

Лектор – Лохвицкий Михаил Сергеевич

Математическая статистика. Описательная статистика.

В теории вероятностей предполагается, что все основные характеристики случайного события, случайной величины или случайного процесса известны. На практике это бывает редко, и все характеристики или их часть (функция распределения, плотность распределения вероятностей, моменты случайной величины, функция корреляции случайного процесса) нужно находить (оценивать) из эксперимента. Этот раздел математической статистики (МС) называется ***описательной статистикой***.

Наряду с упомянутыми вопросами часто возникают задачи другого рода. Из тех или иных соображений выдвигаются некоторые гипотезы, например: случайное событие обладает данной вероятностью; математическое ожидание наблюдаемой случайной величины равно нулю; наблюдаемая случайная величина подчиняется нормальному закону; данный процесс является пуассоновским с постоянной интенсивностью и т.д. Проверка такого рода гипотез по наблюдаемым данным составляет содержание специального раздела МС: ***проверка статистических гипотез***.

Часто необходимо выявить меру и характер зависимости двух или нескольких случайных величин (случайных показателей какого-либо объекта). Такого рода задачи решаются в ***корреляционной теории и регрессионной теории***.

При решении многих задач возникает необходимость выявить степень вклада каких-либо факторов в результирующий фактор (или в результирующие факторы). Такие задачи решаются в ***факторном анализе***.

9.1. Описательная статистика

9.1.1. Выборка. Статистическое распределение. Полигон частот.

Совокупность наблюдаемых случайных величин (x_1, x_2, \dots, x_n)

называется **выборкой**, величины x_i ($i = 1, 2, \dots, n$) – **элементами** выборки, а их число n – **объёмом** (или размером) выборки. Конкретные значения выборки, полученные в результате испытаний, называют **реализацией** выборки и обозначают строчными буквами (x_1, x_2, \dots, x_n) . Если элементы выборки расположены в порядке возрастания, то такая последовательность называется **вариационным рядом**, а номер элемента выборки в вариационном ряду называется **рангом элемента**.

Пусть в результате n наблюдений над дискретной случайной величиной X значение x_1 выпало n_1 раз, x_2 – n_2 раз, ..., x_k – n_k раз ($n_1 + n_2 + \dots + n_k = n$). Подсчитаем соответствующие частоты появления этих значений

$$v_1 = \frac{n_1}{n}, \quad v_2 = \frac{n_2}{n}, \quad \dots, \quad v_k = \frac{n_k}{n}.$$

и составим таблицу

x_1	x_2	...	x_i	...	x_k
v_1	v_2	...	v_i	...	v_k

Эта таблица носит название **статистического распределения**.

Статистическое распределение является оценкой неизвестного закона распределения. В соответствии с **теоремой Бернулли** частоты v_i сходятся по вероятности (при $n \rightarrow \infty$) к соответствующим вероятностям p_i ($i = 1, 2, \dots, k$). Поэтому при больших n статистическое распределение мало отличается от истинного распределения.

Графически статистическое распределение изображается в виде **полигона частот**. По оси абсцисс откладываются значения x_1, x_2, \dots, x_k и в точках x_1, x_2, \dots, x_k откладываются в направлении оси OY величины v_1, v_2, \dots, v_k соответственно. Полученные точки соединяют отрезками прямых. Полигон частот является статистическим аналогом многоугольника распределения (напомним, что при построении

многоугольника распределения по оси ординат откладывают вероятности p_i). Заметим, что $\nu_1 + \nu_2 + \dots + \nu_k = 1$.

Пример 9.1. Наблюдения над дискретной случайной величиной X заданы таблицей.

Возможные значения x_i	3	5	7	9	11
Количества полученных наблюдений n_i	4	10	25	8	3

Построить полигон частот.

