1 Первичная обработка статистических данных

1. Абстрактная и конкретная выборки.
2. Основные числовые характеристики выборки.
3. Вариационные ряды выборки.
4. Гистограмма частот.
5. Эмпирическая функция распределения.

Пусть в одинаковых условиях и независимо друг от друга производится n измерений случайной величины ξ. Назовем случайную величину ξ *теоретической случайной величиной*, а ее функцию распределения $F(x)$ – *теоретической функцией распределения*. Пусть $x_1, x_2, ..., x_n$ – результаты измерений. Набор $X = (x_1, x_2, ..., x_n)$ называется *конкретной выборкой* объема n из распределения $F(x)$.

Абстрактной выборкой объема n называется совокупность n независимых одинаково распределенных случайных величин $\xi_1, ..., \xi_n$, распределение каждой из которых совпадает с распределением теоретической случайной величины ξ.

Если элементы выборки $X = (x_1, x_2, ..., x_n)$ упорядочить по возрастанию, получится новый набор, называемый *вариационным рядом*:

$$x_{(1)} \le x_{(2)} \le ... \le x_{(n)}.$$

Если среди элементов вариационного ряда есть повторяющиеся, то можно выделить $m \le n$ его различных значений, расположив их в порядке возрастания. Обозначим их $z_{(1)}, z_{(2)}, ..., z_{(m)}$ $(z_{(1)} < z_{(2)} < ... < z_{(m)})$. Число k_i, показывающее, сколько раз элемент $z_{(i)}$ встретился в выборке, называется *частотой*, а $\frac{k_i}{n}$ – *относительной частотой* (*частостью*) этого значения, $i = 1, ..., m$ $\left(\sum_{i=1}^{m} k_i = n \right)$.

Статистическим рядом называется таблица, содержащая в первой строке значения $z_{(1)}, z_{(2)}, ..., z_{(m)}$, а во второй строке – частоты значений.

$z_{(1)}$	$z_{(2)}$	...	$z_{(m)}$
k_1	k_2	...	k_m

Случайная величина τ с рядом распределения

τ	$z_{(1)}$	...	$z_{(n)}$
	$\frac{k_1}{n}$	...	$\frac{k_m}{n}$

называется *эмпирической случайной величиной*, а соответствующая ей функция распределения $F_n^*(x)$ – *выборочной или эмпирической функцией распределения:*

$$F_n^*(z)=\begin{cases}0, & z \le z_{(1)}\\ \ldots\\ \dfrac{k_1+\ldots+k_i}{n}, & z_{(i)} < z \le z_{(i+1)}\\ \ldots\\ 1, & z > z_{(n)}\end{cases}$$

Элементы выборки можно объединить в группы и построить *интервальный вариационный ряд*. Для этого отрезок $[x_{(1)}, x_{(n)}]$ разбивается на k равных промежутков $\Delta_1,\ldots,\Delta_k$. Определяются середины промежутков l_i, $i=1,\ldots,k$. Количество промежутков k зависит от объема выборки n и может быть вычислено по формуле *Стерджесса*:

$$k \approx 1+3{,}32\lg n\,.$$

Далее определяются *частоты интервального вариационного ряда* n_i – количество элементов выборки, попавших в i-й промежуток, $i=1,\ldots,k$, $\sum_{i=1}^{k} n_i = n$. *Относительные частоты (частости) интервального вариационного ряда* определяются как $\omega_i = \frac{n_i}{n}$, $i=1,\ldots,k$. Результаты удобно представить в виде таблицы 1:

Таблица 1

Интервал	Δ_1	Δ_2	...	Δ_k
Середина интервала	l_1	l_2	...	l_k
Частота	n_1	n_2	...	n_k
Относительная частота	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_k}{n}$

Заметим, что эмпирическая функция распределения может быть определена как функция распределения случайной величины,

принимающей значения $l_1,...,l_k$ с вероятностями $\frac{n_1}{n},...,\frac{n_k}{n}$ соответственно.

Статистические данные, представленные в виде статистического ряда или интервального вариационного ряда, называют *группированными*.

Гистограмма частот группированной выборки – это график кусочно-постоянной функции, принимающей на каждом из интервалов $\Delta_1,...,\Delta_k$ значение $\frac{n_i}{h}$ ($h=(x_{(n)}-x_{(1)})/k$ – длина интервала), $i=1,...,k$.

Аналогично по значениям $\frac{n_i}{hn}$ строится *гистограмма относительных частот*, $i=1,...,k$.

Полигоном частот для данных, представленных в виде интервального вариационного ряда, называется график ломаной с вершинами в точках $\left(l_i,\frac{n_i}{h}\right)$, а полигоном относительных частот – в точках $\left(l_i,\frac{n_i}{hn}\right)$, $i=1,...,k$.

При увеличении объема выборки и уменьшении интервала группирования гистограмма и полигон относительных частот могут рассматриваться как статистические аналоги теоретической плотности распределения.

