Тема 1. Методы описательной статистики

2.1.1. Эмпирические распределения

Первоначально выборки могут быть представлены в виде таблицы, состоящей из двух строк (табл. 2.1). В первой даны номера измерений, во второй – их результаты.

Таблица 2.1

Простой статистический ряд

I – номер измерений	1	2	…	n
результат измерений	x_1	x_2	…	x_n

Таблица такого вида называется *простым статистическим рядом*. Далее этот ряд преобразуют в *вариационный ряд*, где все наблюдения представляются в порядке возрастания, т. е. в виде

$$x_{1\min}, x_2, \ldots, x_{n\max},$$

где $x_1 \le x_2 \le \ldots \le x_{n\max}$.

На следующем этапе данный вариационный ряд представляется в виде *статистического ряда*. Статистический ряд для дискретной переменной – это сгруппированный (или частотный) ряд следующего вида (табл. 2.2).

Таблица 2.2

Статистический ряд для переменной дискретного типа

Возможные значения переменой x	Частота	Частость (относительная частота)	Накопленная частота	Накопленная частость
x_i	f_i	$\frac{f_i}{n}$	$\sum\limits_{x_i<x} f_i$	$\sum\limits_{x_i<x} \frac{f_i}{n}$
	$\sum f_i = n$	$\sum \frac{f_i}{n} = 1$		

Статистический ряд для непрерывной переменной представляется интервальной таблицей (табл. 2.3).

Таблица 2.3

Интервальная таблица

Класс границ	Частота f_i	Средняя точка класса x_i	Частость (относитель-ные часто-ты)	Накопленные частоты	Накопленные частости
x_i	f_i	$\bar{x}_i$	$\frac{f_i}{n}$	$\sum_{x_i<x} f_i$	$\sum_{x_i<x} \frac{f_i}{n}$
x_i, x_{i+h}	$\sum f_i = n$		$\sum \frac{f_i}{n} = 1$		

Табл. 2.2. и 2.3 дают полный вид статистических рядов (они могут представляться и в усеченном виде, в зависимости от решаемой задачи).

Класс границ для интервального ряда можно изобразить на числовой оси (рис. 2.1).

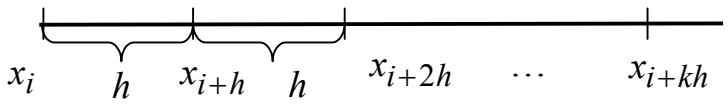


Рис. 2.1. Класс границ для интервального ряда

Значения x_i, x_{i+h}, x_{i+2h}, …, x_{i+kh} определяют границы классов.

Класс границ табл. 2.3 заполняется в столбик в виде $x_i - x_{i+h}$, $x_{i+h} - x_{i+2h}$, …, где каждая пара образует границы k-го интервала, $k = \overline{1,l}$. Значения, стоящие слева, называются *нижними границами классов*, значения, стоящие справа, называются *верхними границами классов.* Верхняя граница предыдущего класса является нижней границей следующего класса.

Рассматривается случай (см. рис. 2.1), когда граничные точки классов отстоят друг от друга на одну и ту же величину, равную шагу h, который равен $h = (x_{\max} - x_{\min})/k$, где k – число классов.

Чтобы определить величину этого шага, необходимо установить, на какое количество классов k разбивается данный ряд наблюдений. Это рекомендуется сделать в соответствии со следующими формулами:

$$k = 1 + 3.322 \lg n \qquad (2.1)$$

или

$$k \le 5 \lg n, \qquad (2.2)$$

где n – число наблюдений (объем выборки).

В литературе предлагается также выбирать число классов в зави-

симости от объема выборки. Для малых выборок $k = 5 \div 7$, для больших – $k = 10 \div 20$. Выбор числа классов является важным моментом. При слишком малом k гистограмма (см. ниже) не будет отражать особенностей распределения, при слишком большом k гистограмма будет излишне изрезанной. Значения h и k обычно округляются до ближайшего целого.

За начало первого интервала берется точка $x_{\min}$ или $x_{\min} - h/2$.

Средняя точка x_i определяется в виде $\bar{x}_{i_k} = x_i + h/2$, где x_i – нижняя граница соответствующего класса.

Частота f_i представляет собой количество наблюдений, соответствующих данному наблюдению для дискретной переменной для сгруппированного ряда, или число наблюдений, попавших в данный интервал для интервального ряда.

Значение соответствующей частоты, деленной на объем выборки, характеризует *частость* попадания x_i в частичные интервалы.

Закон больших чисел в форме Бернулли утверждает, что если эксперимент повторяется n раз при одинаковых условиях, то частость f_i/n сходится по вероятности к p_i, т. е. $\frac{f_i}{n} \xrightarrow{p} p_i$. Следовательно, значения f_i/n являются приближенными значениями вероятностей p_i.

В отличие от теоретических законов распределения для случайных величин, рассмотренных в «Теории вероятностей», для выборки определяется *эмпирический закон распределения*, или *эмпирическое распределение частот*.

Для наглядного представления эмпирических распределений для переменной дискретного типа (см. табл. 2.2) строится график, где по оси X откладываются значения переменной, а по оси Y – значения частот (частостей). Полученные точки соединяют ломаной линией. Этот график называется *полигоном*.

Интервальный ряд графически представляется в виде *гистограммы*. Гистограмма представляет собой ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы $[x_i, x_{i+kh}]$, а их высоты равны f_i/n или f_i, h – длина интервала, или шаг. В первом случае площадь гистограммы равна единице, во втором случае – объему выборки n.

Результаты в столбце «накопленные частости» используются для определения *эмпирической функции распределения*, которая также используется для задания эмпирического распределения.

Эмпирическая функция распределения определяется по значениям

накопленных частостей из следующего соотношения:

$$F_n^*(x) = 1/n \sum_{x_i < x} f_i \,, \qquad (2.3)$$

где суммируются частости тех элементов выборки, для которых выполняется неравенство $x_i < x$ (x – некоторое значение). Из приведенных формул следует, что

$$F_n^*(x) = 0, \quad \text{если } x \le x_1;$$

$$F_n^*(x) = 1, \quad \text{если } x > x_n;$$

$$F_n^*(-\infty) = 0, \quad F_n^*(+\infty) = 1.$$

На промежутке (x_1, x_{i+kh}) $F_n^*(x)$ представляет собой неубывающую кусочно-постоянную функцию.