Решение. Общее число наблюдений

$$n = 4 + 10 + 25 + 8 + 3$$

Подсчитаем частоты:

$$\nu_1 = \frac{n_1}{n} = \frac{4}{50} = 0,08; \quad \nu_2 = \frac{n_2}{n} = \frac{10}{50} = 0,20; \quad \nu_3 = \frac{n_3}{n} = \frac{25}{50} = 0,50;$$

$$\nu_4 = \frac{n_4}{n} = \frac{8}{50} = 0,16; \quad \nu_5 = \frac{n_5}{n} = \frac{3}{50} = 0,06.$$

В точках x_i откладываем частоты ν_i ($i = 1, 2, 3, 4, 5$). Полученные точки соединяем отрезками прямых (рисунок 9.1).

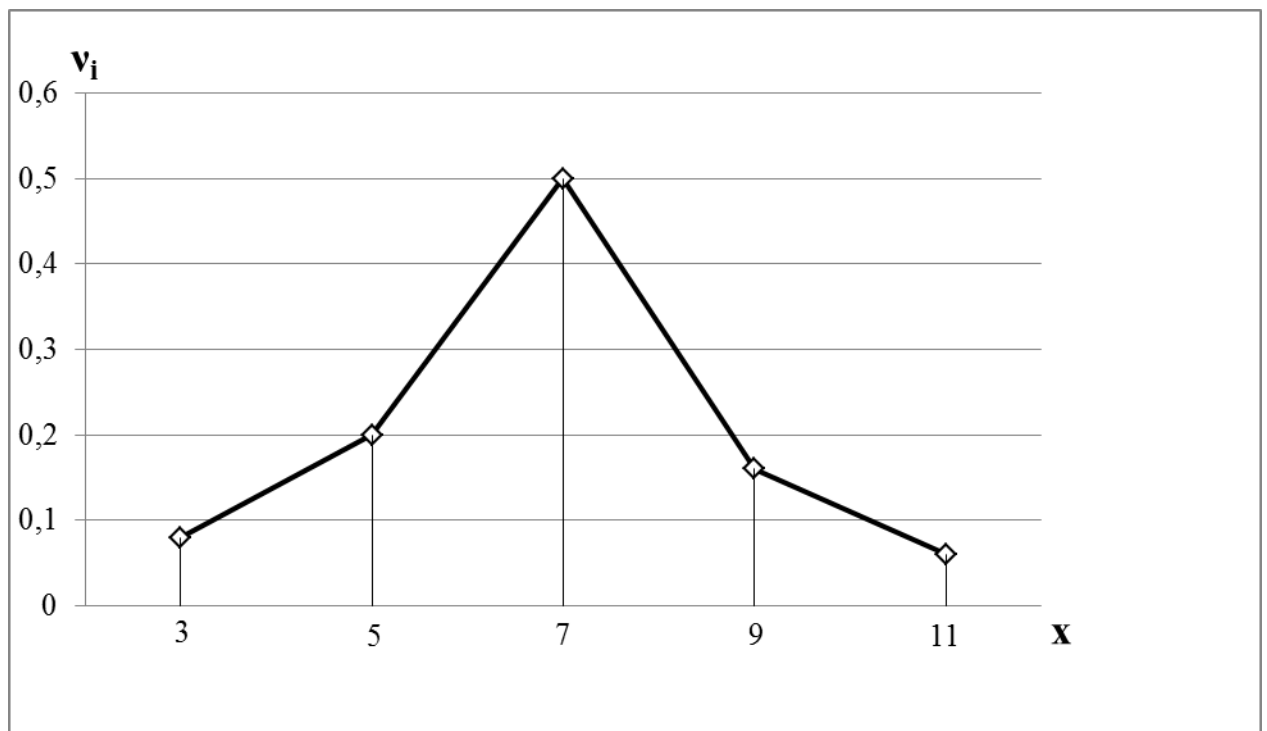


Рисунок 9.1. Полигон частот

9.1.2. Гистограмма.

В теории вероятностей *непрерывная* случайная величина X характеризуется некоторой плотностью распределения вероятностей $p(x)$. Эта плотность, как правило, бывает неизвестной. Чтобы получить о ней некоторое представление, строят *гистограмму*.