В таблице 2 приведены основные числовые характеристики выборки.

Таблица 2 – Основные выборочные характеристики

Выборочное среднее	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$
Выборочная дисперсия	$\tilde{S}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-(\bar{x})^2$
Выборочное среднеквадратическое отклонение	$\tilde{S}=\sqrt{\tilde{S}^2}$
Выборочный начальный момент k-го порядка	$\overline{x^k}=\frac{1}{n}\sum_{i=1}^{n}x_i^k$
Выборочный центральный момент k-го порядка	$\tilde{S}^k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$
Выборочная мода *Mo*	Элемент выборки, встречающийся с наибольшей частотой

Выборочная медиана Me	$Me = \begin{cases} x_{(l+1)}, n = 2l+1; \\ \dfrac{x_{(l)} + x_{(l+1)}}{2}, \ n = 2l. \end{cases}$

Окончание таблицы 2

Выборочный коэффициент асимметрии	$k_a = \dfrac{\widetilde{S}^3}{(\widetilde{S})^3}$
Выборочный коэффициент эксцесса	$k_e = \dfrac{\widetilde{S}^4}{(\widetilde{S})^4} - 3$

Пример 1.1 В результате наблюдений над случайной величиной ξ получена выборка X объема $n = 30$:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1,37	0,11	1,56	–0,11	0,23	–0,76	–0,13	–0,64	–0,46	–0,88

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
–0,56	1,28	1,16	–0,3	–0,31	1,13	–0,17	0,6	–1,16	2,65

x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}
1,55	0,29	–2,16	–0,77	0,93	0,01	–1,56	1,59	–1,13	–1,74

Произвести статистическую обработку результатов:

1) вычислить основные числовые характеристики выборки;

2) построить интервальный вариационный ряд выборки и гистограмму частот;

3) построить эмпирическую функцию распределения, взяв в качестве значений середины интервалов интервального вариационного ряда.

1) Основные числовые характеристики выборки.

Выборочное среднее:

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{1}{30}\left(1{,}37 + 0{,}11 + 1{,}56 + \ldots + (-1{,}74)\right) \approx 0{,}05.$$

Выборочная дисперсия:

$$\widetilde{S}^2 = \frac{1}{30}\sum_{i=1}^{30} x_i^2 - (\bar{x})^2 \approx \frac{1}{30}\left(1{,}37^2 + 0{,}11^2 + 1{,}56^2 + \ldots + (-1{,}74)^2\right) - 0{,}05^2 \approx 1{,}26.$$

Выборочное среднеквадратическое отклонение:

$$\widetilde{S} = \sqrt{\widetilde{S}^2} \approx \sqrt{1{,}26} \approx 1{,}12;$$

Вариационный ряд выборки имеет вид:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
–2,16	–1,74	–1,56	–1,16	–1,13	–0,88	–0,77	–0,76	–0,64	–0,56

$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$	$x_{(19)}$	$x_{(20)}$
–0,46	–0,31	–0,3	–0,17	–0,13	–0,11	0,01	0,11	0,23	0,29

$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	$x_{(26)}$	$x_{(27)}$	$x_{(28)}$	$x_{(29)}$	$x_{(30)}$
0,6	0,93	1,13	1,16	1,28	1,37	1,55	1,56	1,59	2,65

Размах выборки:

$$x_{(30)} - x_{(1)} = 2{,}65 - (-2{,}16) = 4{,}81.$$

Выборочная медиана. Объем выборки $n = 30$ – четное число, поэтому воспользуемся формулой

$$Me = \frac{x_{(15)} + x_{(15+1)}}{2} = \frac{-0{,}13 + (-0{,}11)}{2} = -0{,}12.$$

Перед вычислением выборочных коэффициентов асимметрии и эксцесса найдем выборочные центральные моменты третьего и четвертого порядков:

$$\tilde{S}^3 = \frac{1}{30}\sum_{i=1}^{30}(x_i - \bar{x})^3 \approx \frac{1}{30}\left((1{,}37 - 0{,}05)^3 + (0{,}11 - 0{,}05)^3 + \ldots + (-1{,}74 - 0{,}05)^3\right) \approx 0{,}29;$$

$$\tilde{S}^4 = \frac{1}{30}\sum_{i=1}^{30}(x_i - \bar{x})^4 \approx \frac{1}{30}\left((1{,}37 - 0{,}05)^4 + (0{,}11 - 0{,}05)^4 + \ldots + (-1{,}74 - 0{,}05)^4\right) \approx 3{,}92.$$

Выборочный коэффициент асимметрии:

$$k_a = \frac{\tilde{S}^3}{(\tilde{S})^3} \approx \frac{0{,}29}{(1{,}12)^3} \approx 0{,}21.$$