Согласно теореме Гливенко эмпирическая функция распределения $F_n^*(x)$ является хорошей оценкой генеральной функции распределения $F(x)$ при $n \to \infty$.

Вариационный ряд является статистическим аналогом (реализацией) распределения признака (случайной величины X).

В этом смысле полигон и гистограмма аналогичны кривой распределения, а эмпирическая функция распределения $F_n^*(x)$ – функции распределения $F(x)$ случайной величины X.

График эмпирической функции распределения $F_n^*(x)$ представляется для дискретной случайной величины X в виде неубывающей ступенчатой функции вида, представленного на рис. 2.3. Скачки графика функции $F_n^*(x)$ имеют место в тех точках, которым соответствуют наблюдаемые значения вариантов, при этом величина скачка равна частости варианта. Значения функции $F_n^*(x)$ находятся в интервале [0,1].

Пример. В выборке из 30-ти семей указано число членов в каждой семье.

2	3	1	2	6	4	2	1	5	3	2	3	1	2	2
1	3	1	2	2	4	2	1	2	8	3	2	1	1	3

Построить статистический ряд и полигон.

Построить график эмпирической функции распределения.

Решение. Данная выборка представляет собой наблюдения дискретного типа, поэтому статистический ряд является сгруппированным рядом (табл. 2.4).

Прежде чем заполнять табл. 2.4, представим исходную выборку в виде вариационного ряда

1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 4 4 5 6 8

Таблица 2.4

Статистический ряд для выборки из 30-ти семей

Возможные значения x	Частота f_i	Частость f_i/n	Накопленная частота	Накопленная частость
1	8	8/30	8	8/30
2	11	11/30	19	19/30
3	6	6/30	25	25/30
4	2	2/30	27	27/30
5	1	1/30	28	28/30
6	1	1/30	29	29/30
8	1	1/30	30	30/30

$\sum f_i = 30;$ $\sum f_i/n = 1.$

На рис. 2.2 представлен многоугольник распределения.

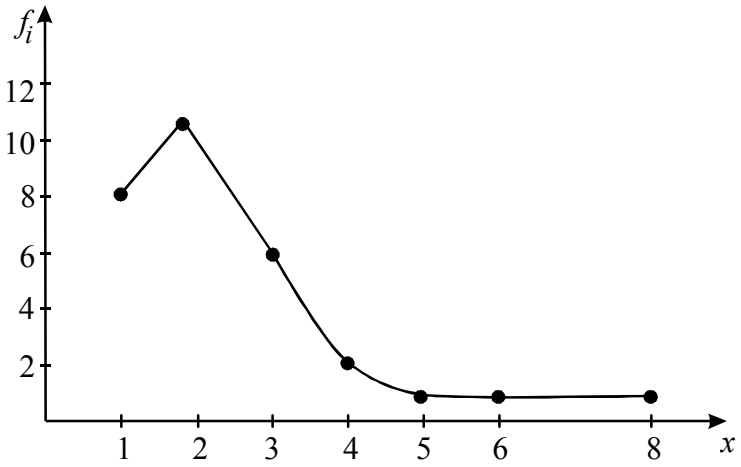


Рис. 2.2. Полигон числа членов семьи для выборки из 30-ти семей

На рис. 2.3 представлен график эмпирической функции распределения.

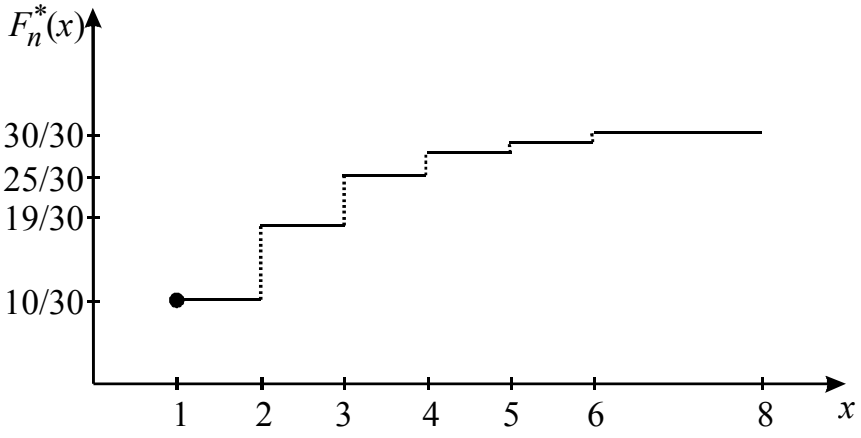


Рис. 2.3. График эмпирической функции распределения числа членов семьи для выборки (из 30-ти семей)

Пример. Выборка представляет собой перечень цен на обезболивающее лекарство, которое продается в разных аптеках города:

12.19 10.09 10.09 13.09 13.45 7.89 12.00 10.49 15.30 13.29

Построить статистический ряд, гистограмму и эмпирическую функцию распределения.

Решение. Данная выборка представляет собой наблюдения непрерывного типа, поэтому статистический ряд представляется интервальной таблицей (см. табл. 2.3).

Вариационный ряд имеет следующий вид:

7.89 10.09 10.09 10.49 12.00 12.19 13.09 13.29 13.45 15.30

Для заполнения столбца «Класс границ» табл. 2.5 необходимо определить число классов K и ширину класса h:

$$k = 1 + 3.322 \lg 10 = 4.322;$$

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{15.30 - 7.89}{4.322} = \frac{7.41}{4.322} = 1.7 \approx 2.$$

Статистический ряд представляет собой следующую интервальную таблицу (табл. 2.5).