Идея построения гистограммы. Гистограмма состоит из k прямоугольников. Прямоугольники выбирают так, чтобы выполнялось свойство плотности распределения: «Вероятность попадания в интервал равна площади под графиком плотности на этом интервале», и для прямоугольника на этом интервале. Вероятность попадания в интервал заменяется её оценкой - относительной частотой попадания в интервал.

Пусть в результате n наблюдений над случайной величиной X мы получили выборочные значения x_1, x_2, \dots, x_k . Область значений случайной величины X разбиваем на k интервалов:

$$(-\infty, \gamma_1), (\gamma_1, \gamma_2), \dots, (\gamma_{i-1}, \gamma_i), \dots, (\gamma_{k-2}, \gamma_{k-1}), (\gamma_{k-1}, \infty).$$

Вычисляем количества наблюдений n_i ($i = 1, 2, \dots, k$), попавших в каждый интервал. При этом получаем числа n_1, n_2, \dots, n_k . По ним подсчитываем частоты $\nu_1, \nu_2, \dots, \nu_k$. Отрезки (γ_{i-1}, γ_i) ($i = 2, 3, \dots, k-1$) во многих случаях будут одинаковой длины, так что $\gamma_i - \gamma_{i-1} = h$. Первый интервал $(-\infty, \gamma_1)$ заменяем отрезком $[\gamma_0, \gamma_1]$ длины h (от точки γ_1 отступаем влево на h и получаем точку γ_0 , а последний интервал (γ_{k-1}, ∞) заменяем отрезком $[\gamma_{k-1}, \gamma_k]$ (от точки γ_{k-1} отступаем вправо на h и получаем точку γ_k). Таким образом, мы получаем k отрезков $[\gamma_0, \gamma_1], [\gamma_1, \gamma_2], \dots, [\gamma_{i-1}, \gamma_i], \dots, [\gamma_{k-1}, \gamma_k]$, длины h . Строим прямоугольники высотой $\delta_i = \frac{\nu_i}{h}$, основаниями которых являются отрезки $[\gamma_{i-1}, \gamma_i]$.

Полученная фигура и называется гистограммой. Гистограмма является статистическим аналогом плотности распределения вероятностей. При

больших n и малых h гистограмма мало отличается от истинной плотности распределения $p(x)$.

На рисунке 9.2 показана гистограмма и истинная плотность $p(x)$.

