

**Теория вероятностей
и
математическая статистика**



Математическая статистика -

- это наука и методах сбора, систематизации и обработке данных научных исследований с целью выявления существующих в них закономерностях.

Выборочный метод

Генеральная совокупность (ГС)

— это вся подлежащая изучению совокупность объектов.

Т.е., совокупность всех мыслимых наблюдений, которые могли быть получены при данном комплексе условий.

Генеральная совокупность

аналогична случайной величине X , поэтому она обладает законом распределения, математическим ожиданием, дисперсией и т. д.

Основная задача МС-

- исследовать ГС статистически, т.е., определение ее основных характеристик, закона распределения и т.п.
- Однако полное исследование ГС либо не представляется возможным, либо неэкономично. Поэтому из нее делают **выборку**, т.е. подвергают исследованию только некоторые объекты ГС.

Выборка

Выборкой называется множество значений x_1, x_2, \dots, x_n ГС, предназначенное для непосредственного исследования.

Количество элементов выборки n – называется **объемом выборки**.

Выборка

Выборка бывает

дискретной и непрерывной
повторной и бесповторной,
одномерной и многомерной.

Выборка

Семинарское занятие - пример **дискретной, повторной** выборки.

Диспансеризация спортсменов раз в полгода, измерение антропометрических данных (рост, вес и т. д.) – пример **непрерывной бесповторной** выборка.

Выборка

Результат измерения роста 20 человек.

176,5; 163,3; 173,4; 182,1; 152,3;
162,2; 200,0; 194,1; 154,4; 170,8;
160,0; 173,3; 167,6; 168,2; 166,1;
176,6; 175,9; 165,8; 151,5; 178,6

Бесповторная непрерывная
выборка объема $n = 20$.

Суть выборочного метода

заключается в том, что на основании выборочных данных делается вывод о генеральной совокупности в целом.

Репрезентативность

Для того, чтобы оценки полученные по выборочным данным были достоверными, необходимо, чтобы выборка была **репрезентативной, (организованной случайным образом)** - каждый элемент ГС должен иметь равную вероятность попасть в выборку.

Крупномасштабный почтовый опрос престижного американского журнала *«The Literary Digest»*, проведенный во время предвыборной кампании 1936 года, занимает важное место в истории эмпирической социологии. Исследование должно было определить, кого хотят видеть американцы своим президентом: Франклина Д. Рузвельта, кандидата от демократической партии, баллотировавшегося на второй срок, или Элфа Лэндона, кандидата республиканской партии.

Итоги электорального опроса не оставляли никаких сомнений: безоговорочную победу на предстоящих выборах одерживал республиканец Лэндон, за которого собирались голосовать 55% респондентов. Рузвельта поддержали участники опроса в количестве только 41%.

Результат выборов стал полной неожиданностью для «The Literary Digest»: действовавший президент Ф. Д. Рузвельт получил 61% голосов избирателей, в то время как его соперник — 37%.

Для составления списка респондентов были использованы **телефонные книги и регистрационные списки владельцев автомобилей** каждого территориального округа во всех сорока восьми штатах.

Самой распространенной точкой зрения на причину неудачи опроса «The Digest» уже более семидесяти лет остается мнение о некорректной процедуре составления выборки. Респонденты были отобраны, главным образом, из обширной картотеки журнала, которую владельцы создали с целью привлечения новых подписчиков. **Как телефон, так и автомобиль в тридцатых годах прошлого века являлись определенным показателем уровня дохода.** Таким образом, необъективно большую долю в выборке представляли состоятельные американцы. Учитывая, что во время выборов 1936 года существовала тесная взаимосвязь между размером доходов и партийными предпочтениями, результат можно было прогнозировать еще до начала опроса. В то же время процедура исключала значительную часть электората — бедняков, которые предположительно обеспечили победу Рузвельта: **многие сторонники президента не участвовали в опросе только потому, что не имели автомобиля и телефона.**

Варианты и частоты

Наблюдаемое значение признака в статистике называется

вариантой и обозначается X_i .

Одна и та же варианта в выборке может встречаться несколько раз – это число называется

частотой n_i .

Относительная частота

$$w_i = n_i / n.$$

Вариационный ряд

Если все значения признака записать в порядке возрастания или убывания , то такое представление выборки называется ***вариационным рядом.***

Выборка

Выборка

1 3 2 2 4 1 5 3 6 1

6 5 6 5 4 3 4 2 1 3

Упорядоченная выборка

Статистическое представление выборки

Если значения вариант
соответствуют значениям
дискретной СВ, то она
называется **дискретной**.

Статистическое представление выборки

Статистическим представлением дискретной выборки называется таблица, в первой строке которой записывают значения вариантов x_i , во второй строке значения соответствующих им частот n_i (относительных частот w_i).