Таблица 2.5

Интервальная таблица для переменной *X*

Класс границ	Частота f_i	Средн. точка класса	Частость f_i/n	Накопленная частота	Накоп-ленная частость
7.89 – 9.89	1	8.89	1/10	1	1/10
9.89 – 11.89	3	10.89	3/10	4	4/10
11.89 – 13.89	5	12.89	5/10	9	9/10
13.89 – 15.89	1	14.89	1/10	10	10/10

$\sum f_i = 10$ $\qquad$ $\sum f_i / n = 1$

На рис. 2.4 представлена гистограмма. По оси X отложены границы классов, по оси $У$ – частоты.

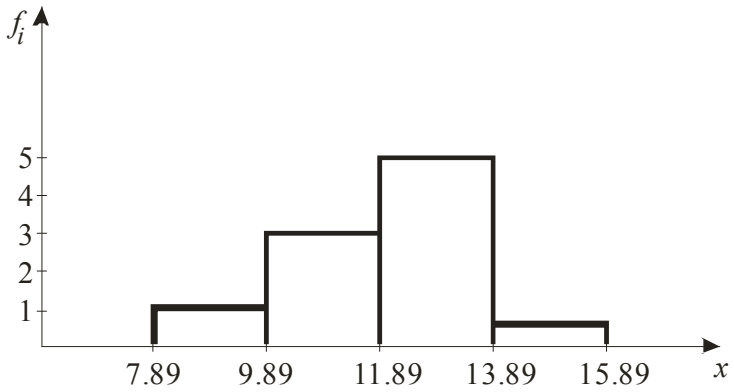


Рис. 2.4. Гистограмма,
построенная для интервальной табл. 2.5

На рис. 2.5 представлен график эмпирической функции распределения, построенный для интервальной табл. 2.5.

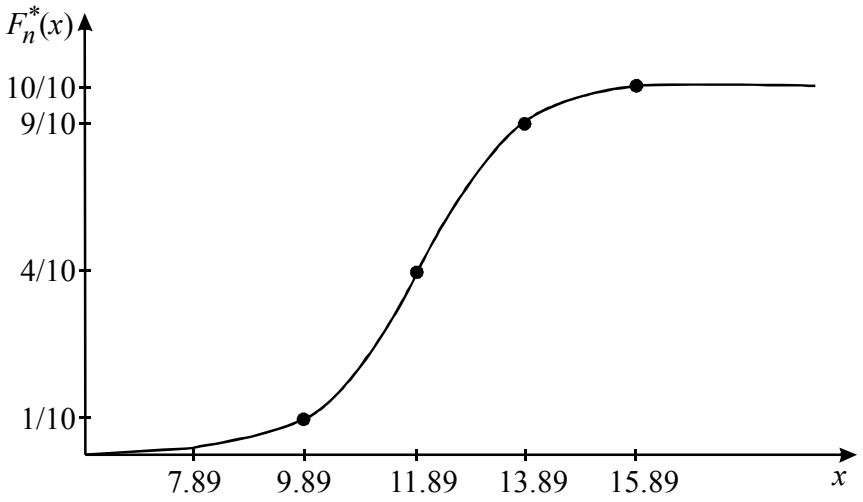


Рис. 2.5. График эмпирической функции распределения, построенный для интервальной табл. 2.5

Для построения графика по оси *X* откладываются границы интервалов, по оси *У* – значения из столбца «Накопленная частость» интервальной таблицы, соответствующие верхним границам интервалов. Полученные точки соединяют плавной линией.

График эмпирической функции распределения называется *кумулятой*. Если оси поменять местами, то полученный график называется *огивой*.

2.1.2. Числовые характеристики

На практике часто оказывается достаточным знание лишь характеристик, например, центральной тенденции для вариационного ряда, характеристик изменчивости и др. Вычисление этих характеристик представляет собой этап обработки данных наблюдений. Поскольку эти характеристики вычисляются по данным, полученным в результате наблюдений (статистическим данным), то их называют *выборочными числовыми характеристиками*, *или статистическими характеристиками, или оценками*.

Некоторые из них характеризуются тем, что вокруг них концентрируются остальные наблюдения. Такие числовые характеристики называются *характеристиками расположения* и к ним относятся такие, как *среднее арифметическое, мода, медиана, процентили, квартили*.

Для оценки изменчивости служат показатели вариации. К ним относятся такие характеристики, как *дисперсия, среднеквадратическое отклонение* и др., которые называются *характеристиками рассеивания*.

Числовые характеристики вводятся через выборочные моменты, которые являются определенными числовыми значениями. Моменты бывают различных порядков: 1-го, 2-го и более. На практике не используются моменты выше 4-го порядка.

2.1.3. Первый выборочный момент

Первым выборочным моментом является

$$\overline{X} = m_1 = \frac{1}{n}\sum_{i=1}^{k} x_i f_i \,, \tag{2.4}$$

называемый *средним арифметическим*. Если в качестве исходных данных имеем интервальную таблицу, то в этом случае k – число классов, x_i – средняя точка класса, f_i – частоты.

Числовая величина 1-го выборочного момента, или *среднее арифметическое* $\overline{X}$, *характеризует точку равновесия оси x гистограммы*.

Если имеем выборку $x_1, x_2, \ldots, x_n$, то среднее арифметическое определяется по формуле

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{2.5}$$

Среднее значение, полученное по генеральной совокупности, называется *математическим ожиданием*, или *генеральным средним*. Генеральное среднее называется *параметром* и является постоянной величиной. Среднее арифметическое $\overline{X}$, определяемое по выборке, является оценкой математического ожидания, $\overline{X}$ – случайная величина.

Пример. Определить среднее арифметическое для исходной выборки (см. с. 72) и интервальной табл. 2.5 (см. с. 73), представляющих собой перечень цен на обезболивающее лекарство.