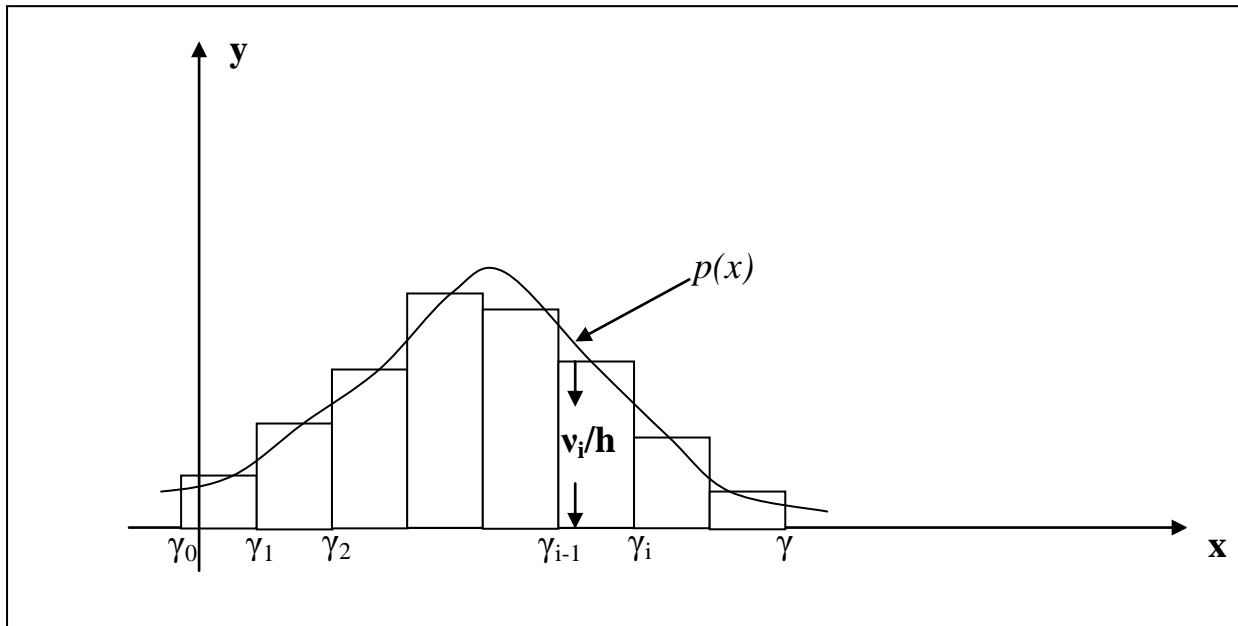


Рисунок 9. 2. Гистограмма и плотность.

Сумма площадей прямоугольников равна:

$$\frac{v_1}{h} h + \frac{v_2}{h} h + \dots + \frac{v_k}{h} h = v_1 + v_2 + \dots + v_k = 1$$

что соответствует условию нормировки $\int_{-\infty}^{\infty} p(x) dx = 1$ для плотности распределения вероятностей $p(x)$.

Пример 9.2. Произведено n наблюдений над непрерывной случайной величиной X . Диапазон изменений величины X разбит на восемь промежутков. Промежутки и количества наблюдений n_i , попавших в каждый из них, указаны в таблице.

$[\gamma_{i-1}, \gamma_i]$	$(-\infty, 3]$	$[3, 5]$	$[5, 7]$	$[7, 9]$	$[9, 11]$	$[11, 13]$	$[13, 15]$	$[15, +\infty)$
----------------------------	----------------	----------	----------	----------	-----------	------------	------------	-----------------

n_i	5	15	38	80	58	28	18	8
-------	---	----	----	----	----	----	----	---

Требуется построить гистограмму.

Решение. Общее число наблюдений n равно

$$n = 5 + 15 + 38 + 80 + 58 + 28 + 18 + 8 = 250.$$

Промежуток $(-\infty; 3]$ заменяем отрезком $[1; 3]$, а промежуток $[15; +\infty)$ - отрезком $[15, 17]$.

Длина отрезков $h = \gamma_i - \gamma_{i-1} = 2$.

Подсчитаем высоты прямоугольников $\delta_i = \frac{v_i}{h} = \frac{n_i}{2n} = \frac{n_i}{500}$:

$$\delta_1 = \frac{5}{500} = 0,010; \quad \delta_2 = \frac{15}{500} = 0,030; \quad \delta_3 = \frac{38}{500} = 0,076; \quad \delta_4 = \frac{80}{500} = 0,160;$$

$$\delta_5 = \frac{58}{500} = 0,116; \quad \delta_6 = \frac{28}{500} = 0,056; \quad \delta_7 = \frac{18}{500} = 0,036; \quad \delta_8 = \frac{8}{500} = 0,016.$$

Строим прямоугольники с высотой δ_i ($i = 1, 2, \dots, 8$), основаниями которых являются заданные отрезки (см. рисунок 9.3). (Проверьте, что сумма площадей прямоугольников равна 1.)