Статистическое представление дискретной выборки (частоты)

X_i	X_1	X_2	...	X_k
n_i	n_1	n_2	...	n_k

Условие нормировки

$$\sum_{i=1}^k n_i = n$$

Статистическое представление дискретной выборки (относительные частоты)

X_i	X_1	X_2	...	X_k
W_i	W_1	W_2	...	W_k

Условие нормировки

$$\begin{aligned} \sum_{i=1}^k w_i &= \sum_{i=1}^k \frac{n_i}{n} = \\ &= \frac{1}{n} \sum_{i=1}^k n_i = \frac{1}{n} \cdot n = 1 \end{aligned}$$

Выборка

Упорядоченная выборка

**1, 1, 1, 1, 2, 2, 2, 3, 3, 3,
3, 4, 4, 4, 5, 5, 5, 6, 6, 6.**

x_i	1	2	3	4	5	6
n_i	4	3	4	3	3	3
w_i	4/20	3/20	4/20	3/20	3/20	3/20

Интервальное представление выборки

Если в выборке имеется большое количество различных значений признака, то ее удобно представлять в виде частичных интервалов.

Интервальное представление выборки

Частота i -го частичного интервала n_i – определяется путем подсчета объектов выборки, значения которых попали в данный интервал $[a_{i-1}; a_i)$.

Статистическое представление интервальной выборки

Частичный интервал	$a_0 - a_1$...	$a_{k+1} - a_k$
n_i	n_1	...	n_k

Пример

Результат измерения роста 20 человек ($n = 20$).

176,5; 163,3; 173,4; 182,1; 152,3;
162,2; 201,0; 194,1; 154,4; 170,8;
165,5; 173,3; 167,6; 168,2; 166,1;
176,6; 175,9; 165,8; 151,5; 178,6

Интервальное представление выборки

~~176,5; 163,3; 173,4; 182,1; 152,3; 162,2;~~
~~200,0; 194,1; 154,4; 170,8; 160,0; 173,3; 167,6;~~
~~168,2; 166,1; 176,6; 175,9; 185,8; 151,5; 178,6~~

$a_{i-1} - a_i$	150-160	160-170	170-180	180-190	190-200	200-210
n_i	3	6	7	2	1	1
w_i	3/20	6/20	7/20	2/20	1/20	1/20

Накопленные частоты

Накопленной частотой i -ой

варианты – называется количество объектов выборки, значение которых не превосходит x_i .

Накопленной частотой i -ого

интервала называется количество выборочных данных, значения которых не превышают конца этого интервала.

Накопленные частоты

Относительной накопленной частотой i -ой группы выборки – называется число

$$W_i^* = \frac{n_i^*}{n}$$

Графические представления выборки

Графически выборку можно
представить в виде: полигона,
гистограммы и кумуляты.

Полигон частот-

это ломаная линия, отрезки которой соединяют вершины (x_i, n_i) или (x_i, w_i) .

По огибающей, проведенной через вершины полигона можно сделать предположение в виде закона распределения ГС, а также определить моду.

