

# Теория вероятностей и математическая статистика



## ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

# Задачи математической статистики



**Математическая статистика** – раздел математики, посвященный методам сбора, анализа и обработки результатов статистических данных наблюдений для научных и практических целей.

Установление закономерностей, которым подчинены массовые случайные явления, основано на изучении статистических данных (результатов наблюдений) **методами теории вероятностей**.

# Задачи математической статистики



**Первая задача** математической статистики—указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов.

**Вторая задача** математической статистики—разработать методы анализа статистических данных в зависимости от целей исследования. Сюда относятся:

- a) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости случайной величины от одной или нескольких случайных величин и др.;
- b) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

# Задачи математической статистики



Современная математическая статистика разрабатывает способы определения **числа** необходимых **испытаний** до начала исследования (планирование эксперимента), в ходе исследования (последовательный анализ) и решает многие другие задачи.

Современную математическую статистику определяют как **науку о принятии решений в условиях неопределенности.**

# Генеральная и выборочная совокупности



Пусть требуется изучить **совокупность однородных объектов** относительно некоторого качественного или количественного признака, характеризующего эти объекты.

Пример: если имеется партия деталей, то качественным признаком может служить стандартность детали, а количественным — контролируемый размер детали.

Если совокупность содержит очень большое число объектов, то провести сплошное обследование физически невозможно. В таких случаях **случайно отбирают** из всей совокупности ограниченное число объектов и подвергают их изучению.

# Генеральная и выборочная совокупности



**Выборочной совокупностью** или просто **выборкой** называют совокупность случайно отобранных объектов.

**Генеральной совокупностью** называют совокупность объектов, из которых производится выборка.

**Объемом совокупности** (выборочной или генеральной) называют число объектов этой совокупности.

**Пример:** из 1000 деталей отобрано для обследования 100 деталей. Объем генеральной совокупности  $N = 1000$ , а объем выборки  $n = 100$ .

# Виды выборки



Выборка должна быть **репрезентативной**, т.е. правильно отражать пропорции генеральной совокупности.

Это достигается **случайностью** отбора, когда все объекты генеральной совокупности имеют одинаковую вероятность быть отобранными.

Выборки подразделяют на повторные и бесповторные.

**Повторной** называют выборку, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.

**Бесповторной** называют выборку, при которой отобранный объект в генеральную совокупность не возвращается.

# Вариационные ряды



Различные значения признака (случайной величины  $X$ ) называются **вариациями** (обозначаем их через  $x$ ).

Наблюдаемые значения признака называются **вариантами**.

Число, показывающее, сколько раз встречается варианта в статистической совокупности, называется **абсолютной частотой варианты**.

Отношение частоты к объему генеральной совокупности называется **относительной частотой (частостью)** или **статистической вероятностью**, и обозначается  $w_i$ :

$$w_i = \frac{n_i}{n}$$



# Вариационные ряды



**Пример.** Для исследования жителей г.Ярославля (генеральная совокупность) на доминирующий цвет волос (качественный признак) рассмотрели 500 человек из разных районов города (выборочная совокупность). Получили следующие результаты:

Блондины – 85 чел.

Брюнеты – 126 чел.

Шатены – 73 чел.

Русоволосые – 154 чел.

Каштановый цвет – 62 чел.

Цвет волос – вариация;

Блондины, брюнеты и т.д. – варианты;

85, 126, и т.д. – абсолютные частоты.

# Вариационные ряды



**Пример.** Измерили рост 50 старшеклассников в сантиметрах:

149	150	150	151	151	152	152	153	154	154
155	155	155	156	156	157	157	157	158	158
159	159	159	159	161	161	161	162	162	162
162	162	165	166	166	166	167	167	169	170
171	171	173	173	173	175	176	178	180	182

Рассмотрение и осмысление этих данных (особенно при большом числе наблюдений  $n$ ) затруднительно, и по ним практически нельзя представить характер распределения признака (случайной величины  $X$  - рост).

# Вариационные ряды



Полученные данные можно представить в виде таблицы

$x_i$	145-149	150-154	155-159	160-164	165-169	170-174	175-179	180-184
$n_i$	1	9	14	8	7	6	3	2
$w_i$	1/50	9/50	14/50	8/50	7/50	6/50	3/50	2/50

Группы роста – **вариации**;

значения вариаций 145-149, 150-154, ... – **варианты**.