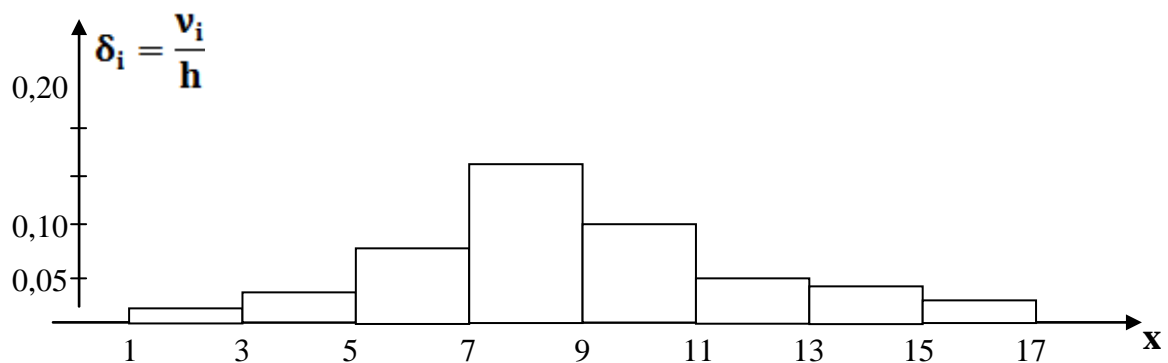


Рисунок 9.3 Гистограмма к примеру 9.2.

9.1.3. Эмпирическая функция распределения.

Определение 9.1.1 Эмпирической функцией распределения $F^*(x)$

называется функция

$$F^*(x) = \frac{n_x}{n}, \quad (9.1)$$

где n_x – число наблюдавшихся значений случайной величины X , меньших числа x ; n – общее число наблюдений.

Очевидно, что $F^*(x)$ удовлетворяет тем же условиям, что и истинная функция распределения $F(x)$, т.е. она является неубывающей функцией и заключена в пределах

$$0 \leq F^*(x) \leq 1.$$

Функция $F^*(x)$ является статистическим аналогом истинной функции распределения $F(x)$.

Функция $F^*(x)$ сходится по вероятности (при $n \rightarrow \infty$) к истинной функции распределения $F(x)$ при каждом x , поскольку

$$F^*(x) = \frac{n_x}{n} = \nu_x,$$

где ν_x – частота события $\{X < x\}$, которая в соответствии с теоремой Бернулли сходится по вероятности (при $n \rightarrow \infty$) к вероятности

$$P(X < x) = F(x).$$

Эмпирическая функция $F^*(x)$ имеет вид ступенчатой функции.

Если наблюдается *непрерывная* случайная величина X и при этом мы получили реализацию вариационного ряда $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, (вариационный ряд – все значения ряда расположены в порядке возрастания), то скачки происходят в точках x_i ($i = 1, 2, \dots, n$), а величина скачков равна $\frac{1}{n}$ (отметим, что сумма всех скачков равна единице). Функция $F^*(x)$ для непрерывной случайной величины X при $n=10$ показана на рисунке 9.4. На этом же рисунке показана функция распределения $F(x)$

Если наблюдается *дискретная* случайная величина и мы получили статистическое распределение

x_1	x_2	...	x_i	...	x_k
ν_1	ν_2	...	ν_i	...	ν_k

то скачки у $F^*(x)$ происходят в точках x_i ($i = 1, 2, \dots, n$), а величины скачков равны ν_i (сумма скачков равна единице).

Пример 9.3. В условиях примера 9.1 построить эмпирическую функцию распределения $F^*(x)$. Частоты ν_i ($i = 1, 2, 3, 4, 5$) были ранее посчитаны

