УНИВЕРСИТЕТ ИТМО

# *Теория вероятностей и математическая статистика*

## *Лекция 5. **Выборки и их характеристики***

# *Генеральная и выборочная совокупности*

Задачами математической статистики являются оценивание законов распределения и основных характеристик случайных величин, проверка статистических гипотез, анализ зависимостей между входными и выходными параметрами систем, прогнозирование, планирование эксперимента и т.д.

Эти и другие статистические выводы относительно свойств полной совокупности данных (**генеральной совокупности**) делают на основе некоторой специальным образом сформированной части данных

$$x_1, x_2, \ldots, x_n,$$

называемой **выборкой объема** $n$.

# *Свойства выборки*

Выборка должна обладать следующими свойствами.

1. Необходимо, чтобы выборка была **репрезентативной**, т. е. достаточно полно, однородно, равномерно и равновероятно по отношению к другим возможным выборкам представляла всю генеральную совокупность.

2. Выборка должна быть **рандомизированной**, т. е. полученной случайным образом в одинаковых условиях в виде последовательности повторных независимых реализаций случайной величины $X$.

3. В рамках вероятностной математической модели, привлекаемой для анализа данных, выборка должна рассматриваться как реализация $n$-мерного случайного вектора $(X_1, X_2, \ldots, X_n)$ с взаимно независимыми и одинаково распределенными компонентами.

# *Генеральная и выборочная совокупности*

**Пример 7.1.** Десять абитуриентов проходят тестирование по математике. Каждый из них может набрать от 0 до 5 баллов включительно. Пусть $X_k$ — количество баллов, набранных $k$-м ($k = 1, 2, \ldots, 10$) абитуриентом.

Тогда значения 0, 1, 2, 3, 4, 5 — все возможные количества баллов, набранных одним абитуриентом, — образуют генеральную совокупность.

Выборка $X_1, X_2, X_3, \ldots, X_{10}$ – – результат тестирования 10 абитуриентов.

Реализациями выборки могут быть следующие наборы чисел: {5, 3, 0, 1, 4, 2. 5. 4. 1. 5} или {4, 4, 5, 3, 3, 1, 5, 5, 2, 5} или {3, 4, 5. 0. 1. 2. 3, 4, 5, 4} и т. д.

# *Вариационный ряд*

Если элементы выборки расположить в порядке возрастания

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$

то получим **вариационный ряд**, элементы которого называют **порядковыми статистиками**. Наименьшее значение в выборке называют **первой порядковой статистикой** $x_{(1)}$, а наибольшее значение $n$**-ой порядковой статистикой** $x_{(n)}$. Разность между наибольшим и наименьшим значениями называют **размахом выборки**, обозначают буквой $w^*$ и вычисляют по формуле $w^* = x_{(n)} - x_{(1)}$ .

# *Статистический ряд*

**Статистическим рядом** называют систему пар чисел

$$(z_i, n_i), \quad i = 1, 2, \ldots, k,$$

где $z_i$ различные элементы выборки, расположенные в порядке возрастания, $n_i$ частота элемента в выборке, т. е. число повторений элемента. Обычно статистический ряд представляют в виде **таблицы**, где первая строка содержит элементы $z_i$, а вторая их частоты. Если в выборке нет одинаковых элементов, то статистический и вариационный ряды совпадают. По вариационному или статистическому ряду строится эмпирическая (выборочная) функция распределения $F_n^*(x)$, которая является оценкой функции распределения $F_X(x)$ случайной величины $X$, сформировавшей данную выборку.

# Статистическое распределение выборки

Пусть изучается некоторая с. в. $X$. С этой целью над с. в. $X$ производится ряд независимых опытов (наблюдений). В каждом из этих опытов величина $X$ принимает то или иное значение.

Пусть она приняла $n_1$ раз значение $x_1$, $n_2$ раз — значение $x_2$, ..., $n_k$ раз — значение $x_k$. При этом $n_1+n_2+\ldots+n_k=n$ — объем выборки. Значения $x_1, x_2, \ldots, x_k$ называются *вариантами* с. в. $X$.

Вся совокупность значений с. в. $X$ представляет собой первичный статистический материал, который подлежит дальнейшей обработке, прежде всего — упорядочению.