Выборочный коэффициент эксцесса:

$$k_e = \frac{\tilde{S}^4}{(\tilde{S})^4} - 3 \approx -0{,}52.$$

2) Построим интервальный вариационный ряд выборки. Число интервалов вычислим по формуле Стерджесса: $k = 1 + 3{,}32 \lg 30 \approx 6$. Разобьем отрезок $[x_{(1)}, x_{(30)}] = [-2{,}16; 2{,}65]$ на 6 равных интервалов. Длина интервала $h = \frac{x_{(30)} - x_{(1)}}{k} = \frac{2{,}65 - (-2{,}16)}{6} \approx 0{,}8$. Результаты представим в

виде таблицы 3:

Таблица 3

Интервал	Середина интервала z_i	Частота n_i	Относительная частота $\frac{n_i}{n}$
$[-2{,}16; -1{,}36)$	–1,76	3	$\frac{3}{30}=\frac{1}{10}$
$[-1{,}36; -0{,}56)$	–0,96	6	$\frac{6}{30}=\frac{1}{5}$
$[-0{,}56; 0{,}24)$	–0,16	10	$\frac{10}{30}=\frac{1}{3}$
$[0{,}24; 1{,}04)$	0,64	3	$\frac{3}{30}=\frac{1}{10}$
$[1{,}04; 1{,}84)$	1,44	7	$\frac{7}{30}$
$[1{,}84; 2{,}65]$	2,245	1	$\frac{1}{30}$

На рисунке 1 изображена гистограмма частот.

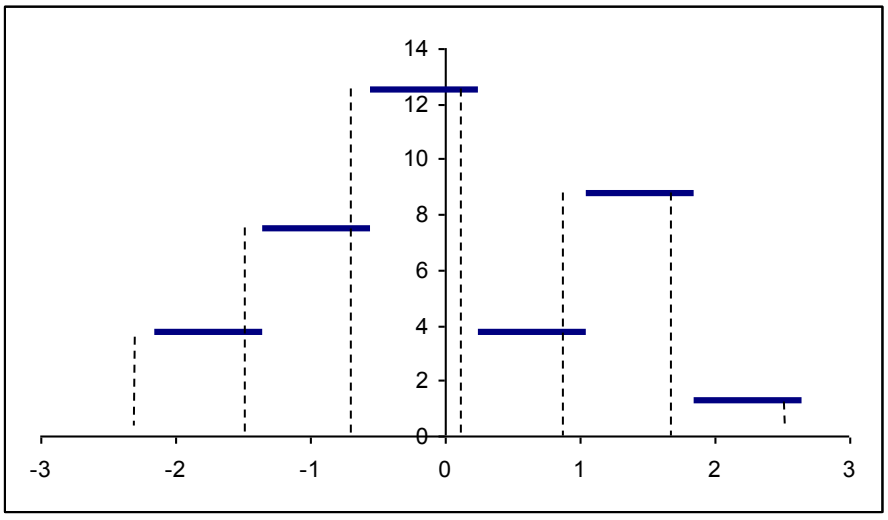

Рисунок 1

3) Построим эмпирическую функцию распределения, взяв в качестве значений середины интервалов интервального вариационного

ряда

$$F^*(z)=\begin{cases}0, & z\le -1{,}76,\\ \dfrac{1}{10}, & -1{,}76<z\le -0{,}96,\\ \dfrac{3}{10}, & -0{,}96<z\le -0{,}16,\\ \dfrac{19}{30}, & -0{,}16<z\le 0{,}64,\\ \dfrac{11}{15}, & 0{,}64<z\le 1{,}44,\\ \dfrac{29}{30}, & 1{,}44<z\le 2{,}245,\\ 1, & z>2{,}245.\end{cases}$$

График эмпирической функции распределения изображен на рисунке 2.

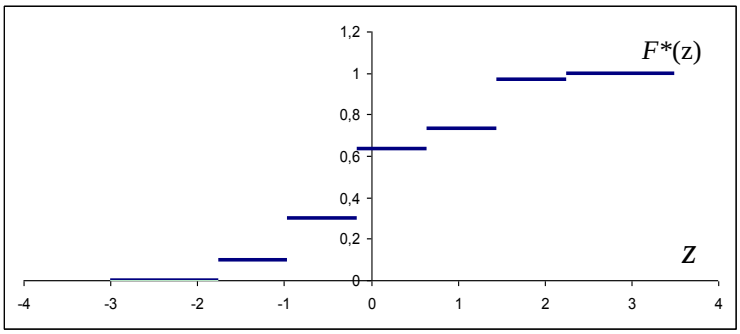

Рисунок 2

Вопросы для самоконтроля

1. Дайте определение абстрактной и конкретной выборок.
2. Укажите основные числовые характеристики выборки: размах выборки, выборочное среднее, выборочная дисперсия, выборочная медиана, выборочные коэффициенты асимметрии и эксцесса.
3. Как построить интервальный вариационный ряд выборки?
4. Как построить гистограмму частот?
5. Что называется эмпирической функцией распределения?