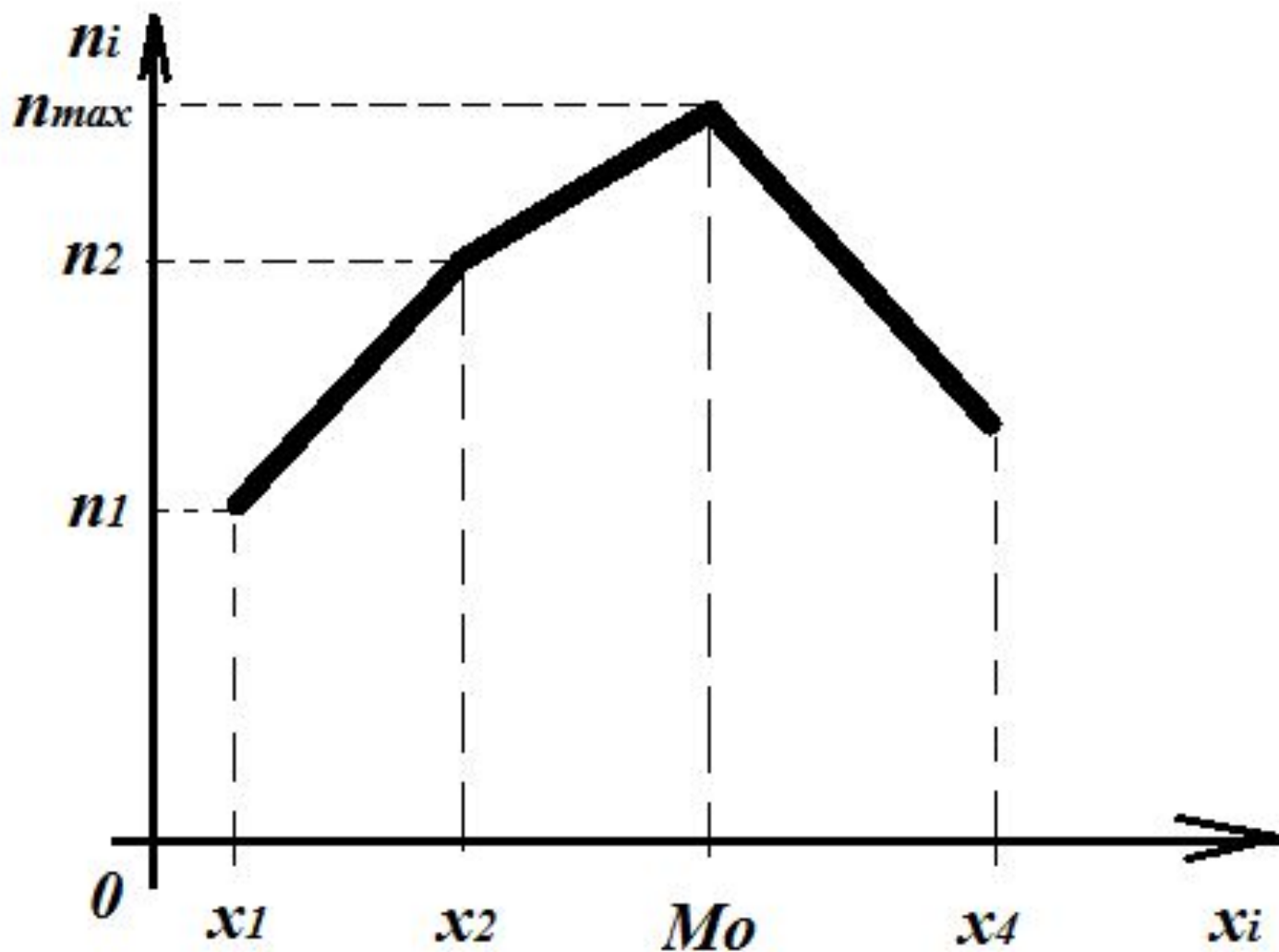


Рис.1. Полигон частот

Пример бимодального распределения

x_i	1	2	3	4	5	6
n_i	4	3	4	3	3	3

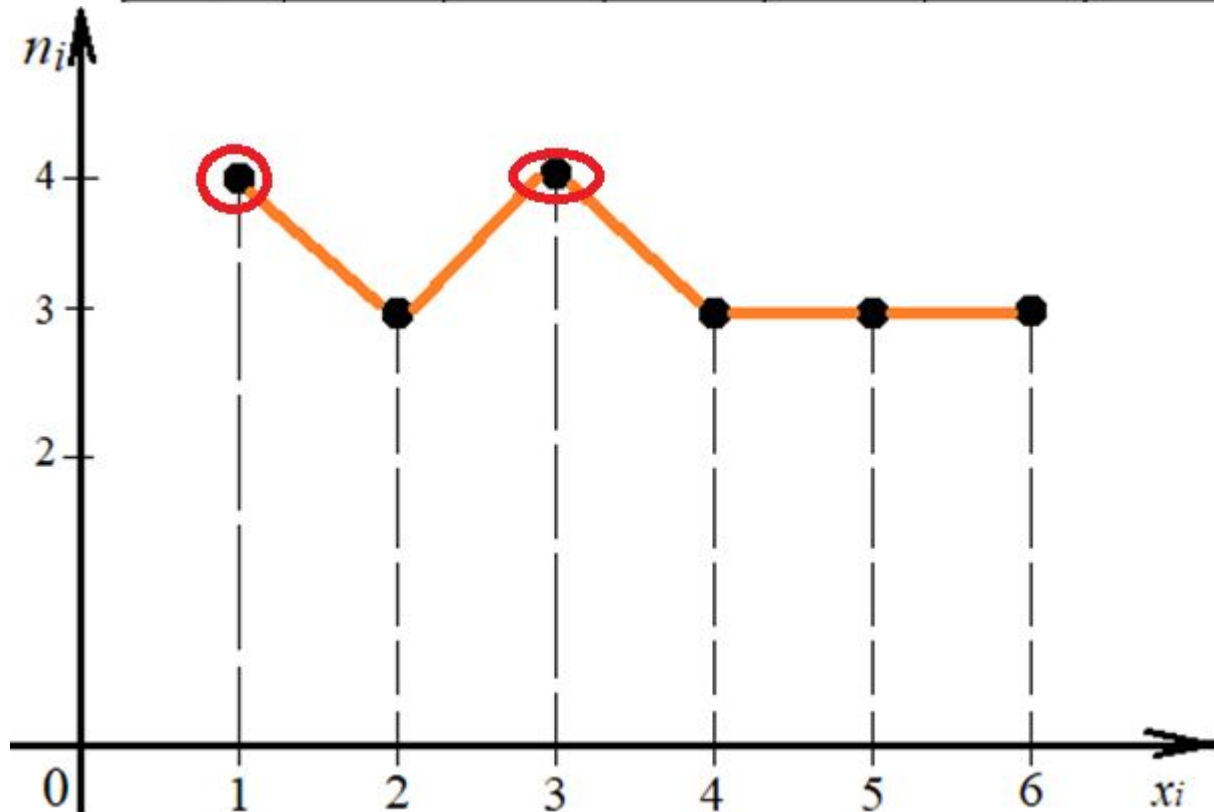


Рис. 2. Полигон частот для дискретного вариационного ряда – число очков на кости

Гистограмма частот-

это множество прямоугольников, в основании которых лежат частичные интервалы, а высоты соответствуют частоте (относительной частоте).

