Тема

Основы математической статистики

К.Ф. Гаусс
(1777-1855)

Математика – царица наук!

1795 г. - на основе теории вероятностей исследовал и обосновал ***метод наименьших квадратов***
С этой работы **математическая статистика** начинается как наука

I. Основные понятия

Статистика
– это область науки, изучающая сбор, анализ и интерпретацию данных.

От лат. *status* - «состояние, положение вещей»

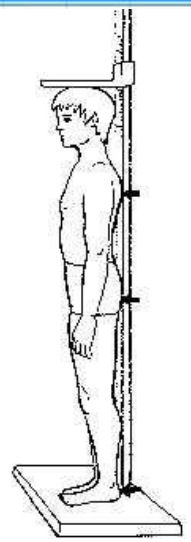
1746 г. – Г.Ахенваль ввел термин в науку

Г. Ахенвалль (1719—1772)

Пример 1.

В девятых классах «А» и «Б» измерили рост 50 учеников.
Получились следующие результаты:

162, 168, 157, 176, 185, 160, 162, 158, 181, 179, 164, 176, 177, 180, 181, 179, 175, 180, 176, 165, 168, 164, 179, 163, 160, 176, 162, 178, 164, 190, 181, 178, 168, 165, 176, 178, 185, 179, 180, 168, 160, 176, 175, 177, 176, 165, 164, 177, 175, 181.



Недостатки данной информации:

- Трудно «читается»
- Не наглядна
- Занимает много места

Выход:

— *преобразовать* данные, получить небольшое количество *характеристик* начальной информации.

⇒ Одна из **основных задач** статистики: ***обработка информации***.

Другие задачи статистики:

- получение и хранение информации
- выработка различных прогнозов
- оценка их достоверности

Новый термин	Простое описание	Более научный термин	Определение
Общий ряд данных	То, откуда выбирают	Генеральная совокупность	Множество всех в принципе возможных результатов измерения
Выборка	То, что выбрали	Статистическая выборка, статистический ряд	Множество результатов, реально полученных в данном измерении
Варианта	Значение одного из результатов измерения	Варианта	Одно из значений элементов выборки
Ряд данных	Значения всех результатов измерения, перечисленные по порядку	Вариационный ряд	Упорядоченное множество всех вариант

Пример 1.
В девятых классах «А» и «Б» измерили рост 50 учеников. Получились следующие результаты:
162, 168, 157, 176, 185, 160, 162, 158, 181, 179, 164, 176, 177, 180, 181, 179, 175, 180, 176, 165, 168, 164, 179, 163, 160, 176, 162, 178, 164, 190, 181, 178, 168, 165, 176, 178, 185, 179, 180, 168, 160, 176, 175, 177, 176, 165, 164, 177, 175, 181.

1. С некоторым запасом можно считать, что рост девятиклассника находится в пределах от 140 до 210 см.
⇒
Общий ряд данных этого измерения: **140; 141; 142; …; 208; 209; 210**

2. ***Выборка*** — это данные реального измерения роста (выписаны выше)

3. ***Варианта*** — это любое из чисел выборки

4. ***Ряд данных*** — все реальные результаты измерения, выписанные в определенном порядке *без повторений*, например, по возрастанию:
157; 158; 160; 162; 163; 164; 165; 168; 175; 176; 177; 178; 179; 180; 181; 185; 190

Пример 2.

30 абитуриентов на четырех вступительных экзаменах набрали в сумме такие количества баллов (оценки на экзаменах выставлялись по пятибалльной системе):

20; 19; 12; 13; 16; 17; 15; 14; 16; 20; 15; 19; 20; 20; 15; 13; 19; 14; 18; 17; 12; 14; 12; 17; 18; 17; 20; 17; 16; 17.

Составьте общий ряд данных, выборку из результатов, стоящих на четных местах и соответствующий ряд данных.

Решение:

1) После получения двойки дальнейшие экзамены не сдаются, поэтому сумма баллов не может быть меньше 12 (12 — это 4 «тройки»).

   ⇒ **Общий ряд данных:** 12; 13; 14; 15; 16; 17; 18; 19; 20

2) **Выборка** состоит из 15 результатов: 19; 13; 17; 14; 20; 19; 20; ..., расположенных на четных местах

3) **Ряд данных:** 13; 14; 17; 19; 20

Составим ***таблицу распределения*** выборки и частот выборки

Пример 2.

30 абитуриентов на четырех вступительных экзаменах набрали в сумме такие количества баллов (оценки на экзаменах выставлялись по пятибалльной системе):

20; **19**; 12; **13**; 16; **17**; 15; **14**; 16; **20**; 15; **19**; 20; **20**; 15; **13**; 19; **14**; 18; **17**; 12; **14**; 12; **17**; 18; **17**; 20; **17**; 16; **17**.

Составьте общий ряд данных, выборку из результатов, стоящих на четных местах и соответствующий ряд данных.

Решение:

Составим ***таблицу распределения*** выборки и часто выборки

Варианта	13	14	17	19	20	Всего: 5 вариант
Кратность варианты	2	3	6	2	2	Сумма =15 (объем выборки)
Частота варианты	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{6}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	Сумма =1 (так всегда)

$$\text{Частота варианты} = \frac{\text{Кратность варианты}}{\text{Объём выборки}}$$

Иногда измеряется в процентах (·100%)

II. Графическое представление информации

Таблицы образуют «мостик», по которому от выборок данных можно перейти к функциям и их графикам.