1, 9, 14, и т.д. – **абсолютные частоты**;

$\frac{1}{50}$ ,  $\frac{9}{50}$ ,  $\frac{14}{50}$  ... – **относительные частоты (частоты)**.

# Вариационные ряды



**Определение.** **Вариационным рядом** называется ранжированный в порядке возрастания (или убывания) ряд вариантов с соответствующими им абсолютными или относительными частотами.

Вариационные ряды бывают **дискретными** и **интервальными**.

**Дискретные вариационные ряды** строят обычно в том случае, если значения изучаемого признака могут отличаться друг от друга не менее чем на некоторую конечную величину. В дискретных вариационных рядах задаются точечные значения признака.

**Интервальные вариационные ряды** строят обычно в том случае, если значения изучаемого признака могут отличаться друг от друга на сколь угодно малую величину. Значения признаков в них задаются в виде интервалов.

# Вариационные ряды



<b>Варианты <math>x_i</math></b>	$x_1$	$x_2$	...	$x_k$
<b>Частоты <math>n_i</math></b>	$n_1$	$n_2$	...	$n_k$

Общий вид дискретного ряда

<b>Варианты <math>x_i</math></b>	$x_1 - x_2$	$x_2 - x_3$	...	$x_{k-1} - x_k$
<b>Частоты <math>n_i</math></b>	$n_1$	$n_2$	...	$n_k$

Общий вид интервального ряда

# Вариационные ряды



Пример дискретного (точечного) вариационного ряда

Тарифный разряд $x_i$	1	2	3	4	5	6	$\Sigma$
Частота (количество рабочих) $n_i$	2	3	6	8	22	9	50

Пример интервального вариационного ряда

$x_i$	145-149	150-154	155-159	160-164	165-169	170-174	175-179	180-184
$w_i$	1/50	9/50	14/50	8/50	7/50	6/50	3/50	2/50

# Вариационные ряды



<i>i</i>	<i>Выработка в отчетном году в процентах к предыдущему x</i>	<i>Частота (количество рабочих) n<sub>i</sub></i>
1	94,0—100,0	3
2	100,0—106,0	7
3	106,0—112,0	11
4	112,0—118,0	20
5	118,0—124,0	28
6	124,0—130,0	19
7	130,0—136,0	10
8	136,0—142,0	2
	$\Sigma$	100

Пример интервального ряда

# Вариационные ряды



В интервальных вариационных рядах в каждом интервале выделяют верхнюю и нижнюю границы.

Разность между верхней и нижней границами интервала называется интервальной разностью или **длиной интервала**. В общем виде интервальную разность  $k_i$  представим как

$$k_i = X_{i \text{ (max)}} - X_{i \text{ (min)}}$$

Первый и последний интервалы могут быть **открытыми**, т.е. иметь только одну границу.

Число интервалов  $k$  следует брать не очень большим, чтобы после группировки ряд не был громоздким, и не очень малым, чтобы не потерять особенности распределения признака.



# Вариационные ряды



Разность между наибольшим и наименьшим значением вариант  $x_{\max} - x_{\min}$  называется **размахом выборки**.

Согласно формуле Стерджеса рекомендуемое число интервалов  $k = 1 + 3,322 * \lg n$ ,

а длина интервала:

$$h = \frac{x_{\max} - x_{\min}}{k}$$

где  $n$  число единиц совокупности;  
 $x_{\max}$  и  $x_{\min}$  – наибольшее и наименьшее значения вариационного ряда.

За начало первого интервала рекомендуется брать величину, равную  $x_{\text{нач}} = x_{\min} - h/2$

# Вариационные ряды



**Пример.** Необходимо изучить изменение выработки на одного рабочего механического цеха в отчетном году по сравнению с предыдущим. Получены следующие данные о распределении 100 рабочих цеха по выработке в отчетном году (в процентах к предыдущему году):

**97,8; 97,0; 101,7; 132,5; ...; 142,3; 104,2; 141,0; 122,1**

**100 значений**

# Вариационные ряды



Разобьем варианты на отдельные интервалы, т.е. проведем их группировку:  $x_{\max} = 142,3$      $x_{\min} = 97,0$ .

По ф. Стерджеса: 
$$h = \frac{x_{\max} - x_{\min}}{1 + 3.322 \cdot \lg n} = \frac{142.3 - 97.0}{1 + 3.322 \cdot \lg 100} = \frac{45.3}{7.644} \approx 5.93$$

Примем  $h = 6,0$ .