По ***гистограмме*** можно сделать предположение и виде закона распределения ГС и найти ***моду*** интервального ряда.

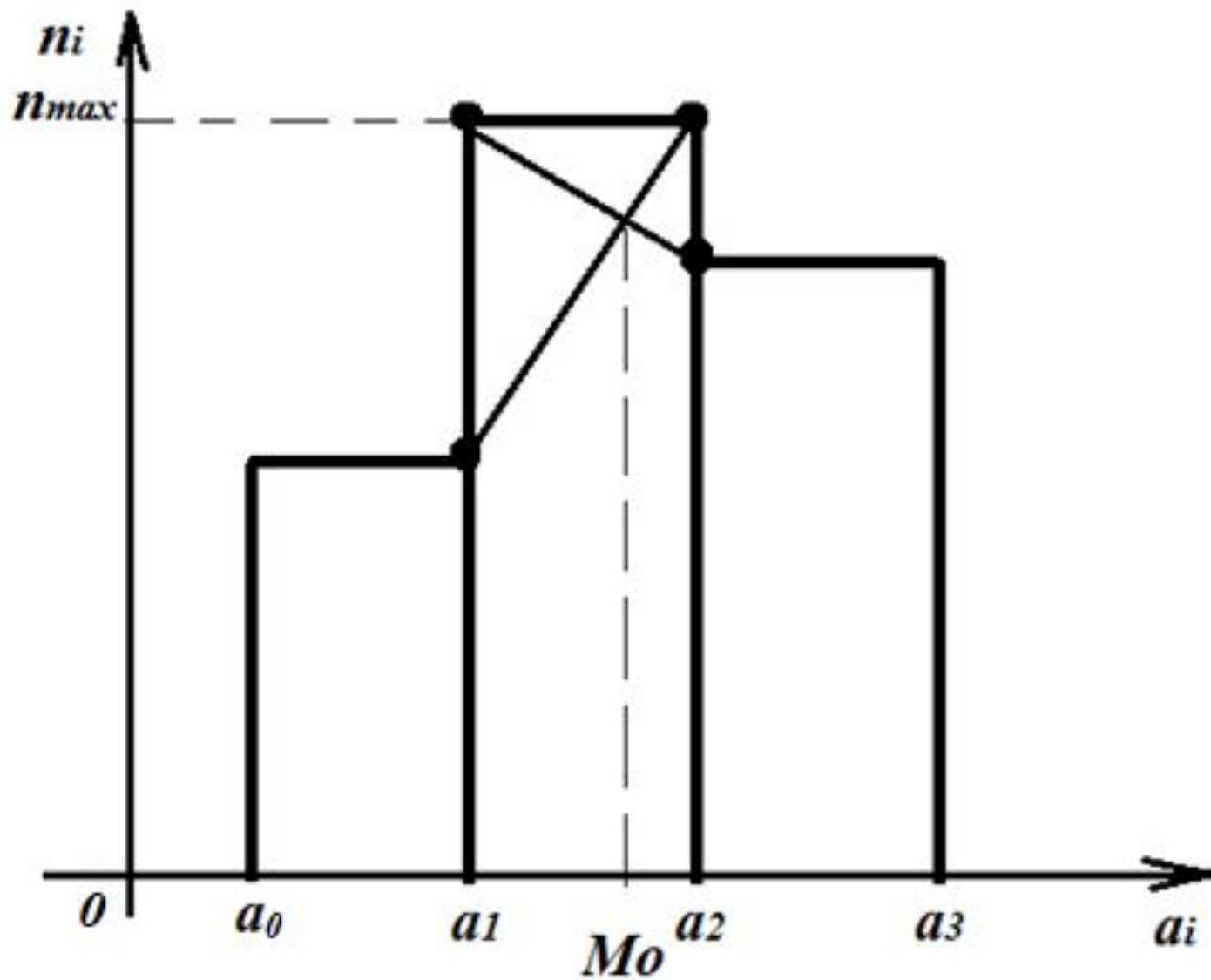


Рис.3. Гистограмма частот

Гистограмма относительных частот

При построении **гистограммы относительных частот** высоты прямоугольников соответствуют относительной частоте.

