

Лекция № 2.

**Числовые
характеристики
выборки**

Выборочное среднее. Выборочная дисперсия. Выборочное среднее квадратическое отклонение

В теории вероятностей определили числовые характеристики для случайных величин, с помощью которых можно сравнивать однотипные случайные величины. Аналогично можно определить ряд числовых характеристик и для выборки. Поскольку эти характеристики вычисляются по статистическим данным (по данным, полученным в результате наблюдений), их называют **статистическими характеристиками**.

**Пусть дано статистическое распределение
выборки объёма n :**

X_1	X_1	X_2	X_3	X_4	...	X_m
-------	-------	-------	-------	-------	-----	-------

n_1	n_1	n_2	n_3	n_4	...	n_m
-------	-------	-------	-------	-------	-----	-------

где m – число вариантов.

Выборочным средним \bar{x}_e называется среднее арифметическое всех значений выборки.

$$\bar{x}_e = \frac{1}{n} \sum_{i=1}^m x_i n_i$$

Выборочное среднее можно записать и так: $\bar{x}_e = \frac{1}{n} \sum_{i=1}^m x_i p_i^*$, где:

p_i^* - частота.

В случае интервального статистического ряда в качестве x_i берут середины интервалов, а n_i – соответствующие им частоты.

Выборочной дисперсией D_v называется среднее арифметическое квадратов отклонений выборки от

выборочного среднего \overline{x}_v :

$$D_v = \frac{1}{n} \sum_{i=1}^m (x_i - \overline{x}_v)^2 \cdot n_i \quad \text{или} \quad D_v = \frac{1}{n} \sum_{i=1}^m (x_i - \overline{x}_v)^2 \cdot p_i^*$$

Выборочное среднее квадратическое выборки определяется формулой:

$$\sigma_v = \sqrt{D_v}$$

Особенность σ_v состоит в том, что оно измеряется в тех же единицах, что и данные выборки.

Если объём выборки мал ($n \leq 30$), то пользуются **исправленной выборочной дисперсией:**

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

Величина $S = \sqrt{S^2}$ называется **исправленным средним квадратическим отклонением.**

Выборочные начальные и центральные моменты. Асимметрия. Эксцесс.

Приведём краткий обзор характеристик, которые наряду с уже рассмотренными применяются для анализа статистических рядов и являются аналогами соответствующих числовых характеристик случайной величины.

Среднее выборочное и выборочная дисперсия являются частным случаем более общего понятия – момента статистического ряда.

Начальным выборочным моментом порядка называется среднее арифметическое l -х-степеней всех значений выборки:

$$v_l^* = \frac{1}{n} \sum_{i=1}^m x_i^l \cdot n_i \quad \text{или} \quad v_l^* = \sum_{i=1}^m x_i^l \cdot p_i^* .$$

Из определения следует, что начальный выборочный момент первого порядка:

$$v_1^* = \frac{1}{n} \sum_{i=1}^m x_i \cdot n_i = \overline{x}_e$$

Центральным выборочным моментом порядка l

называется среднее арифметическое l -х-степеней отклонений наблюдаемых значений выборки от выборочного среднего \overline{x}_e .

$$\mu_l^* = \frac{1}{n} \sum_{i=1}^m \left(x_i - \overline{x}_e \right)^l \cdot n_i \quad \text{или} \quad \mu_l^* = \sum_{i=1}^m \left(x_i - \overline{x}_e \right)^l \cdot p_i^*$$

Из определения следует, что центральный выборочный момент второго порядка :

$$\mu_2^* = \frac{1}{n} \sum_{i=1}^m (x_i - \bar{x}_e)^2 \cdot n_i = D_e = \sigma_e^2$$

Выборочным коэффициентом асимметрии называется

число A_S^* , определяемое формулой:
$$A_S^* = \frac{\mu_3^*}{\sigma_v^3}$$

Выборочный коэффициент асимметрии служит для характеристики асимметрии полигона вариационного ряда. Если полигон асимметричен, то одна из его ветвей, начиная с вершины, имеет больший «спуск», чем другая. Если $A_S^* < 0$, то более пологий «спуск» полигона наблюдается слева; если $A_S^* > 0$ - справа. В первом случае асимметрию называют **левосторонней**, а во втором – **правосторонней**.