За начало первого интервала рекомендуется брать величину  $x_{\text{нач}} = x_{\min} - h/2 = 97,0 - 6/2 = 94,0$ .

# Вариационные ряды



Сгруппированный ряд можно представить в виде таблицы.

<i>i</i>	<i>Выработка в отчетном году в процентах к предыдущему x</i>	<i>Частота (число рабочих) n<sub>i</sub></i>	<i>Частость (доля рабочих) w<sub>i</sub> = <math>\frac{n_i}{n}</math></i>
1	94,0—100,0	3	0,03
2	100,0—106,0	7	0,07
3	106,0—112,0	11	0,11
4	112,0—118,0	20	0,20
5	118,0—124,0	28	0,28
6	124,0—130,0	19	0,19
7	130,0—136,0	10	0,10
8	136,0—142,0	2	0,02
	$\Sigma$	100	1,00

# Вариационные ряды



**Пример.** Для контроля качества в 40 пробах стали GS50 определялось содержание углерода (%C) и прочность на разрыв (Н/мм). Данные оформлены в виде таблицы чисел:

$x_i$	$z_i$	$x_i$	$z_i$	$x_i$	$z_i$	$x_i$	$z_i$	$x_i$	$z_i$	$x_i$	$z_i$	$x_i$	$z_i$	$x_i$	$z_i$
0,30	589	0,35	535	0,37	602	0,29	572	0,29	537	0,32	562	0,34	596	0,38	557
0,33	614	0,32	593	0,33	544	0,30	555	0,34	574	0,38	601	0,36	605	0,37	558
0,37	612	0,39	582	0,34	545	0,33	555	0,39	570	0,37	587	0,33	575	0,34	587
0,36	572	0,30	538	0,33	562	0,32	518	0,37	540	0,38	587	0,34	570	0,35	580
0,31	548	0,32	566	0,30	576	0,32	539	0,38	575	0,33	614	0,36	550	0,36	560

Представить данные в виде вариационных рядов данные для выборки, составленной из данных измерений содержания углерода, и для выборки, составленной из измерений прочности на разрыв.

# Вариационные ряды



**Решение.** Дана независимая выборка:

0.3, 0.33, 0.37, 0.36, 0.31, 0.29, 0.34, 0.39, 0.37, 0.38, 0.35,  
0.32, 0.39, 0.3, 0.32, 0.32, 0.38, 0.37, 0.38, 0.33, 0.37, 0.33,  
0.34, 0.33, 0.3, 0.34, 0.36, 0.33, 0.34, 0.36, 0.29, 0.3, 0.33,  
0.32, 0.32, 0.38, 0.37, 0.34, 0.35, 0.36.

$$x_{\max} = 0,39 \quad x_{\min} = 0,29.$$

По ф. Стерджеса: 
$$h = \frac{x_{\max} - x_{\min}}{1 + 3.322 \cdot \lg n} = \frac{0.39 - 0.29}{1 + 3.322 \cdot \lg 40} = \frac{0.1}{6.32} \approx 0.016$$

За начало первого интервала возьмем величину

$$x_{\text{нач}} = x_{\min} - h/2 = 0,29 - 0,008 = 0,282.$$

$x_i$	0,282-0,298	0,299-0,315	0,316-0,332	0,333-0,349	0,35-0,366	0,367-0,383	0,384-0,4
$n_i$	2	5	11	5	6	9	2
$w_i$	0,05	0,125	0,275	0,125	0,15	0,225	0,05

# Эмпирическая функция распределения



Все важнейшие характеристики случайной величины могут быть выражены в терминах ее функции распределения.

В задачах математической статистики функция распределения генеральной совокупности (теоретическая) всегда является неизвестной.

Основываясь на выборке, можно построить хорошее приближение для неизвестной функции распределения. Так как эта функция находится эмпирическим (опытным) путем, то ее называют **эмпирической**.

# Эмпирическая функция распределения



**Эмпирической функцией распределения** (функцией распределения выборки) называют функцию  $F^*(x)$ , определяющую для каждого значения  $x$  относительную частоту события  $X < x$ .

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

где  $n_x$  — число наблюдений, при которых наблюдалось значение признака, меньше  $x$ ;  $n$  — общее число наблюдений (объем выборки).