Решение. Для выборки

$$\overline{X} = \frac{1}{10}(12.19 + 10.09 + 10.09 + 13.09 + 13.45 + 7.89 +$$
$$+ 12.00 + 10.49 + 15.30 + 13.29) = 11.788.$$

Для интервальной таблицы

$$\overline{X} = \frac{1}{10}(8.89 \cdot 1 + 10.89 \cdot 3 + 12.89 \cdot 5 + 14.89 \cdot 1) = 12.09 \,.$$

2.1.4. Основные свойства средней арифметической

1. *Среднее арифметическое постоянной равно самой постоянной*

$$\overline{X} = \frac{1}{n}\sum C = \frac{nc}{n} = C\,.$$

2. *Если все значения x_i увеличить* (*уменьшить*) *в одно и то же число раз, то средняя арифметическая увеличится* (*уменьшится*) *во столько же раз:*

$$\overline{kx} = k\overline{x}\,, \quad \text{или} \quad \frac{\sum kx_i}{n} = k\frac{\sum x_i}{n}\,.$$

3. *Если все значения x_i увеличить* (*уменьшить*) *на одно и то же число, то средняя арифметическая увеличится* (*уменьшится*) *на то же число*.

$$\overline{X + C} = \overline{X} + C\,, \quad \text{или} \quad \frac{\sum (x_i + C)}{n} = \frac{\sum x_i}{n} + \frac{nc}{n} = \overline{X} + C\,.$$

4. *Средняя арифметическая отклонений значения x_i от средней арифметической равна нулю*:

$$\overline{X - \overline{X}} = 0\,, \quad \text{или} \quad \frac{\sum x_i - \overline{X}}{n} = \frac{\sum x_i}{n} - \frac{n\overline{X}}{n} = \overline{X} - \overline{X} = 0\,.$$

5. *Если ряд состоит из нескольких групп, общая средняя равна средней арифметической групповых средних, причем весами являются объемы групп:*

$$\overline{X} = \frac{\sum_{i=1}^{l} \overline{X}_i n_i}{n}\,,$$

где $\overline{X}$ – общая средняя всего ряда;

$\overline{X}_i$ – групповая средняя *i*-й группы объема n_i;

l – число групп.

2.1.5. Мода и медиана

К характеристикам расположения относятся также мода и медиана, которые называют еще *структурными средними*.

Мода M_0 – наиболее часто встречающееся значение в вариационном ряду. Для интервальной таблицы мода определяется в виде

$$M_0 = x_k + \frac{f_k - f_{k-1}}{2f_k - (f_{k-1} + f_{k+1})} \cdot h\,, \qquad (2.6)$$

где входящие в формулу величины определяются из фрагмента гистограммы, представляющей собой интервал с наибольшей частотой и два соседних с ним интервала (рис. 2.6),

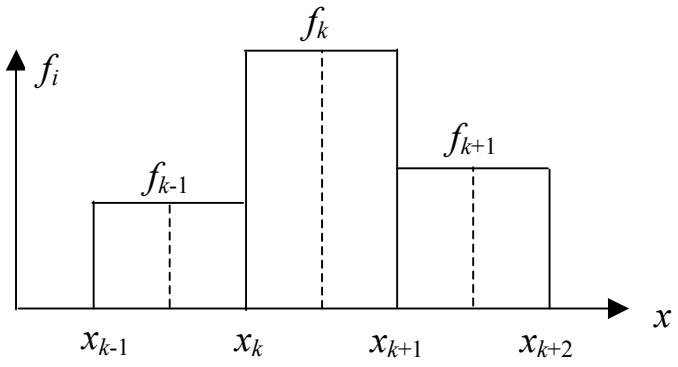


Рис. 2.6

где x_k – нижняя граница интервала с наибольшей частотой;

f_k – частота указанного выше интервала;

f_{k-1} – частота интервала, находящегося слева от интервала с наибольшей частотой;

x_{k-1} – нижняя граница указанного выше интервала;

x_{k+1} – нижняя граница интервала, находящегося справа от интервала с наибольшей частотой;

f_{k+1} – частота указанного выше интервала;

h – шаг.

Мода – мера центральной тенденции, и *она полезна, когда представляет значительную долю своей популяции*.

Медиана – это серединная точка в вариационном ряду, она делит вариационный ряд на две равные по числу членов части.

Для вариационного ряда медиана определяется в зависимости от того, является ли объем выборки n числом четным или нечетным.

Для n – четного $$M_e = \frac{x_{n/2} + x_{n/2+1}}{2}; \qquad (2.7)$$

для n – нечетного $$M_e = x_{(n+1)/2}. \qquad (2.8)$$

Пример

– для вариационного ряда 5 7 11 13 15 17

$$M_e = \frac{x_{6/2} + x_{6/2+1}}{2} = \frac{x_3 + x_4}{2} = \frac{11+13}{2} = 12;$$

– для вариационного ряда 15 17 19 20 25 27 30

$$M_e = x_{(7+1)/2} = x_4 = 20.$$

Для интервальной таблицы

$$M_e = x_{k+1} + \frac{n/2 - S_k}{f_{k+1}} \cdot h, \qquad (2.9)$$

где S_k такое значение накопленных частот, что $S_k \le n/2$ и $S_{k+1} > n/2$,

где n – объем выборки;

f_k – значение частоты для интервала с $S_{k+1} > n/2$;

x_k – нижняя граница интервала с $S_{k+1} > n/2$;

h – шаг.

Медиана обладает следующим свойством. *Сумма абсолютных величин отклонений значения признака от медианы меньше, чем от любой другой величины*, т. е. $\sum_{i=1}^{n} |x_i - c|$ достигнет минимума, если $c = M_e$.

Пример. Определить моду и медиану для интервальной таблицы, представляющей собой перечень цен на обезболивающее лекарство (см. табл. 2.5, с. 73).

Решение. Согласно формуле (2.6)

$$M_0 = 11.89 + \frac{5-3}{10-(3+1)} \cdot 2 = 12.56.$$

Согласно формуле (2.9)

$$M_e = 11.89 + \frac{5-4}{9} \cdot 2 = 12.11.$$

Примечание. При выборе числовых характеристик центральной тенденции следует помнить, что на среднее арифметическое оказывают влияние все члены вариационного ряда, в то время как медиана не подвержена этому влиянию.

Выборочные моменты более высоких порядков, чем 1-й вводятся следующим образом.