Операция расположения значений случайной величины (признака) по неубыванию называется *ранжированием* статистических данных. Полученная таким образом последовательность $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ значений с. в. $X$ (где $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ и $x_{(1)} = \min\limits_{1\leqslant i\leqslant n} X_i$, ..., $x_{(n)} = \max\limits_{1\leqslant i\leqslant n} X_i$) называется *вариационным рядом*.

# Статистическое распределение выборки

Числа $n_i$, показывающие, сколько раз встречаются варианты $x_i$ в ряде наблюдений, называются *частотами*, а отношение их к объему выборки — *частостями* или *относительными частотами* $(p_i^*)$, т. е.

$$p_i^* = \frac{n_i}{n},$$

где $n = \sum_{i=1}^{k} n_i$.

Перечень вариантов и соответствующих им частот или частостей называется *статистическим распределением выборки* или *статистическим рядом*.

Записывается статистическое распределение в виде таблицы. Первая строка содержит варианты, а вторая — их частоты $n_i$ (или частости $p_i^*$).

# *Интервальный статистический ряд*

В случае, когда число значений признака (с. в. $X$) велико или признак является непрерывным (т. е. когда с. в. $X$ может принять любое значение в некотором интервале), составляют *интервальный статистический ряд*. В первую строку таблицы статистического распределения вписывают частичные промежутки $[x_0, x_1)$, $[x_1, x_2), \ldots, [x_{k-1}, x_k)$, которые берут обычно одинаковыми по длине: $h = x_1 - x_0 = x_2 - x_1 = \ldots$. Для определения величины интервала $(h)$ можно использовать формулу Стерджеса:

$$h = \frac{x_{\max} - x_{\min}}{1 + \log_2 n},$$

где $x_{\max} - x_{\min}$ — разность между наибольшим и наименьшим значениями признака, $m = 1 + \log_2 n$ — число интервалов $(\log_2 n \approx 3{,}322 \lg n)$.

За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{\min} - \frac{h}{2}$. Во второй строчке статистического ряда вписывают количество наблюдений $n_i$ $(i = \overline{1, k})$, попавших в каждый интервал.

# Эмпирическая функция распределения

*Эмпирической (статистической) функцией распределения* называется функция $F_n^*(x)$, определяющая для каждого значения $x$ частость события $\{X < x\}$:

$$F_n^*(x) = p^*\{X < x\}.$$

Для нахождения значений эмпирической функции удобно $F_n^*(x)$ записать в виде

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

где $n$ — объем выборки, $n_x$ — число наблюдений, меньших $x$ ($x \in \mathbb{R}$).

Очевидно, что $F_n^*(x)$ удовлетворяет тем же условиям, что и истинная функция распределения $F(x)$ (см. п. 2.3).

При увеличении числа $n$ наблюдений (опытов) относительная частота события $\{X < x\}$ приближается к вероятности этого события

# *Теорема Гливенко*

**Теорема (Гливенко).** *Эмпирическая функция распределения* $F_n^*(x)$ *при неограниченном увеличении объема выборки сходится по вероятности при любом значении* $x \in \mathbb{R}$ *к теоретической функции распределения* $F_X(x)$ *генеральной совокупности.*