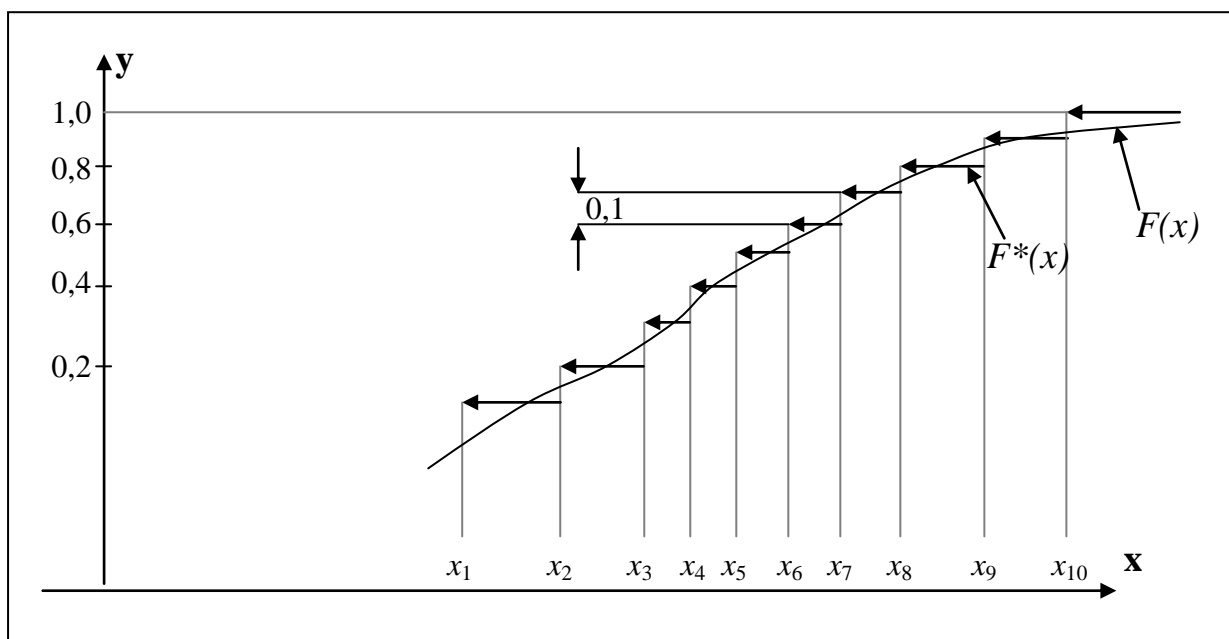


Рисунок 9.4 Эмпирическая функция распределения.

Строим ступенчатую функцию. Скачки величиной ν_i имеют место в точках x_i ($i = 1, 2, 3, 4, 5$).

9.1.4. Точечные оценки характеристик и параметров распределения.

Если нам неизвестны характеристики или параметры распределения, то мы их оцениваем по результатам наблюдений.

Определение 9.1.2. Функция результатов наблюдений (т.е. функция от выборки) называется **статистикой**.

Определение 9.1.3. Оценкой некоторой характеристики или параметра называется статистика, которая в определенном смысле близка к истинному значению этой характеристики или параметра.

Вопрос о том, в каком смысле оценка «близка» к оцениваемой величине допускает различные подходы. Соответственно «хорошая» оценка по одному критерию может не быть таковой по другому.

Определение 9.2.4. Оценка неизвестного математического ожидания MX случайной величины называется **выборочным средним значением** и вычисляется по формуле

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (9.2)$$

Для группированной случайной величины, когда значения x_i появились n_i раз, эту формулу можно упростить как

$$\bar{X} = \frac{1}{n} \sum_{i=1}^k x_i n_i, \quad (9.3)$$

Для оценки дисперсии используют две статистики.

Определение 9.2.5. В качестве оценки дисперсии принята **несмещенная оценка дисперсии** (эту оценку часто называют **исправленной дисперсией**)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2 \quad (9.4)$$

В некоторых случаях используют так называемую **выборочную дисперсию** $D_B = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$, (9.5)

которая всегда дает заниженную оценку реальной дисперсии. Особенно существенно различие при малых n .

Для группированных данных имеем $S^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i - \bar{X})^2 n_i$ и $D_B = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{X})^2 n_i$ соответственно.

При достаточно больших n выборочное среднее \bar{X} и несмещенная дисперсия S^2 мало отличаются от математического ожидания m и дисперсии σ^2 .

Оценка D_B не является несмещённой, а только асимптотически несмещенной. Поэтому и была введена другая оценка для дисперсии – S^2 :

$$S^2 = \frac{n-1}{n} D_B$$

Оценка S^2 является несмещенной оценкой дисперсии.