Выборочный момент 2-го порядка

$$m_2 = \frac{1}{n} \sum_{i=1}^{k} f_i x_i^2. \qquad (2.10)$$

Выборочный момент 3-го порядка

$$m_3 = \frac{1}{n} \sum_{i=1}^{k} f_i x_i^3. \qquad (2.11)$$

Выборочный момент 4-го порядка

$$m_4 = \frac{1}{n}\sum_{i=1}^{k} f_i x_i^4 , \qquad (2.12)$$

где x_i – средняя точка класса интервальной таблицы;

f_i – частоты;

k – число классов;

n – объем выборки.

2.1.6. Процентили

Процентили являются характеристиками, которые делят ряд наблюдений на 100 частей. Для этого требуется 99 процентилей. Процентиль характеризует значение, достигаемое заданным процентом общего количества данных в вариационном ряду.

P-ый процентиль – это такая величина, что P % данных являются меньше этой величины и (100 – P) % являются больше.

Процентили широко используются в различного рода отчетах. Для того чтобы определить P-й процентиль, необходимо выполнить следующее:

- представить результаты наблюдений в виде вариационного ряда;
- вычислить номер P-го процентиля в вариационном ряду

$$i = \frac{P}{100}(n), \qquad (2.13)$$

где P – значения процентиля;

n – объем выборки;

i – номер процентиля в ряду наблюдений;

- определить значение P-го процентиля:

а) если i – целое, то P-й процентиль является средней величиной i-го и (i+1)-го наблюдений в вариационном ряду;

б) если i не является целым числом, то номер P-го процентиля определяется как целая часть от значения (i+1).

Пример. Определить 30-й процентиль для следующего ряда наблюдений:

14 12 19 23 5 13 28 17.

Решение. Вариационный ряд

5 12 13 14 17 19 23 28;

$$i = \frac{30}{100} \cdot 8 = 2.4 .$$

Так как i не является целым числом, то номер 30-го процентиля в данном вариационном ряду определяется как целая часть от значения 2.4+1=3.4, т. е. 3. Следовательно 30-м процентилем является значение $x_3 = 13$.

Квартили – это характеристики, которые делят ряд наблюдений на 4 части. Для этого необходимы 3 квартиля, которые обозначаются Q_1, Q_2, Q_3.

Q_1 является 25-м процентилем, т. е. $Q_1 = P_{25}$.

Q_2 является 50-м процентилем, т. е. $Q_2 = P_{50}$, или медианой.

Q_3 – это 75-й процентиль, т. е. $Q_3 = P_{75}$.

Пример. Определить Q_1, Q_2, Q_3 для следующей выборки:

109 121 122 129 106 116 125 114.

Решение. Вариационный ряд

106 109 114 116 121 122 125 129.

Так как $Q_1 = P_{25}$, то определяем 25-й процентиль. Для $n = 8$

$$i = \frac{25}{100} \cdot 8 = 2 .$$

Так как i – целое число, то P_{25} определяется как среднее второго и третьего значений в вариационном ряду:

$$P_{25} = \frac{109 + 114}{2} = 111.5 ;$$

$Q_1 = 111.5$, $Q_2 = P_{50}$ и является медианой. Так как n – четное, то

$$Q_2 = \frac{116 + 121}{2} = 118.5 ; \quad Q_3 = P_{75} ;$$

$$i = \frac{75}{100} \cdot 8 = 6 , \; P_{75} = \frac{122 + 125}{2} = 123.5 , \; Q_3 = 123.5.$$

Пять базовых показателей включают наименьшее значение $x_{\min}$, нижний квартиль Q_1, медиану Q_2, верхний квартиль Q_3 и наибольшее значение $x_{\max}$. Вместе эти характеристики дают достаточно ясное представление об особенностях еще не обработанного набора данных (см. с. 90).

2.1.7. Характеристики рассеивания

Второй выборочный момент (*2.11*) *используется для описания рассеивания данных относительно среднего арифметического* $\overline{X}$. Для этого вводится

$$\overset{\circ}{x}_i = x_i - \overline{X}, \tag{2.14}$$

которая называется *центрированной величиной*, тогда формула (2.11) имеет вид

$$m_2 = \frac{1}{n}\sum_{i=1}^{k} f_i \dot{x}_i^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{X})^2 f_i, \tag{2.15}$$

где $\overline{X}$ – среднее арифметическое, полученное для интервальной таблицы.

Из приведенной формулы видно, что второй выборочный момент характеризует разброс наблюдений относительно среднего арифметического и называется *дисперсией* S^2.

Эта формула может быть выражена через m_1 и m_2 (см. 2.14 и 2.15):

$$\begin{aligned} S^2 &= 1/n \sum_{i=1}^{k}(x_i - \overline{X})^2 f_i = 1/n\left(\sum_{i=1}^{k}(x_i^2 - 2x_i\overline{X} + \overline{X}^2) f_i\right) = \\ &= 1/n \sum_{i=1}^{k} x_i^2 f_i - 1/n \sum_{i=1}^{k} 2x_i\,\overline{X} f_i + 1/n \sum_{i=1}^{k} \overline{X}^2 f_i = \\ &= m_2 - 2m_1^2 + m_1^2 = m_2 - m_1^2. \end{aligned} \tag{2.16}$$

Выборочные моменты являются оценками генеральных или теоретических моментов, т. е. дисперсия S^2 является оценкой генеральной дисперсии σ^2. Чтобы эти оценки были надежными, к ним предъявляются требования состоятельности, несмещенности и эффективности. Подробнее о свойствах оценок речь пойдет в разделе «Понятие оценки параметров» (см. п. 2.2.2, с. 94).

Чтобы дисперсия S^2 являлась несмещенной оценкой генеральной дисперсии, ее значение домножается на множитель $\frac{n}{n-1}$, в результате имеем

$$S^2 = \frac{1}{n}\cdot\frac{n}{n-1}\sum_{i=1}^{k}(x_i - \overline{X})^2 f_i = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \overline{X}) f_i. \tag{2.17}$$

S^2 называется *выборочной дисперсией*. Множитель $\frac{1}{n-1}$ особенно важен для выборок малого объема.