Выборочным коэффициентом эксцесса или коэффициентом крутости называется число E_k^* , определяемое формулой:

$$E_k^* = \frac{\mu_4^*}{\sigma^4} - 3 .$$

Выборочный коэффициент эксцесса служит для сравнения на «крутость» выборочного распределения с нормальным распределением. Коэффициент эксцесса для случайной величины, распределённой по нормальному закону, равен 0.

Поэтому за стандартное значение выборочного эксцесса принимают $E_k^* = 0$. Если $E_k^* < 0$, то полигон имеет более пологую вершину по сравнению с нормальной кривой; если $E_k^* > 0$, то полигон более крутой по сравнению с нормальной кривой.

Вычисление числовых характеристик выборки

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}_e$	$(x_i - \bar{x}_e)^2 n_i$	$(x_i - \bar{x}_e)^3 n_i$	$(x_i - \bar{x}_e)^4 n_i$
x_1	n_1					
[[
x_m	n_m					
	$\sum_{i=1}^m n_i$	$\sum_{i=1}^m x_i n_i$		$\sum_{i=1}^m (x_i - \bar{x}_e)^2 n_i$	$\sum_{i=1}^m (x_i - \bar{x}_e)^3 n_i$	$\sum_{i=1}^m (x_i - \bar{x}_e)^4 n_i$

x_i - середина интервалов; n_i - частоты;

$\sum_{i=1}^m n_i = n$ - объём выборки; с помощью суммы $\sum_{i=1}^m x_i n_i$ находим \overline{x}_e

с помощью суммы $\sum_{i=1}^m (x_i - \overline{x}_e)^2 n_i$ находим D_e и $\sigma_e = \sqrt{D_e}$

с помощью суммы $\sum_{i=1}^m (x_i - \overline{x}_e)^3 n_i$ находим A_s^*

с помощью суммы $\sum_{i=1}^m (x_i - \overline{x}_e)^4 n_i$ находим E_k^*

Упрощённый способ вычисления статистических характеристик вариационных рядов

При больших значениях вариантов и соответствующих им частот вычисление выборочного среднего, дисперсии и выборочного моментов по приведённым ниже формулам приводит к громоздким вычислениям.

В этом случае условные варианты u_i , определяемые по формулам ,

$u_i = \frac{x_i - c}{h}$ где числа c и h выбираются произвольно. Чтобы упростить вычисления, в качестве c выбирают вариант, который имеет наибольшую частоту или находится в середине ряда. Число c называется «ложным нулём». В качестве h выбирают число равное длине интервала (в случае интервального ряда) или наибольший общий делитель разностей $(x_i - c)$.

Для вычисления числовых характеристик выборки составляем таблицу

u_i	n_i	$u_i n_i$	$u_i^2 n_i$	$u_i^3 n_i$	$u_i^4 n_i$	$(u_i + 1)^4 n_i$
u_1	n_1					
\sqsubset	\sqsubset					
u_m	n_m					
	$\sum_{i=1}^m n_i = n$	$\sum_{i=1}^m u_i n_i$	$\sum_{i=1}^m u_i^2 n_i$	$\sum_{i=1}^m u_i^3 n_i$	$\sum_{i=1}^m u_i^4 n_i$	$\sum_{i=1}^m (u_i + 1)^4 n_i$

Контроль:

$$\sum_{i=1}^m (u_i + 1)^4 n_i = \sum_{i=1}^m u_i^4 n_i + 4 \sum_{i=1}^m u_i^3 n_i + 6 \sum_{i=1}^m u_i^2 n_i + 4 \sum_{i=1}^m u_i n_i + n$$

С помощью сумм, полученных в нижней строке таблицы, находим условные моменты:

$$M_1^* = \frac{1}{n} \sum_{i=1}^m u_i n_i \quad M_2^* = \frac{1}{n} \sum_{i=1}^m u_i^2 n_i$$

$$M_3^* = \frac{1}{n} \sum_{i=1}^m u_i^3 n_i \quad M_4^* = \frac{1}{n} \sum_{i=1}^m u_i^4 n_i$$