Аналогично для среднего квадратического отклонения существуют σ_X две оценки. Лучшей оценкой σ_X является корень из несмещенной дисперсии $S = \sqrt{S^2}$. Вторая оценка – выборочное среднее квадратическое отклонение равно по определению корню из выборочной дисперсии:

$$\sigma_B = \sqrt{D_B}. \quad (9.6)$$

Пример 9.4. В условиях примера 9.1 подсчитать выборочное среднее значение, выборочную дисперсию и выборочное среднее квадратическое отклонение.

Решение. Находим

$$\bar{X} = \frac{3 \cdot 4 + 5 \cdot 10 + 7 \cdot 25 + 9 \cdot 8 + 11 \cdot 3}{4 + 10 + 25 + 8 + 3} = \frac{342}{50} = 6,84.$$

Покажем, как вычислить D_B по сгруппированным данным.

Для выборочной дисперсии справедлива вычислительная формула:

$$D_B = \overline{X^2} - (\bar{X})^2 \quad (9.7)$$

Для этого составим таблицу

x_i	3	5	7	9	11
x_i^2	9	25	49	81	121

Вычислим

$$\overline{X^2} = \frac{9 \cdot 4 + 25 \cdot 10 + 49 \cdot 25 + 81 \cdot 8 + 121 \cdot 3}{50} = \frac{2522}{50} = 50,44;$$

$$(\bar{X})^2 = (6,84)^2 = 46,756 \approx 46,79.$$

$$D_B = 50,44 - 46,79 = 3,65.$$

Пример 9.5. Выборы производились по девяти избирательным округам, однородным по составу жителей. По итогам были получены следующие данные по явке избирателей (%):

Округ	1	2	3	4	5	6	7	8	9
Явка избирателей, %	32.4	36.1	28.5	29.6	34.3	49.1	33.4	31.8	35.1

Средняя явка избирателей составила:

$$\bar{X} = \frac{1}{9} (32.4 + 36.1 + 28.5 + 29.6 + 34.3 + 49.1 + 33.4 + 31.8 + 35.1) \approx 34,48$$

Оценки дисперсии вычисляем по двум формулам

$$D_B = \frac{1}{n} \cdot \sum_{i=1}^n X_i^2 - (\bar{X})^2, \quad S^2 = \frac{n}{n-1} \cdot D_B$$

Получаем

$$D_B = \frac{1}{9} (32.4^2 + 36.1^2 + 28.5^2 + 29.6^2 + 34.3^2 + 49.1^2 + 33.4^2 + 31.8^2 + 35.1^2) - 34.48^2 \approx 32.12 \quad \sigma_B \approx 5.67,$$

$$S^2 = \frac{9}{8} \cdot D_B \approx 36.13, \quad S \approx 6.01.$$

Обратите внимание на существенное отличие D_B от S^2 (напомним, что более точное представление о дисперсии даёт именно S^2 , а D_B всегда её занижает.)

Но исходные данные содержат выброс – показатель явки в 6-ом округе не правдоподобно выделяется на фоне всех остальных. При ближайшем рассмотрении вышестоящая избирательная комиссия обнаружила многочисленные нарушения и аннулировала результаты выборов по этому округу. Давайте посмотрим, как это сказалось на общих характеристиках. Количество округов уменьшилось до 8, остальные характеристики также претерпели изменение:

$$\bar{X} = 32,65, \quad D_B = 6,0625, \quad \sigma_B = 2,426, \quad S^2 \approx 6,929, \quad S \approx 2,632.$$

Удаление выброса уменьшило оценки дисперсии в 5-6 раз! Этот пример наглядно показывает, что выбросы в любом случае требуют к себе особого внимания: либо они отражают реальное, но очень редкое событие и требуют отдельного исследования, либо образуют «статистическую грязь» и должны быть удалены до начала анализа. Часто причиной появления загрязняющих данных является ошибка ввода.