Другой характеристикой рассеивания является среднеквадратическое отклонение S, которое является корнем квадратным из дисперсии:

$$S = \sqrt{S^2}\,. \tag{2.18}$$

Поскольку дисперсия измеряется в квадратах наименований исходных наблюдений (кг^2, см^2 и т. д.), то удобнее для интерпретаций результатов использовать среднеквадратическое отклонение, которое измеряется в тех же единицах, что и исходные данные.

2.1.8. Формулы для определения дисперсии

Если исходные данные представляют собой выборку $x_1, x_2, \ldots, x_n$, то формула для определения выборочной дисперсии имеет вид

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2\,, \tag{2.19}$$

где x_i – значения наблюдений;

$\overline{X}$ – среднее арифметическое для выборки.

Для интервальной таблицы

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \overline{X})^2 f_i\,, \tag{2.20}$$

где $\overline{X}$ – среднее арифметическое для интервального ряда;

f_i – частоты;

x_i – объем выборки.

На практике удобно использовать вычислительные формулы для дисперсии и стандартного отклонения или среднеквадратического отклонения.

Вычислительные формулы для исходной выборки $x_1, x_2, \ldots, x_n$ для определения выборочной дисперсии и среднеквадратического отклонения:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}; \qquad S = \sqrt{S^2}\,. \tag{2.21}$$

Если известно среднее арифметическое, то формулы для выборочной дисперсии и среднеквадратического отклонения имеют вид

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n(\overline{X})^2}{n-1}; \qquad S = \sqrt{S^2}\,. \tag{2.22}$$

Вычислительные формулы для интервальной таблицы для выборочной дисперсии и среднеквадратического отклонения имеют вид

$$S^2 = \frac{\sum_{i=1}^{k} f_i x_i - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1}; \qquad S = \sqrt{S^2}, \qquad (2.23)$$

где x_i – средние точки классов;

k – число классов;

f_i – частоты;

n – объем выборки.

2.1.9. Основные свойства дисперсии

1. *Дисперсия постоянной равна нулю.*

2. *Если все значения x_i увеличить (уменьшить) в одно и то же число k раз, то дисперсия увеличится (уменьшится) в k^2 раз, т. е. если наблюдения $x_1,\ldots,x_n$ увеличить в k раз, то получим выборку вида $kx_1,\ldots,kx_n$, тогда*

$$S^2_{kx} = k^2 S^2_x, \quad \text{или} \quad \frac{\sum (kx_i - k\bar{X})^2}{n-1} = k^2 \frac{\sum (x_i - \bar{X})^2}{n-1}.$$

3. *Если все значения x_i увеличить (уменьшить) на одно и то же число раз, то дисперсия не изменится:*

$$S^2_{x+c} = S^2_x = S^2, \quad \text{или} \quad \frac{\sum [(x_i + c) - (\bar{X} + c)]^2}{n-1} = \frac{\sum (x_i - \bar{X})^2}{n-1}.$$

4. *Дисперсия равна разности между средней арифметической квадратов значений x_i и квадратом средней арифметической:*

$$S^2 = \bar{X}^2 - (\bar{X})^2; \qquad \bar{X}^2 = \frac{\sum x_i^2}{n};$$

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

$$= \bar{X}^2 - 2\bar{X} \cdot \bar{X} + \bar{X}^2 = \bar{X}^2 - (\bar{X})^2.$$

5. *Свойство сложения дисперсий.*

Если ряд состоит из нескольких групп наблюдений, то общая дисперсия равна сумме средней арифметической групповых дисперсий и межгрупповой дисперсии:

$$S^2 = \bar{S}_i^2 + \delta^2, \qquad (2.24)$$

где S^2 – общая дисперсия всего ряда;

$\bar{S}_i^2 = \dfrac{\sum_{i=1}^{l} S_i^2 n_i}{n}$ – средняя арифметическая групповых дисперсий,

где $\bar{S}_i^2$ – дисперсия в i-й группе объема n_i;

l – число групп, n – объем выборки;

$\delta^2 = \dfrac{\sum(\bar{x}_i - \bar{X})^2}{n}$ – межотраслевая дисперсия,

где $\bar{X}$ – общая средняя (см. свойство средней арифметической).

Это свойство известно в статистике как «правило сложения дисперсий» и имеет важное значение в статистическом анализе.

Пример. Имеются следующие данные о средних дисперсиях заработной платы двух групп рабочих, представленных в табл. 2.6.

Таблица 2.6

Данные о средних дисперсиях заработной платы двух групп рабочих

Группа рабочих	Число рабочих	Средняя заработная плата одного рабочего в группе, р.	Дисперсия заработной платы
Работающие на одном станке	40	2400	180000
Работающие на двух станках	60	3200	200000
	Всего 100	$\delta^2 = 153600$	$\bar{S}^2 = 192000$

Найти общую дисперсию распределения числа рабочих по заработной плате.

Решение. Согласно свойству среднего арифметического общая средняя определяется в виде

$$\bar{X} = \frac{2400 \cdot 40 + 3200 \cdot 60}{100} = 2880 \text{ р.}$$

Дисперсия по свойству сложения дисперсий определяется в виде

$$S^2 = \bar{S}_i^2 + \delta^2.$$

Средняя групповых дисперсий $\bar{S}_i^2$ определяется в виде

$$\overline{S}_i^2 = \frac{180000 \cdot 40 + 200000 \cdot 60}{100} = 192000.$$

Межгрупповая дисперсия

$$\delta^2 = \frac{(2400 - 2880)^2 \cdot 40 + (3200 - 2880)^2 \cdot 60}{100} = 153600.$$

Используя правило сложения дисперсий, имеем

$$S^2 = 192000 + 153600 = 345600.$$

2.1.10. Коэффициент вариации

Коэффициент вариации – это отношение стандартного отклонения S к среднему значению, выраженному в процентах:

$$V = \frac{S}{\overline{X}} \cdot 100. \qquad (2.25)$$

Коэффициент вариации и среднеквадратическое отклонение могут использоваться как меры риска, например, при финансовых операциях.

Коэффициент вариации может быть использован при сравнении стандартных отклонений, которые вычислены по данным, имеющим различные средние.

Пример. Предположим, что цены на ценные бумаги широко колеблются. Инвестор, который покупает акции по низкой цене, а продает по высокой, имеет хороший доход. Однако если цены на акции падают ниже стоимости, по которой инвестор купил, то он теряет доход.