# Эмпирическая функция распределения



При больших  $n$   $F^*(x) \rightarrow F(x)$ .

$F^*(x)$  обладает всеми свойствами  $F(x)$ .

1. значения эмпирической функции принадлежат отрезку  $[0, 1]$ ;
2.  $F^*(x)$  — неубывающая функция;
3. если  $x_1$  — наименьшая варианта, то  $F^*(x) = 0$  при  $x \leq x_1$ ,  
если  $x_k$  — наибольшая варианта, то  $F^*(x) = 1$  при  $x > x_k$ .

Эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.



# Эмпирическая функция распределения



**Пример.** Построить эмпирическую функцию по данному распределению выборки:

варианты $x_i$	2	6	10
частоты $n_i$	12	18	30.

**Решение.** Найдем объем выборки:  $12 + 18 + 30 = 60$ .

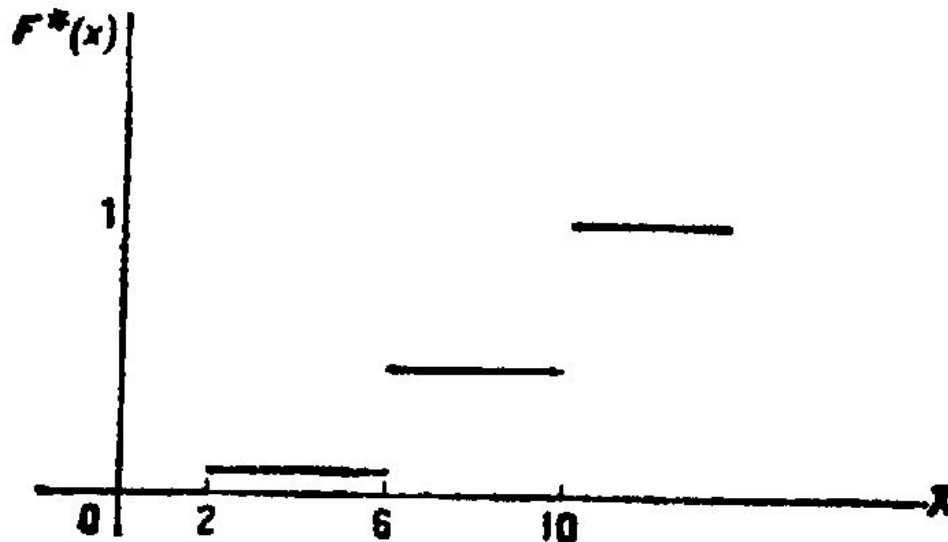
1. Наименьшая варианта равна 2, следовательно,  $F^*(x)=0$  при  $x \leq 2$ .
2. Значение  $X < 6$ , а  $x_1 = 2$ , наблюдалось 12 раз   
 $F^*(x) = 12/60 = 0,2$  при  $2 < x \leq 6$ .
3. Значения  $X < 10$ , а именно  $x_1 = 2$  и  $x_2 = 6$ , наблюдались  $12 + 18 = 30$  раз   
 $F^*(x) = 30/60 = 0,5$  при  $6 < x \leq 10$ .
4. Так как  $x=10$  - наибольшая варианта, то  $F(x)=1$  при  $x > 10$ .

# Эмпирическая функция распределения



Искомая эмпирическая функция

$$F^*(x) = \begin{cases} 0 & \text{при } x \leq 2, \\ 0,2 & \text{при } 2 < x \leq 6, \\ 0,5 & \text{при } 6 < x \leq 10, \\ 1 & \text{при } x > 10. \end{cases}$$



# Эмпирическая функция распределения



При изучении вариационных рядов наряду с понятием частоты используется понятие накопленной частоты (обозначаем  $n_i^{\text{нак}}$  ).

**Накопленная частота** показывает, сколько наблюдалось вариантов со значением признака, меньшим  $x$ .

Отношение накопленной частоты  $n_i^{\text{нак}}$  к общему числу наблюдений  $n$  называется **накопленной частотой**

$$w_i^{\text{нак}} = \frac{n_i^{\text{нак}}}{n}$$

Накопленные частоты (частоты) для каждого интервала находятся последовательным суммированием частот (частостей) всех предшествующих интервалов, включая данный.