Замечание: Статистики \bar{x} , D_B , S^2 , γ_{1B} , γ_{2B} и им подобные очень чувствительны к наличию среди данных выбросов, т.е. ошибочных данных, существенно отличающихся от всех остальных. Поэтому желательно предварительно построить вариационный ряд и удалить из него ошибочные данные.

\bar{X}, S^2, σ_B^2 – это *точечные оценки* соответствующих неизвестных MX, DX и σX .

Без указания степени точности такие оценки мало информативны. Поэтому рассматривают еще и интервальные оценки неизвестных параметров (раздел 7.1.6).

Для оценки связи между наблюдаемыми в эксперименте случайными величинами широко используются методы корреляционного и регрессионного анализа. В случае парных количественных наблюдений $(X_i, Y_i), 1 \leq i \leq n$ применяется коэффициент выборочной корреляции Пирсона

$$\hat{r} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}}, \quad (9.21)$$

который показывает, насколько хорошо зависимость между случайными величинами может быть описана линейной функцией. Для качественной оценки тесноты связи измеряемых величин используют шкалу Чеддока, приведенную в табл. 1.

Таблица 1.

Шкала Чеддока для оценки линейной связи двух случайных величин

Значение модуля коэффициента корреляции r	Теснота связи
0,1 – 0,3	Слабая
0,3 – 0,5	Умеренная
0,5 – 0,7	Заметная
0,7 – 0,9	Высокая
0,9 – 0,99	Весьма высокая

5. Найти выборочный коэффициент корреляции для пары случайных величин:

	Y	0	1	2
X				

0	30	10	10
2	10	20	20

Решение:

	Y	0	1	2	n_x
X					
0		30	10	10	50
2		10	20	20	50
n_y		40	30	30	N=100

Для данной выборки вычислим следующие числовые характеристики:

1) выборочные средние

$$\bar{x} = \frac{\sum_{i=1}^{m_1} x_i \cdot n_i}{N} = \frac{0 \cdot 50 + 2 \cdot 50}{100} = 1$$

$$\bar{y} = \frac{\sum_{j=1}^{m_2} y_j \cdot n_j}{N} = \frac{0 \cdot 40 + 1 \cdot 30 + 2 \cdot 30}{100} = 1$$

где N — объем выборки (в нашем случае $N = 100$).

Несмещенные оценки дисперсий

$$S_x^2 = \frac{\sum_{i=1}^{m_1} n_x (x_i - \bar{x})^2}{N - 1} = \frac{50 \cdot (0 - 1)^2 + 50 \cdot (2 - 1)^2}{100 - 1} = \frac{50 + 50}{99} = 1,01;$$

$$S_y^2 = \frac{\sum_{j=1}^{m_2} n_y (y_j - \bar{y})^2}{N - 1} = \frac{40 \cdot (0 - 1)^2 + 30 \cdot (1 - 1)^2 + 30 \cdot (2 - 1)^2}{100 - 1} = \frac{60}{99} = 0,61.$$

Несмещенные выборочные среднеквадратические отклонения:

$$S_x = \sqrt{1,01} = 1,005; \quad S_y = \sqrt{0,61} = 0,78.$$

$$\sum_{i=1}^N x_i \cdot y_i = \sum_{i=1}^m m_{xy} \cdot x_i \cdot y_i = 0 \cdot (0 \cdot 30 + 2 \cdot 10) + 1 \cdot (0 \cdot 10 + 2 \cdot 20) +$$

$$+ 2 \cdot (0 \cdot 10 + 2 \cdot 20) = 40 + 80 = 120$$

Подставим данные коэффициенты в формулу для вычисления выборочного коэффициента корреляции r_B

$$r_B = \frac{\sum_{i=1}^m m_{xy} \cdot x_i \cdot y_i - N \cdot \bar{x} \cdot \bar{y}}{N \cdot S_x \cdot S_y} = \frac{120 - 100 \cdot 1 \cdot 1}{100 \cdot 1,005 \cdot 0,78} = \frac{20}{78,39} = 0,255$$

Ответ: 0,255