Если интервалы имеют разную длину, то по оси ординат откладывают величины частот деленные на длину i -ого интервала (h_i – длина i -го интервала):

$$\frac{n_i}{h_i} \quad \text{ИЛИ} \quad \frac{w_i}{h_i} = \frac{n_i}{nh_i}.$$

Пример построения гистограммы частот по ростовым данным

$a_{i-1} - a_i$	150-160	160-170	170-180	180-190	190-200	200-210
n_i	3	6	7	2	1	1

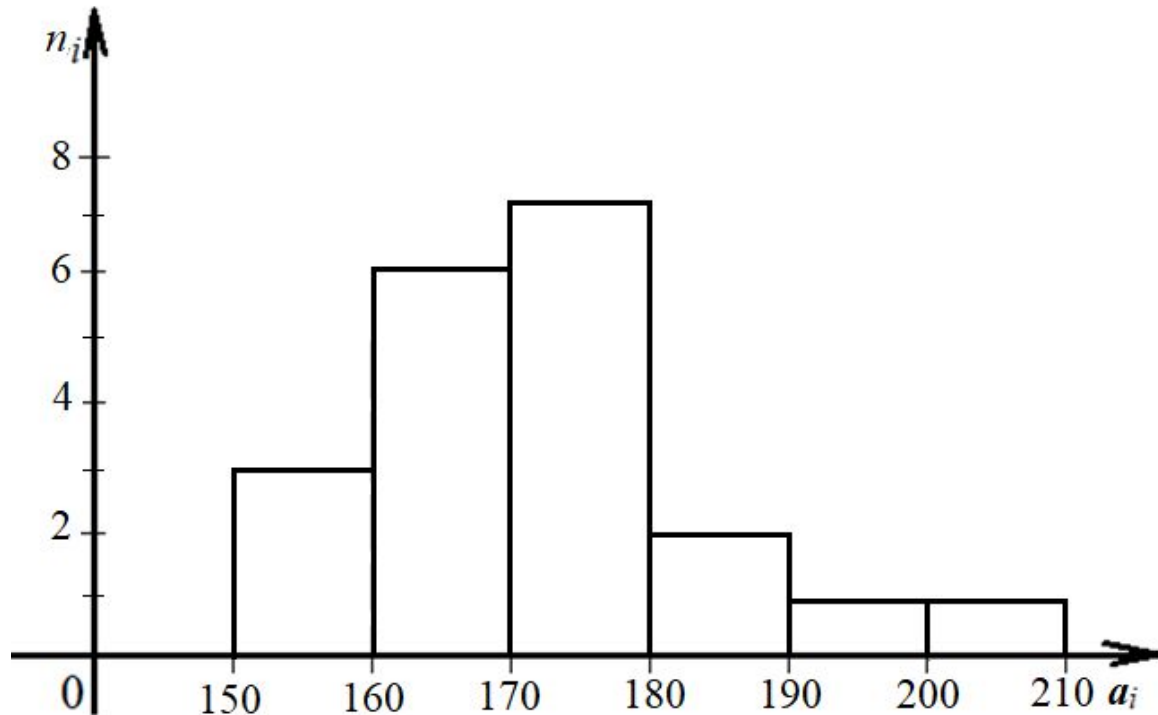


Рис. 4. Гистограмма частот

Кумулята –

это ломаная линия, отрезки которой соединяют точки (x_i, n_i^*) (значение варианты, значение накопленной частоты). По *кумуляте* можно найти *медиану выборки*