# Эмпирическая функция распределения



<i>i</i>	<i>Выработка в отчетном году в процентах к предыдущему x</i>	<i>Частота (количество рабочих) n<sub>i</sub></i>	<i>Частость (доля рабочих)</i> $w_i = \frac{n_i}{n}$	<i>Накопленная частота</i> $n_i^{\text{нак}}$	<i>Накопленная частость</i> $w_i^{\text{нак}} = \frac{n_i^{\text{нак}}}{n}$
1	94,0—100,0	3	0,03	3	0,03
2	100,0—106,0	7	0,07	10	0,10
3	106,0—112,0	11	0,11	21	0,21
4	112,0—118,0	20	0,20	41	0,41
5	118,0—124,0	28	0,28	69	0,69
6	124,0—130,0	19	0,19	88	0,88
7	130,0—136,0	10	0,10	98	0,98
8	136,0—142,0	2	0,02	100	1,00
	$\Sigma$	100	1,00	—	—

# Эмпирическая функция распределения

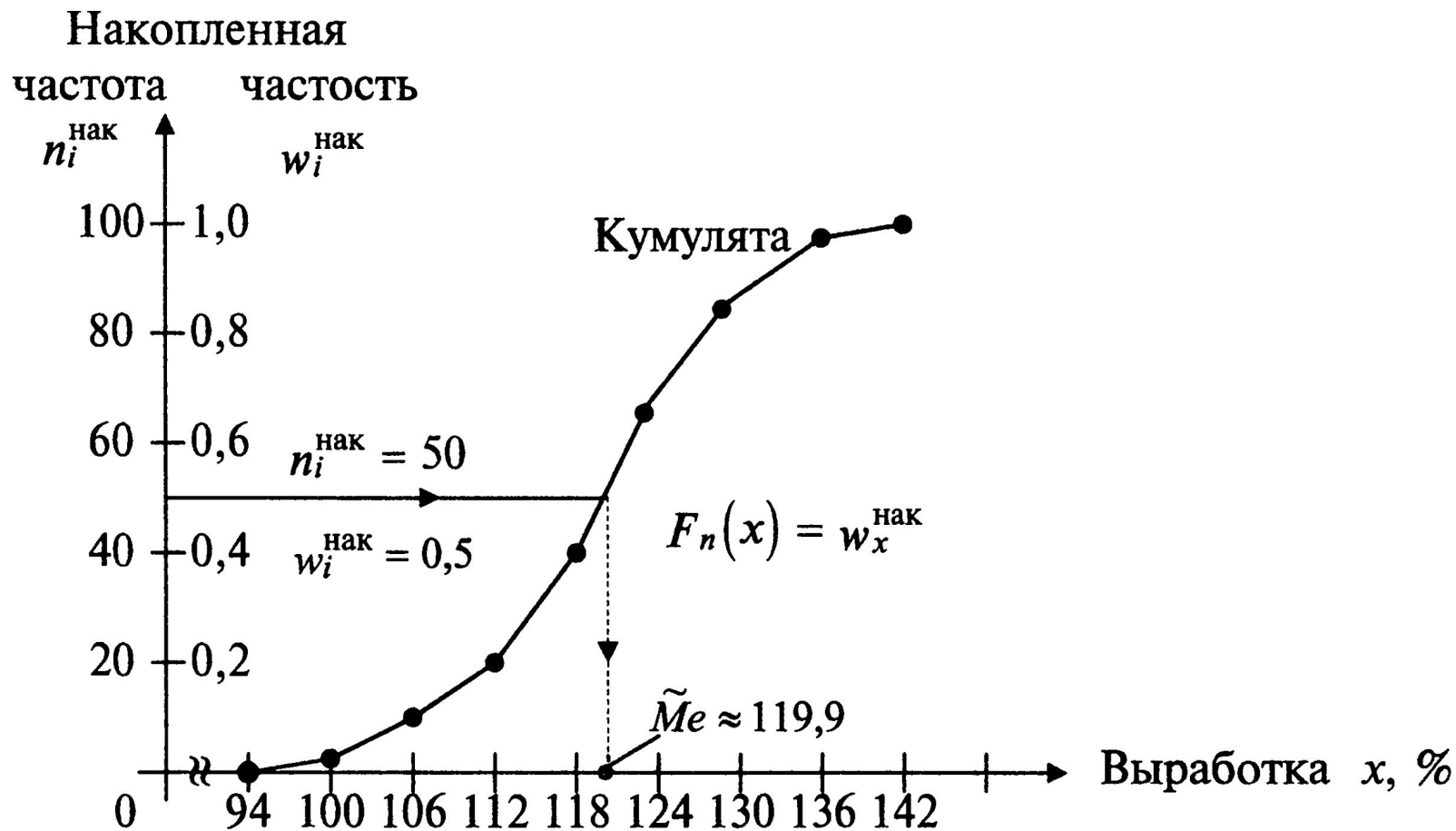


**Кумулятивная кривая** (кумулята) — кривая накопленных частот (частостей).

Для дискретного ряда кумулята представляет ломаную, соединяющую точки  $(x_i; n_i^{\text{нак}})$  или  $(x_i; w_i^{\text{нак}})$ ,  $i = 1, 2, \dots, k$ .

Для интервального вариационного ряда ломаная начинается с точки, абсцисса которой равна началу первого интервала, а ордината — накопленной частоте (частости), равной нулю. Другие точки этой ломаной соответствуют концам интервалов.

# Эмпирическая функция распределения



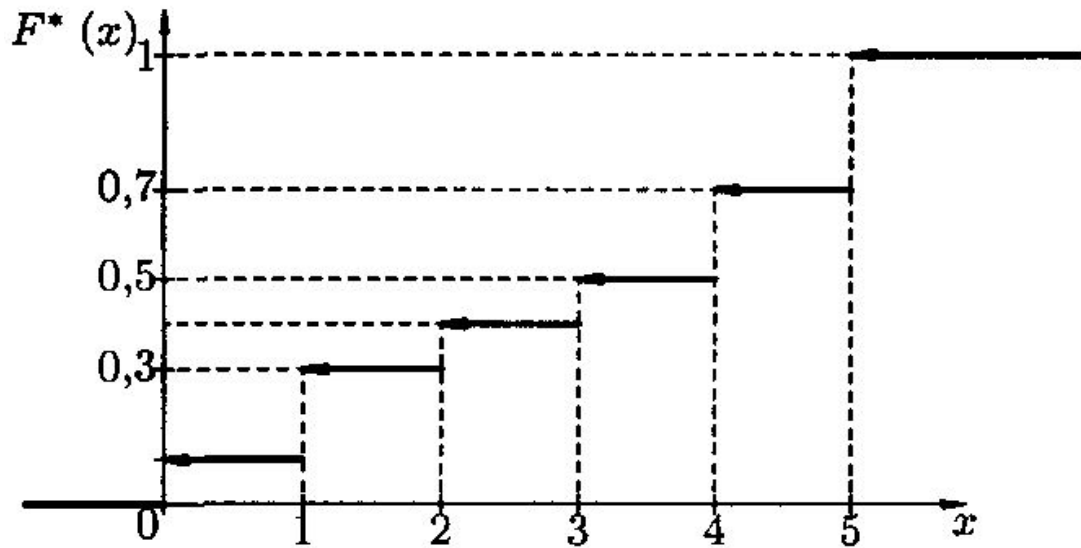