Чтобы оценить меру риска, инвестор может использовать коэффициент вариации и среднеквадратическое отклонение.

Какую информацию о степени риска может дать коэффициент вариации по сравнению со среднеквадратическим отклонением?

Допустим, за пять недель цены:

на акции 1 представлялись в виде \$57, 68, 64, 71, 62;

на акции 2 представлялись в виде \$12, 17, 8, 15, 13.

Средняя цена на акции 1 $\overline{X} = \$64.40$ и $S = \$4.84$.

Средняя цена на акции 2 $\overline{X} = \$13.00$ и $S = \$3.03$.

Со среднеквадратическим отклонением как мерой риска акции 1 более рискованные. Однако среднее арифметическое акций 1 почти в 5 раз больше среднего арифметического акций 2. Коэффициент вариации, используемый в данном случае, дает следующие результаты:

$$V_1 = \frac{4.84}{64.40} \cdot 100 = 7.52\ \%;$$

$$V_2 = \frac{3.03}{13} \cdot 100 = 23.31\ \%.$$

Для акций 2 коэффициент вариации почти в три раза больше, чем коэффициент вариации для акций 1.

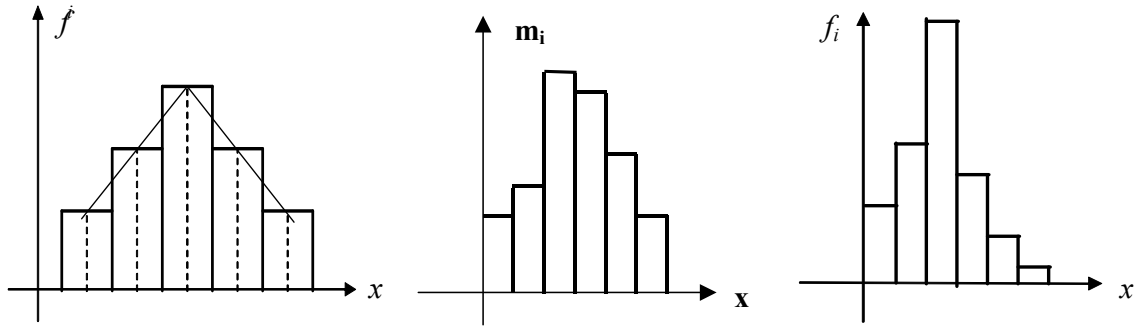
Используя коэффициент вариации в данном случае, можно сделать заключение, что покупать акции 2 более рискованно.

2.1.11. Числовые характеристики формы распределения

К числовым характеристикам формы частотных распределений относятся выборочный коэффициент асимметрии A_x и эксцесс E_x.

Симметричным называется частотное распределение, если относительно наибольшей частоты остальные частоты справа и слева для соответствующих интервалов равны.

Такое распределение будет изображаться гистограммой или полигоном, имеющим симметричную форму, случай *a* на рис. 2.7.



a – симметричное распределение

b – левосторонняя асимметрия

c – правосторонняя асимметрия

Рис. 2.7

Случаи *b* и *c* дают представление о левосторонней и правосторонней асимметрии.

Коэффициент асимметрии определяется через третий выборочный момент

$$m_3 = \frac{1}{n}\sum_{i=1}^{k} x_i^3 f_i\,, \qquad (2.26)$$

где вместо x_i используется центрированная $\dot{x}_i = x_i - \overline{X}$. Формула для выборочного коэффициента асимметрии имеет вид

$$A_x = \frac{m_3}{S^3} = \frac{1/n \sum (x_i - \overline{X})^3 f_i}{S^3}, \qquad (2.27)$$

где S – выборочное среднеквадратическое отклонение.

Существуют относительные коэффициенты асимметрии, одним из таких, часто используемым на практике, является коэффициент асимметрии Пирсона, который имеет вид

$$A_x = \frac{3(\overline{X} - M_e)}{S}, \qquad (2.28)$$

где M_e – медиана;

S – среднеквадратическое отклонение;

$\overline{X}$ – среднее арифметическое.

Если $A_x = 0$, то имеем симметричное распределение;

$A_x < 0$ – распределение имеет левостороннюю асимметрию;

$A_x > 0$ – распределение имеет правостороннюю асимметрию.

Для симметричных распределений выполняется соотношение $\overline{X} = M_0 = M_e$.

Для левосторонней асимметрии $M_0 < M_e \le \overline{X}$.

Для правосторонней асимметрии $M_0 > M_e \ge \overline{X}$.

Пример. Проанализировать форму частотного распределения для интервальной табл. 2.5 (см. с. 73), представляющей цены на обезболивающее лекарство.

Решение. Согласно формуле (2.28)

$$A_x = \frac{3(12.09 - 12.11)}{S} = \frac{3(12.09 - 12.11)}{1.49} = -0.04;$$

$$\hat{\hat{S}}^2 = \frac{1}{9}[(8.89 - 12.09)^2 + (10.89 - 12.09)^2 + (12.89 - 12.09)^2 +$$

$$+ (14.89 - 12.09)^2] = \frac{1}{9}(10.24 + 1.44 + 0.64 + 7.84) = 2.24;$$

$$S = \sqrt{2.24} = 1.49.$$

Числовой характеристикой, оценивающей крутость распределения, является эксцесс, который определяется через четвертый выборочный момент для центрированной $\dot{x}_i = x_i - \overline{X}$,

$$m_4 = 1/n \sum_{i=1}^{k} x_i^4 f_i \qquad (2.29)$$

в виде

$$E_x = \frac{m_4}{S^4} - 3. \qquad (2.30)$$

При $E_x = 0$ имеем нормальную крутизну.

При $E_x < 0$ имеем крутизну меньше нормальной.

При $E_x > 0$ имеем крутизну, превышающую нормальную (рис. 2.8).