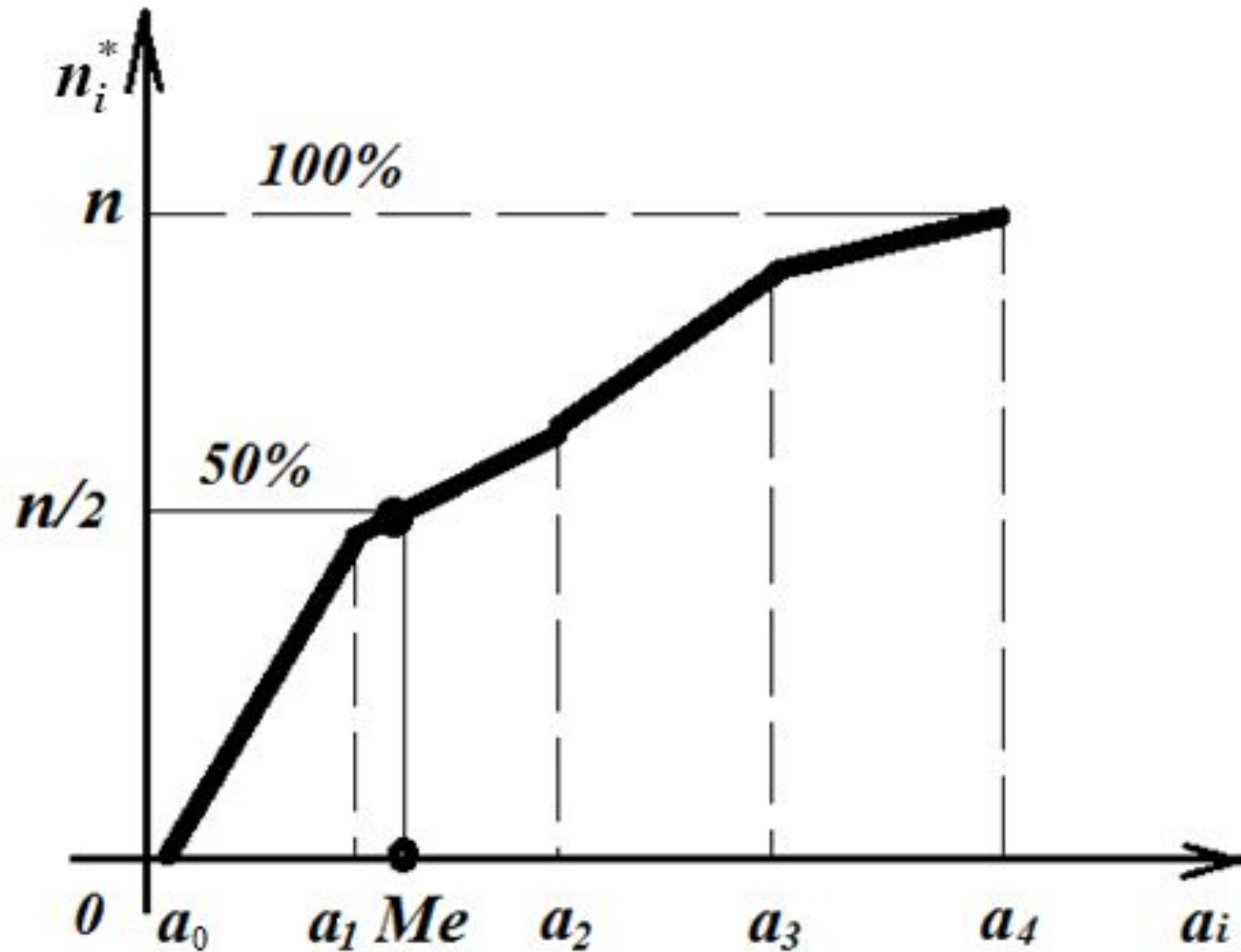


Рис. 5. Кумулята интервального ряда

Пример построения кумуляты

$a_{i-1} - a_i$	1-2	2-3	3-4	4-5
n_i	4	8	6	2

Построим дополнительную таблицу для построения кумулятивной кривой.

a_i	1	2	3	4	5
n_i^*	0	4	12	18	20

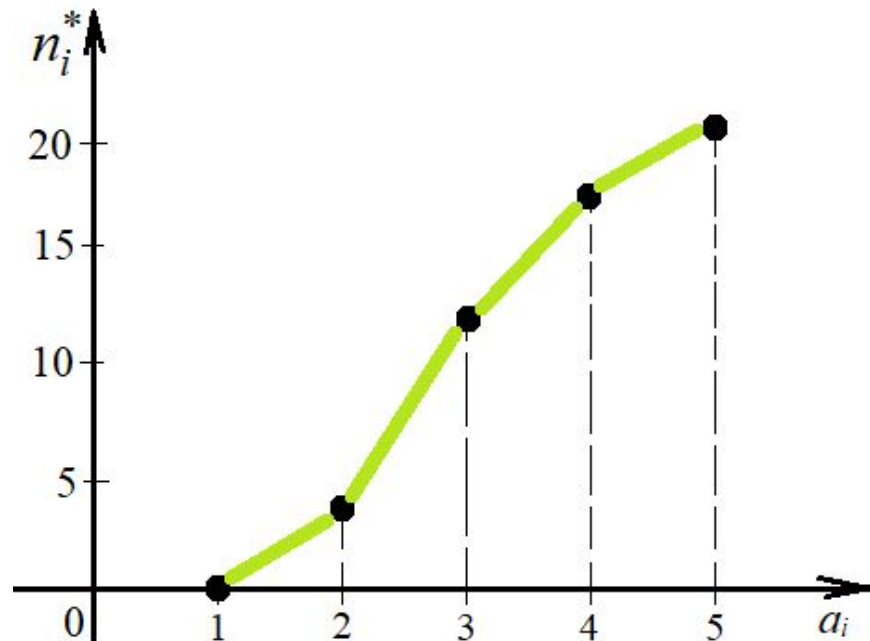


Рис. 6. Пример построения кумуляты

Эмпирическая функция распределения

Эмпирическая функция
распределения находится по
формуле:

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

Здесь n – это объем выборки; n_x -
это число выборочных данных,
строго меньших x .