# Эмпирическая функция распределения

**Пример.** Построить эмпирическую функцию по данному распределению выборки:

$x_i$	0	1	2	3	4	5
$n_i$	1	2	1	1	2	3

**Ответ:**

$$F^*(x) = \begin{cases} 0, & \text{при } x \leq 0, \\ 0,1, & \text{при } 0 < x \leq 1, \\ 0,3, & \text{при } 1 < x \leq 2, \\ 0,4, & \text{при } 2 < x \leq 3, \\ 0,5, & \text{при } 3 < x \leq 4, \\ 0,7, & \text{при } 4 < x \leq 5, \\ 1, & \text{при } 5 < x. \end{cases}$$





# Графическое изображение статистического распределения



Для графического изображения вариационных рядов наиболее часто используются полигон, гистограмма, кумулятивная кривая.

**Полигоном частот** называют ломаную, отрезки которой соединяют точки  $(x_1; n_1)$ ,  $(x_2; n_2)$ , ...,  $(x_k; n_k)$ .

Для построения полигона частот на оси абсцисс откладывают варианты  $x_i$ , а на оси ординат— соответствующие им частоты  $n_i$ . Точки  $(x_i; n_i)$  соединяют отрезками прямых и получают полигон частот.

# Графическое изображение статистического распределения



**Полигоном относительных частот** называют ломаную, отрезки которой соединяют точки  $(x_1; w_1)$ ,  $(x_2; w_2)$ , ...  $(x_k; w_k)$ .