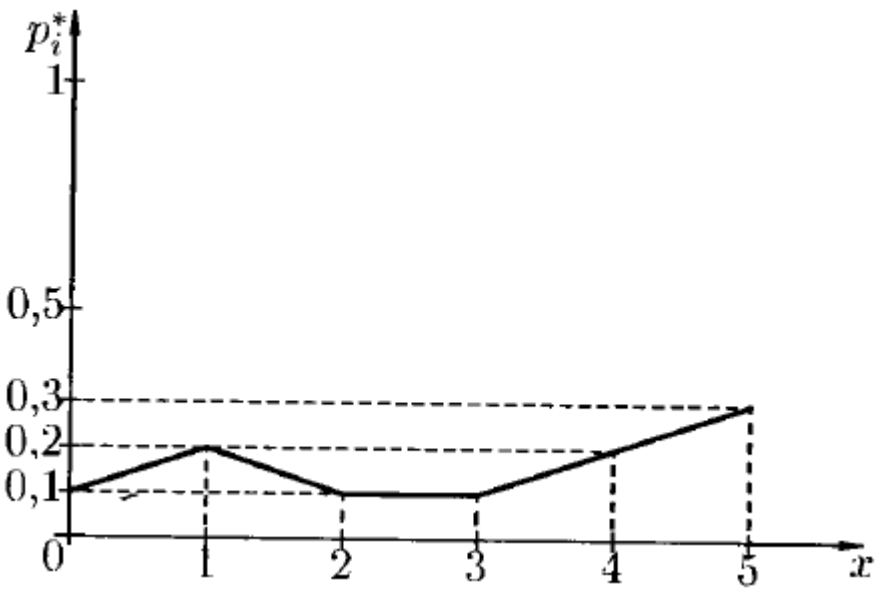
Таким образом, при большом объеме выборки эмпирическая функция распределения $F_n^*(x)$ является достаточно точным приближением для неизвестной заранее теоретической функции распределения $F_X(x)$.

# *Графическое изображение статистического распределения*

Статистическое распределение изображается графически (для наглядности) в виде так называемых полигона и гистограммы. Полигон, как правило, служит для изображения дискретного (т. е. варианты отличаются на постоянную величину) статистического ряда.

*Полигоном частот* называют ломаную, отрезки которой соединяют точки с координатами $(x_1, n_1), (x_2, n_2), \ldots, (x_k, n_k)$; *полигоном частостей* — с координатами $(x_1, p_1^*), (x_2, p_2^*), \ldots, (x_k, p_k^*)$.

Варианты $(x_i)$ откладываются на оси абсцисс, а частоты и, соответственно, частости — на оси ординат.

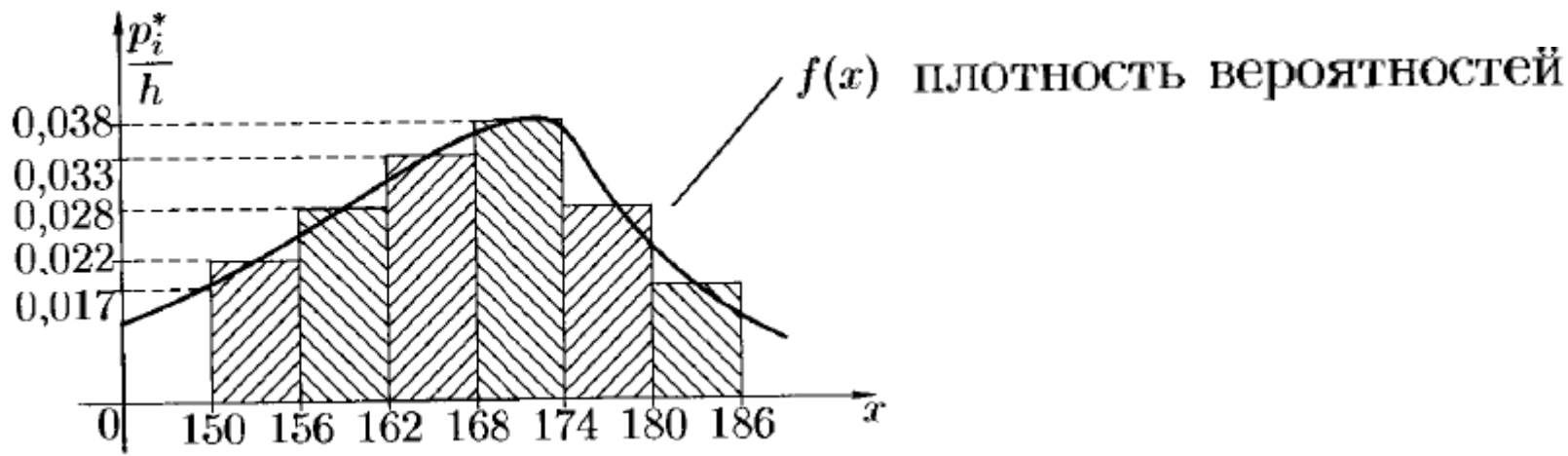


# *Графическое изображение статистического распределения*

Для непрерывно распределенного признака (т. е. варианты могут отличаться один от другого на сколь угодно малую величину) можно построить полигон частот, взяв середины интервалов в качестве значений $x_1, x_2, \ldots, x_k$. Более употребительна так называемая гистограмма.