Свойства функции эмпирической функции распределения

1. $0 \leq F_n(x) \leq 1$
2. неубывающая функция, то есть
$$x_1 < x_2 \Rightarrow F_n(x_1) \leq F_n(x_2)$$
3. $F_n(x) = 0, x \leq x_{min}$
$$F_n(x) = 1, x > x_{max}$$

Эмпирическая функция распределения – ступенчатая. Необходимо разбить ось на интервалы точками x_i , и воспользоваться формулой для каждого интервала в отдельности.

x_i	2	4	6	8	10
n_i	4	6	5	3	1

Найдем объем выборки.

$$n = \sum_{i=1}^k n_i = 4 + 6 + 5 + 3 + 1 = 19$$

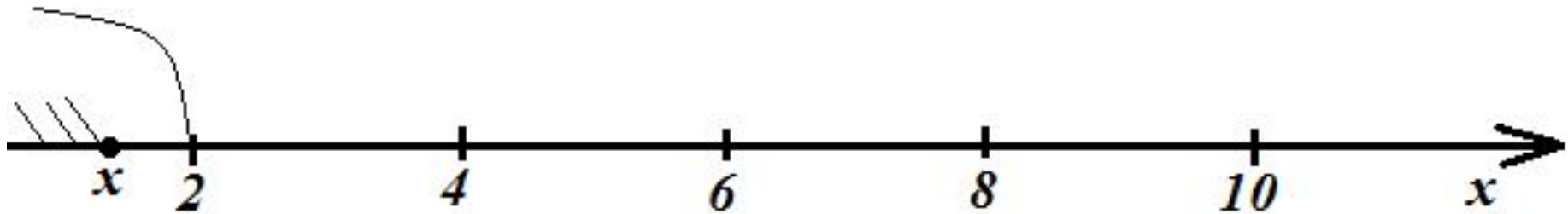
x_i	2	4	6	8	10
n_i	4	6	5	3	1

Эмпирическая функция
распределения – ступенчатая
функция.

Разобьем ось на интервалы
точками 2, 4, 6, 8, 10, и применим
данную формулу для каждого
интервала в отдельности.