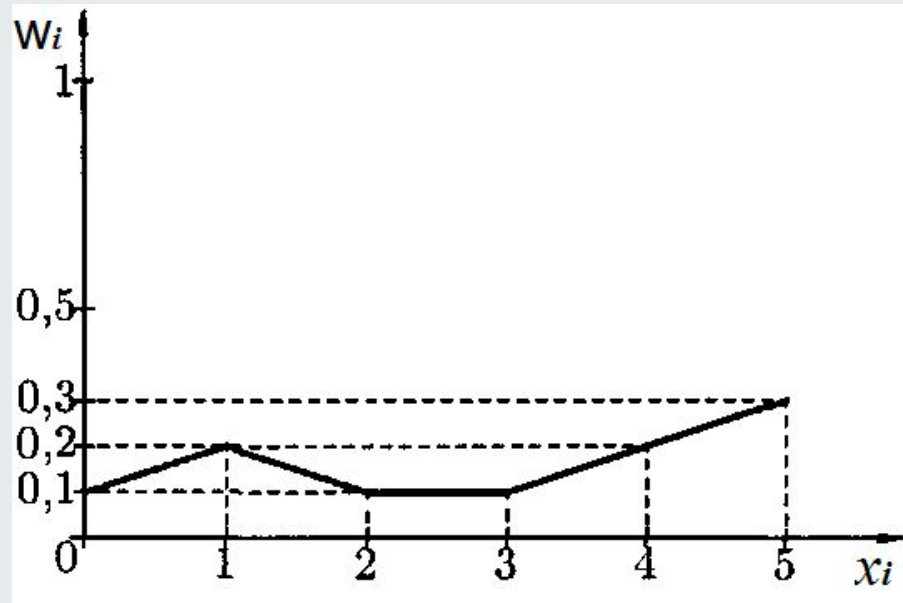
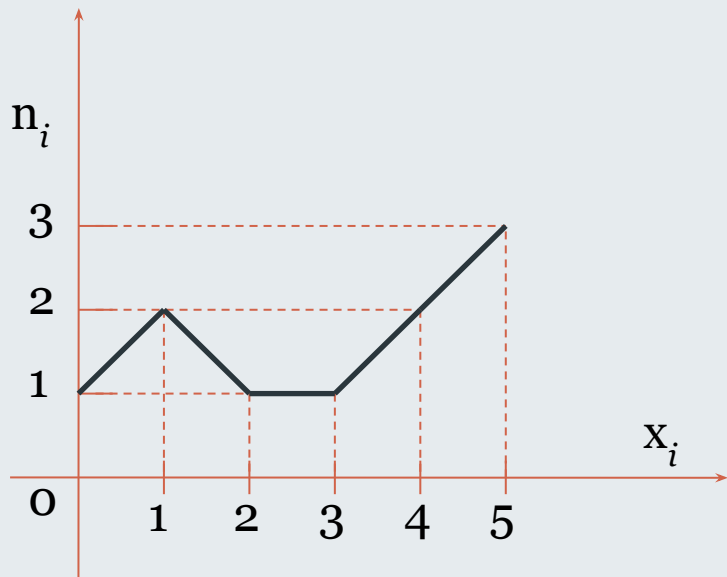
Для построения полигона относительных частот на оси абсцисс откладывают варианты  $x_i$ , а на оси ординат— соответствующие им относительные частоты  $w_i$ . Точки  $(x_i; w_i)$  соединяют отрезками прямых и получают полигон относительных частот.

# Графическое изображение статистического распределения



**Пример.** Для вариационного ряда построить полигон частот и полигон относительных частот.

$x_i$	0	1	2	3	4	5
$n_i$	1	2	1	1	2	3



# Графическое изображение статистического распределения



В случае непрерывного признака целесообразно строить гистограмму.

**Гистограммой частот** называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной  $h$ , а высоты равны отношению  $\frac{n_i}{h}$  (плотность частоты).

# Графическое изображение статистического распределения



**Пример.** Построить гистограммы частот и относительных частот распределения:

$x_i$	145-149	150-154	155-159	160-164	165-169	170-174	175-179	180-184
$n_i$	1	9	14	8	7	6	3	2
$w_i$	1/50	9/50	14/50	8/50	7/50	6/50	3/50	2/50