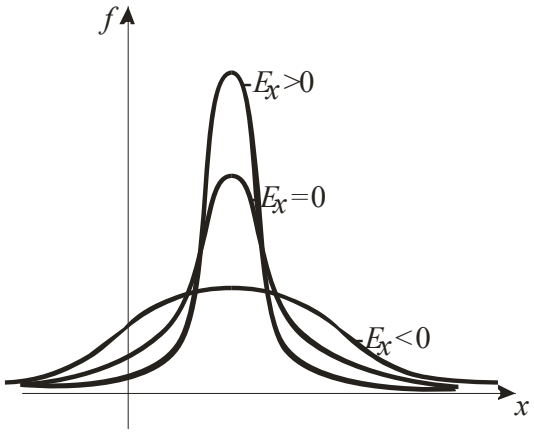


Рис. 2.8

2.1.12. Графический способ box and whisker plot (ящик с усами)

Box and whisker plot – это диаграмма, с помощью которой можно сделать выводы:

– о симметричности распределения;

– о наличии резковыделяющихся наблюдений в выборке, которые являются неоднородными к данному ряду наблюдений.

Данная диаграмма определяется пятью величинами:

– медианой Q_2;

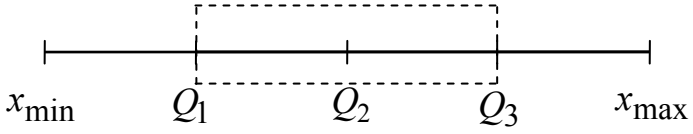
– квартилем Q_1;

– квартилем Q_3;

– наименьшим значением из наблюдений $x_{\min}$;

– наибольшим значением из наблюдений $x_{\max}$.

На числовой оси откладываются значения $x_{\min}$, $x_{\max}$, Q_1, Q_3 и Q_2.



Q_1 и Q_3 определяют границы ящика, внутри которого находится 50 % наблюдений, Q_1 – нижняя граница, Q_3 – верхняя граница. Определяется место медианы Q_2 в ящике:

– Если медиана попадает в центр ящика, то распределение симметрично.

– Если медиана находится в правой (верхней) половине ящика, то распределение имеет левостороннюю асимметрию.

– Если медиана попадает в левую (нижнюю) часть ящика, то распределение имеет правостороннюю асимметрию.

Для выводов о наличии резковыделяющихся наблюдений в выборке определяется

$$IQR = Q_3 - Q_1, \qquad (2.31)$$

где IQR – интервальный размах.

Определяются значения $1.5IQR$ и $3IQR$.

Значения $1.5IQR$ и $3IQR$ откладываются от границы Q_1 влево и от границы Q_3 вправо.

Значения $1.5IQR$ справа от Q_3 и слева от Q_1 определяют границы внутреннего забора (inner fence).

Значения $3IQR$ справа от Q_3 и слева от Q_1 определяют границы внешнего забора (outer fence).

Те наблюдения, которые находятся между Q_1 и $1.5IQR$ и Q_3 и $1.5IQR$, являются однородными с наблюдениями данной выборки.

Те наблюдения, которые находятся между внутренним и внешним заборами, являются слабыми выбросами, которые нужно дополнительно проанализировать.

Те наблюдения, которые находятся за внешним забором, являются резковыделяющимися наблюдениями, и эти наблюдения подлежат строгому анализу.

2.1.13. Задание для самостоятельной работы

1. Тренер по легкой атлетике должен решить, кого из двух спортсменов выбрать для стометровой дистанции в предстоящих соревнованиях. Тренер свое решение должен принять на основании пяти забегов между атлетами:

Анна (сек.)	12,1	12,0	12,0	16,8	12,1
Ирина (сек.)	12,3	12,4	12,4	12,5	12,4

а) Основываясь на этих данных, кого из атлетов следует выбрать тренеру и почему?

б) Если тренер знал о падении Анны на старте в четвертом забеге, то следует ли учесть это?

в) Обсудить концепции среднего арифметического и медианы как мер центральной тенденции. Как это связано с *А* и *Б* ?

2. Предположим, что благодаря ошибке данные, содержащие сведения о заработной плате (недельной) в девяти торговых компаниях, имеют вид

13, 15, 14, 17, 13, 16, 15, 16, 61.

а) Показать, как эта ошибка влияет на среднее значение и на медиану.

б) Ошибку 61 следует заменить на 16. Вычислить указанные статистики для плохих и хороших данных и сделать выводы.

3. Средний возраст в классе из 20-ти мальчиков – 12 лет 6 месяцев. Четыре новых мальчика появились в классе. Их средний возраст – 12 лет. На сколько понизился средний возраст в классе?

4. Данные представляют собой плату за обучение в выборке из 15-ти подготовительных курсов на севере страны и из 15-ти подготовительных курсов в средней полосе страны.

а) Какой совет Вы бы дали своему кузену, который хочет выбрать подготовительные курсы для обучения, используя числовые характеристики?

б) Проанализировать форму распределений, используя числовые характеристики.

S		
10,5	8,9	9,6
10,1	9,3	9,1
10,0	9,7	11,2
10,0	10,4	10,5
9,8	10,0	9,9

M		
7,9	10,6	8,4
8,2	10,0	9,2
9,1	8,5	10,7
9,3	7,5	9,5
8,8	9,3	9,8

5. В таблице представлены накопленные частоты 560-ти абитуриентов:

– Построить кумуляту.

– Сколько абитуриентов не прошли конкурс, если проходной балл 45?

– Каким должен быть проходной балл, если 60 % абитуриентов должны пройти конкурс?

– Определить медиану.

10	18
20	43
30	78
40	130
50	240
60	372
70	462
80	523
90	552
100	560

6. Симметричны ли данные 7; 5; 6; 6; 6; 4; 8; 6; 9; 3?
Для ответа на вопрос использовать:
– ящик с усами;
– коэффициент асимметрии;
– меры центральной тенденции.

7. Фирма имеет парк из ста автомобилей. В течение месяца число километров, которые проехал каждый грузовик в выборке из 10-ти автомобилей, представлены в виде 3;4;0;8;0;0;0;8;5;7.
Определить следующие статистики:
– среднее арифметическое;
– медиану;
– моду;
– размах;
– дисперсию;
– стандартное отклонение.
Определить форму распределения.