*Гистограммой частот* (*частостей*) называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длины $h$. а высоты равны отношению $\frac{n_i}{h}$ — плотность частоты ($\frac{p_i^*}{h}$ или $\frac{n_i}{n \cdot h}$ — плотности частости).

Очевидно, площадь гистограммы частот равна объему выборки, а площадь гистограммы частостей равна единице.



# Числовые характеристики статистического распределения

Пусть статистическое распределение выборки объема $n$ имеет вид:

| $x_i$ | $x_1$ | $x_2$ | $x_3$ | $\dots$ | $x_k$ |
|---|---|---|---|---|---|
| $n_i$ | $n_1$ | $n_2$ | $n_3$ | $\dots$ | $n_k$ |

*Выборочным средним* $\overline{x}_{\text{в}}$ называется среднее арифметическое всех значений выборки:

$$\overline{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i.$$

Выборочное среднее можно записать и так: $\overline{x}_{\text{в}} = \sum_{i=1}^{k} x_i \cdot p_i^*$,

где $p_i^* = \frac{n_i}{n}$ — частость. Для обозначения выборочного среднего используют следующие символы: $\overline{x}$. $M^*(X)$. $m_x^*$.

Отметим, что в случае интервального статистического ряда в качестве $x_i$ берут середины его интервалов, а $n_i$ соответствующие им частоты.

# Числовые характеристики статистического распределения

*Выборочной дисперсией* $D_{\text{в}}$ называется среднее арифметическое квадратов отклонений значений выборки от выборочной средней $\overline{x}_{\text{в}}$, т. е.

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x}_{\text{в}})^2 \cdot n_i$$

или, что то же самое.

$$D_{\text{в}} = \sum_{i=1}^{k}(x_i - \overline{x}_{\text{в}})^2 \cdot p_i^*.$$

Можно показать, что $D_{\text{в}}$ может быть подсчитана также по формуле:

$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot n_1 - (\overline{x}_{\text{в}})^2$, т. е.

$$D_{\text{в}} = \overline{x^2} - (\overline{x})^2,$$

здесь $\overline{x} = \overline{x}_{\text{в}}$.

# Числовые характеристики статистического распределения

*Выборочное среднее квадратическое отклонение* выборки определяется формулой

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}}.$$

Особенность выборочного с. к. о. ($\sigma_{\text{в}}$) состоит в том, что оно измеряется в тех же единицах, что и изучаемый признак.

При решении практических задач используется и величина

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{k} (x_i - \overline{x}_{\text{в}})^2 \cdot n_i,$$

т. е.

$$S^2 = \frac{n}{n-1} D_{\text{в}},$$

которая называется *исправленной выборочной дисперсией*

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

Для непрерывно распределенного признака формулы для выборочных средних будут такими же, но за значения $x_1, x_2, \ldots, x_k$ надо брать не концы промежутков $[x_0, x_1), [x_1, x_2), \ldots$, а их середины

# Числовые характеристики статистического распределения

*Размахом вариации* называется число $R = x_{(n)} - x_{(1)}$, где $x_{(1)} = \min\limits_{1 \leqslant k \leqslant n} x_k$, $x_{(n)} = \max\limits_{1 \leqslant k \leqslant n} x_k$ или $R = x_{\max} - x_{\min}$. где $x_{\max}$ — наибольший, $x_{\min}$ — наименьший вариант ряда.

*Модой* $M_o^*$ вариационного ряда называется вариант, имеющий наибольшую частоту.

*Медианой* $M_e^*$ вариационного ряда называется значение признака (с. в. $X$), приходящееся на середину ряда.

Если $n = 2k$ (т. е. ряд $x_{(1)}, x_{(2)}, \ldots, x_{(k)}, x_{(k+1)}, \ldots, x_{(2k)}$ имеет четное число членов), то $M_e^* = \dfrac{x_{(k)} + x_{(k+1)}}{2}$; если $n = 2k + 1$, то $M_e^* = x_{(k+1)}$.