Числовые характеристики выборки вычисляем по формулам:

$$\bar{x}_e = M_1^* h + c \quad D_e = \left(M_2^* - (M_1^*)^2 \right) \cdot h^2 \quad \sigma_e = \sqrt{D_e}$$

$$A_s^* = \frac{\mu_3^*}{\sigma_e^3} \quad E_k^* = \frac{\mu_4^*}{\sigma_e^4} - 3$$

где μ_3^* и μ_4^* находим по формулам: $\mu_3^* = \left(M_3^* - 3M_1^* M_2^* + 2(M_1^*)^3 \right) \cdot h^3$

$$\mu_4^* = \left(M_4^* - 4M_1^* M_3^* + 6(M_1^*)^2 \cdot M_2^* - 3(M_1^*)^4 \right) \cdot h^4$$

Пример. Вычислить числовые характеристики выборки, рассмотренной в примере 2.

В качестве вариантов x_i возьмём середины интервалов.

Перейдём к условным вариантам.

Вариант, значение которого 0,04, имеет наибольшую частоту и находится в середине модального ряда. Примем его за «ложный ноль» (начало отсчёта).

Условные варианты найдём по формуле:

$$u_i = \frac{x_i - c}{h}, \text{ где}$$

$$c = 0,04 \quad h = 0,6$$

Составим расчётную таблицу:

x_i	n_i	u_i	$u_i \cdot n_i$	$u_i^2 \cdot n_i$	$u_i^3 \cdot n_i$	$u_i^4 \cdot n_i$	$(u_i + 1)^4 \cdot n_i$
-1,76	2	-3	-6	18	-54	162	32
-1,16	6	-2	-12	24	-48	96	6
-0,56	11	-1	-11	11	-11	11	0
0,04	15	0	0	0	0	0	15
0,64	11	1	11	11	11	11	176
1,24	3	2	6	12	24	48	234
1,84	2	3	6	18	54	162	512
Σ	50		-6	94	-24	490	984

Контроль:
$$\sum_{i=1}^m (u_i + 1)^4 n_i = \sum_{i=1}^m u_i^4 n_i + 4 \sum_{i=1}^m u_i^3 n_i + 6 \sum_{i=1}^m u_i^2 n_i + 4 \sum_{i=1}^m u_i n_i + n =$$

$$= 490 + 4 \cdot (-24) + 6 \cdot 94 + 4 \cdot (-6) + 50 = 984 \rightarrow \text{расчёты проведены верно.}$$

По данным таблицы находим условные моменты:

$$M_1^* = -\frac{6}{50} = -0,12 \qquad M_2^* = \frac{94}{50} = 1,88$$

$$M_3^* = -\frac{24}{50} = -0,48 \qquad M_4^* = \frac{490}{50} = 9,8$$

Находим числовые характеристики выборки:

$$\overline{x_g} = M_1^* h + c = -0,12 \cdot 0,6 + 0,04 = -0,032$$

$$D_e = \left(M_2^* - (M_1^*)^2 \right) \cdot h^2 = \left(1,88 - (-0,12)^2 \right) \cdot 0,6^2 = 0,6716$$

$$\sigma_e = \sqrt{D_e} = \sqrt{0,672} = 0,8195$$

Вычислим центральные моменты третьего и четвёртого порядка:

$$\mu_3 = \left(M_3^* - 3M_1^*M_2^* + 2(M_1^*)^3 \right) \cdot h^3 = \left(-0,48 - 3 \cdot (-0,12) \cdot 1,88 + 2(-0,12)^3 \right) \cdot 0,6^3 = 0,0418$$

$$\mu_4^* = \left(M_4^* - 4M_1^*M_3^* + 6(M_1^*)^2 \cdot M_2^* - 3(M_1^*)^4 \right) \cdot h^4 =$$

$$= \left(9,8 - 4 \cdot (-0,12) \cdot (-0,48) + 6(-0,12)^2 \cdot 1,88 - 3(-0,12)^4 \right) \cdot 0,6^4 = 1,2127$$

Вычислим выборочные коэффициенты асимметрии и эксцесса:

$$A_s^* = \frac{\mu_3^*}{\sigma_e^3} = \frac{0,0418}{0,8195^3} = 0,0759 \quad E_k^* = \frac{\mu_4^*}{\sigma_e^4} - 3 = \frac{1,2127}{0,8195^4} - 3 = -0,3112$$