найти доверительный интервал, накрывающий с надежностью 0,9 неизвестную вероятность изготовления бракованной детали.



По условию задачи $n = 400$, $k = 10$, $\gamma = 0,9$. По заданной надежности $\gamma = 0,9$ получаем уравнение $\Phi_0(x) = \frac{0,9}{2} = 0,45$. По таблице значений функции Лапласа $\Phi_0(x)$ находим, что $x \approx 1,645$. По формулам (2) находим границы доверительного интервала:

$$a_1 = \frac{10}{400} - \frac{1,645}{400} \sqrt{\frac{10 \times 390}{400}} \approx 0,012,$$

$$a_2 = \frac{10}{400} + \frac{1,645}{400} \sqrt{\frac{10 \times 390}{400}} \approx 0,038.$$

Таким образом, если считать неизвестную вероятность равной 0,025, то с вероятностью 0,9 ошибка не превзойдет 0,013.



КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Дайте определения: а) несмещенной оценки; б) состоятельной оценки.

2. Объясните, что значит, что доверительный интервал $(a_1; a_2)$ накрывает неизвестный параметр a с вероятностью γ .

7.3. ОБРАБОТКА РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

Пусть изучается зависимость между величинами X и Y в случае, когда имеется возможность измерять значения величины Y при различных значениях величины X . Если величина X может принимать лишь

конечное множество значений и если при всех этих значениях возможно точное измерение величины Y , то после конечного числа измерений мы получим таблицу соответствующих значений X и Y . Очевидно, эта таблица описывает искомую функциональную зависимость. Однако даже в случае, когда величина X принимает конечное множество значений, не всегда имеется возможность провести измерения величины Y для всех значений величины X , так как этих значений может быть очень много или процесс измерения связан с уничтожением изучаемого объекта. Кроме того, при одном и том же значении величины X , но при разных измерениях из-за неточности измерения могут получаться разные приближенные значения величины Y . В общем случае приближенные значения величины Y при фиксированном X являются значениями некоторой случайной величины.

Из сказанного следует, что при помощи конечного числа измерений, как правило, нельзя установить точную функциональную зависимость между изучаемыми величинами.

Одной из важнейших задач математической статистики является установление связи между двумя величинами по известным выборкам их значений. В общем случае эта задача формулируется следующим образом.

Пусть для n значений величины X после серии из n измерений получена соответствующая выборка значений величины Y : $y_1, y_2, \dots, y_n$. По этим данным требуется найти приближенный закон функциональной зависимости между величинами X и Y , причем этот приближенный закон должен быть в некотором смысле наилучшим.

Если нет никакой дополнительной информации о виде искомой функциональной зависимости, то поставленная задача является неразрешимой. Поэтому на практике задается вид функциональной зависимости с точностью до некоторых параметров. Кроме того, определяется понятие «наилучшего приближения функциональной зависимости».

Рассмотрим частный случай поставленной задачи, имеющей большое прикладное значение: среди всех линейных функций $y = ax + b$, где a и b — неизвестные параметры, найти ту, которая дает наилучшее приближение функциональной зависимости между величинами X и Y .

Для решения поставленной задачи используется *метод наименьших квадратов*. В этом методе наилучшими значениями для параметров a и b считаются значения, при которых является наименьшей сумма квадратов разностей

$$y_1 - ax_1 - b, \quad y_2 - ax_2 - b, \quad \dots, \quad y_n - ax_n - b,$$

где $x_1, x_2, \dots, x_n$ — выборка значений величины X , а $y_1, y_2, \dots, y_n$ — полученная выборка соответствующих значений величины Y . Другими словами, находят значения $\bar{a}$ и $\bar{b}$ параметров a и b , при которых

$$F(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$$

принимает свое наименьшее значение. Прямая называется *прямой линией регрессии*.

Значения и параметров a и b можно найти из системы уравнений

$$\begin{cases} \overline{x^2} + \bar{x}b = \bar{xy}, \\ \bar{x}a + b = \bar{y}. \end{cases} \quad (1)$$

Здесь введены обозначения:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i, & \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i, \\ \overline{x^2} &= \frac{1}{n} \sum_{i=1}^n x_i^2, & \overline{xy} &= \frac{1}{n} \sum_{i=1}^n x_i y_i. \end{aligned}$$

Замечание 1. Так как выборка, получаемая в результате измерений, является случайной, то и значения $\bar{a}$ и $\bar{b}$, являются случайными. Во многих практически важных случаях, когда между величинами X и Y существует линейная зависимость $Y = aX + b$, значения $\bar{a}$ и $\bar{b}$ являются несмещенными и состоятельными оценками параметров a и b .

Замечание 2. Методом наименьших квадратов могут быть получены оценки параметров для более сложных функциональных зависимостей: квадратичной, гиперболической, показательной и др.

Пример. Результаты пяти измерений некоторой величины Y , зависящей от величины X , приведены в таблице:

x	-2	-1,5	0	1	2
y	1,25	1,40	1,50	1,75	2,25

Найти уравнение прямой линии регрессии.



По результатам измерений вычисляем:

$$\bar{x} = \frac{1}{5}(-2 - 1,5 + 0 + 1 + 2) = -0,1,$$

$$\bar{y} = \frac{1}{5}(1,25 + 1,40 + 1,50 + 1,75 + 2,25) = 1,63,$$

$$\overline{x^2} = \frac{1}{5}((-2)^2 + (-1,5)^2 + 0^2 + 1^2 + 2^2) = 2,25,$$

$$\overline{xy} = \frac{1}{5}(-2 \times 1,25 - 1,5 \times 1,40 + 0 \times 1,50 + 1 \times 1,75 + 2 \times 2,25) = 0,33.$$

Подставив данные значения $\overline{x}$, $\overline{y}$, $\overline{x^2}$, $\overline{xy}$ в систему (1) и решив ее, получим: $a = 0,22$; $b = 1,652$. Следовательно, прямая линия регрессии имеет уравнение

$$y = 0,22x + 1,652.$$



КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что называется прямой линией регрессии?
2. Как составляется нормальная система для определения прямой линии регрессии?
3. Как находятся оценки параметров неизвестной линейной зависимости между величинами методом наименьших квадратов?

УПРАЖНЕНИЯ

1. Дана выборка: 0, -1, 1, 2, 0, 1, 2, 3, 5, 5, 5, 2.
Найти объем выборки, размах выборки; записать вариационный ряд, статистический ряд, выборочное распределение; построить полигон частот.
2. Для выборки, заданной статистическим рядом (10; 3), (20; 3), (50; 4), найдите выборочное среднее $\overline{x}$ и выборочную дисперсию S_0 .
3. Постройте гистограмму частот для выборки:
17, 19, 20, 10, 14, 16, 21, 21, 22, 22,
35, 27, 32, 24, 24, 24, 24, 27, 27, 27.
4. В 10 000 сеансах игры с автоматом выигрыш отмечен в 400 случаях. Укажите несмещенную и состоятельную оценку для вероятности выигрыша при игре с таким автоматом.
- 5 В результате измерений некоторой величины Y , зависящей от температуры X , даны в таблице:

x_i	-5	-4	-3	-2	-1	0	1	2	3	4	5
y_i	0,2	0,3	0,6	1,0	1,8	2,7	3,8	5,1	6,7	8,3	10,2

Найдите прямую линию регрессии.