x_i	2	4	6	8	10
n_i	4	6	5	3	1

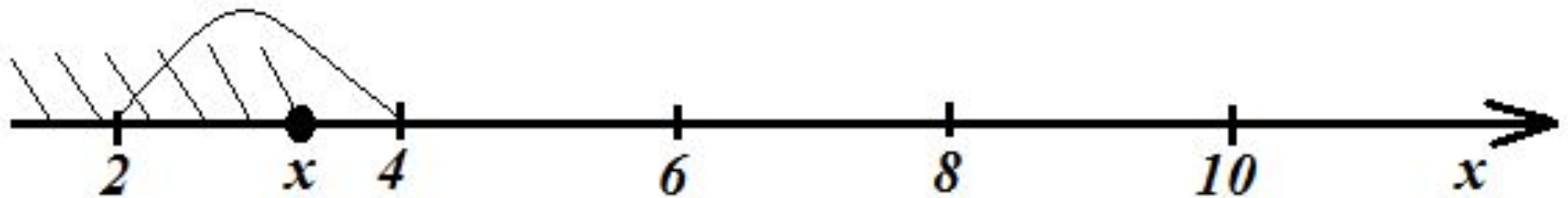
1) $x \leq 2$



$$F_n(x) = \frac{n_x}{n} = \frac{0}{19} = 0$$

x_i	2	4	6	8	10
n_i	4	6	5	3	1

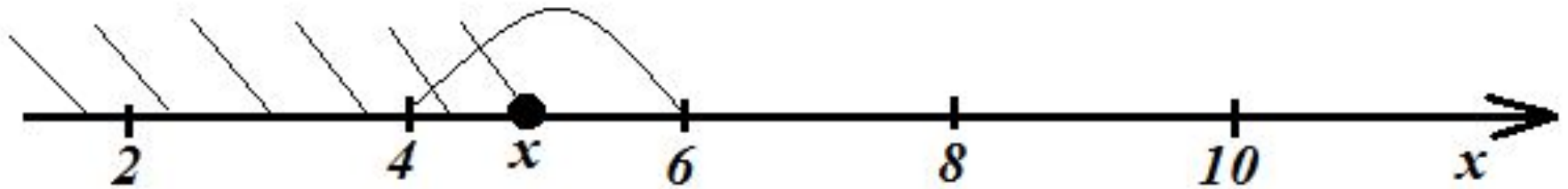
$$2) 2 < x \leq 4$$



$$F_n(x) = \frac{n_x}{n} = \frac{4}{19}$$

x_i	2	4	6	8	10
n_i	4	6	5	3	1

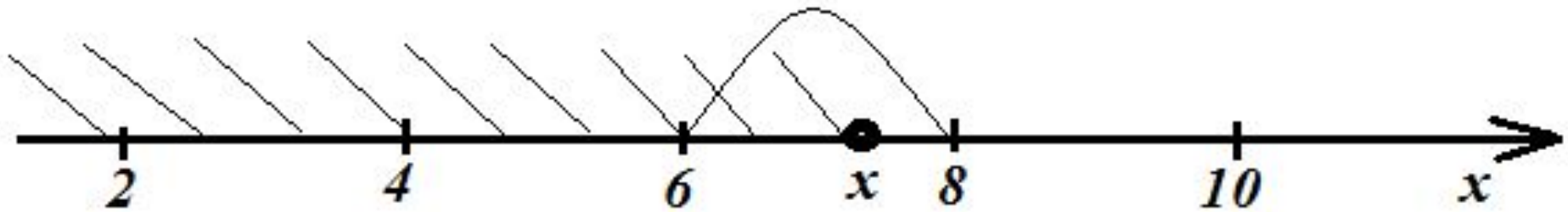
$$3) \quad 4 < x \leq 6$$



$$F_n(x) = \frac{n_x}{n} = \frac{4 + 6}{19} = \frac{10}{19}$$

x_i	2	4	6	8	10
n_i	4	6	5	3	1

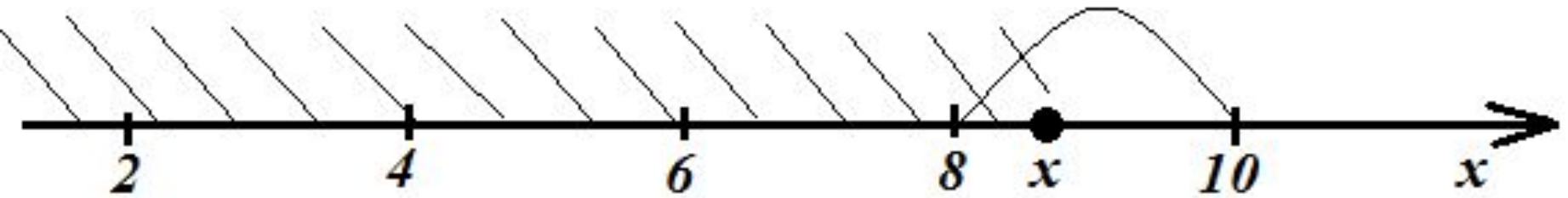
$$4) \quad 6 < x \leq 8$$



$$F_n(x) = \frac{n_x}{n} = \frac{4 + 6 + 5}{19} = \frac{10 + 5}{19} = \frac{15}{19}$$

x_i	2	4	6	8	10
n_i	4	6	5	3	1

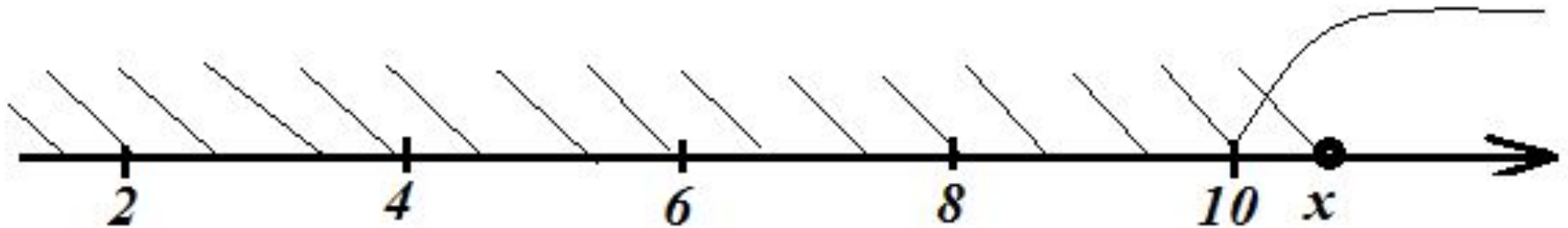
$$5) \quad 8 < x \leq 10$$



$$F_n(x) = \frac{n_x}{n} = \frac{15 + 3}{19} = \frac{18}{19}$$

x_i	2	4	6	8	10
n_i	4	6	5	3	1

6) $x > 10$



$$F_n(x) = \frac{n_x}{n} = \frac{18+1}{19} = \frac{19}{19} = 1$$

$$F_n(x) = \begin{cases} 0, & x \leq 2; \\ \frac{4}{19}, & 2 < x \leq 4; \\ \frac{10}{19}, & 4 < x \leq 6; \\ \frac{15}{19}, & 6 < x \leq 8; \\ \frac{18}{19}, & 8 < x \leq 10; \\ 1, & x > 10. \end{cases}$$

x_i	2	4	6	8	10
n_i	4	6	5	3	1

Нахождение значений функции распределения можно осуществить с помощью таблицы:

x_i	2	4	6	8	10
n_i^*	4	10	15	18	19
w_i^*	4/19	10/19	15/19	18/19	1

Затем, используя свойства эмпирической функции распределения, записывают формулу.