Варианта	13	14	17	19	20	Всего: 5 вариант
Кратность варианты	2	3	6	2	2	Сумма =15 (объем выборки)
Частота варианты	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{6}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	Сумма =1 (так всегда)

Алгоритм получения графика распределения выборки:

1)Отложить по оси абсцисс значения из первой строки таблицы
2)Отложить по оси ординат — значения из ее второй строки
3)Построить соответствующие точки в координатной плоскости
4)Построенные точки для наглядности соединить отрезками

Примечание:

Если заменить вторую строку таблицы ее третьей строкой, то получится **график распределения частот выборки**.

Термин «**график распределения частот выборки**» заменяют кратким — **многоугольник частот** или **полигон частот**. (polygon – многоугольник)

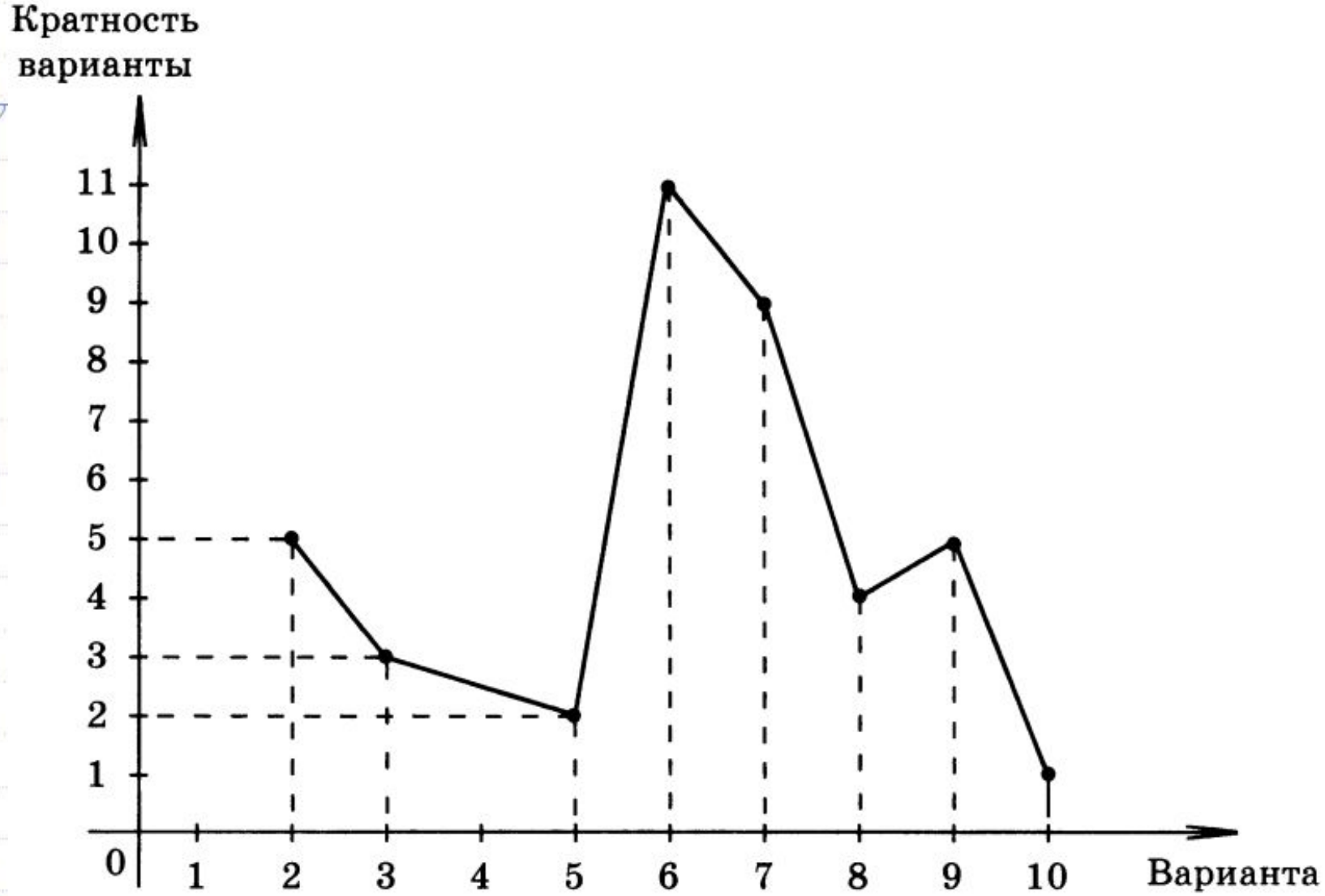
Пример 3.
Постройте график распределения и многоугольник частот для следующих результатов письменного экзамена по математике:

6	7	7	8	9	2	10	6	5	6
7	3	7	9	9	2	3	2	6	6
6	7	8	8	2	6	7	9	7	5
9	8	2	6	6	3	7	7	6	6

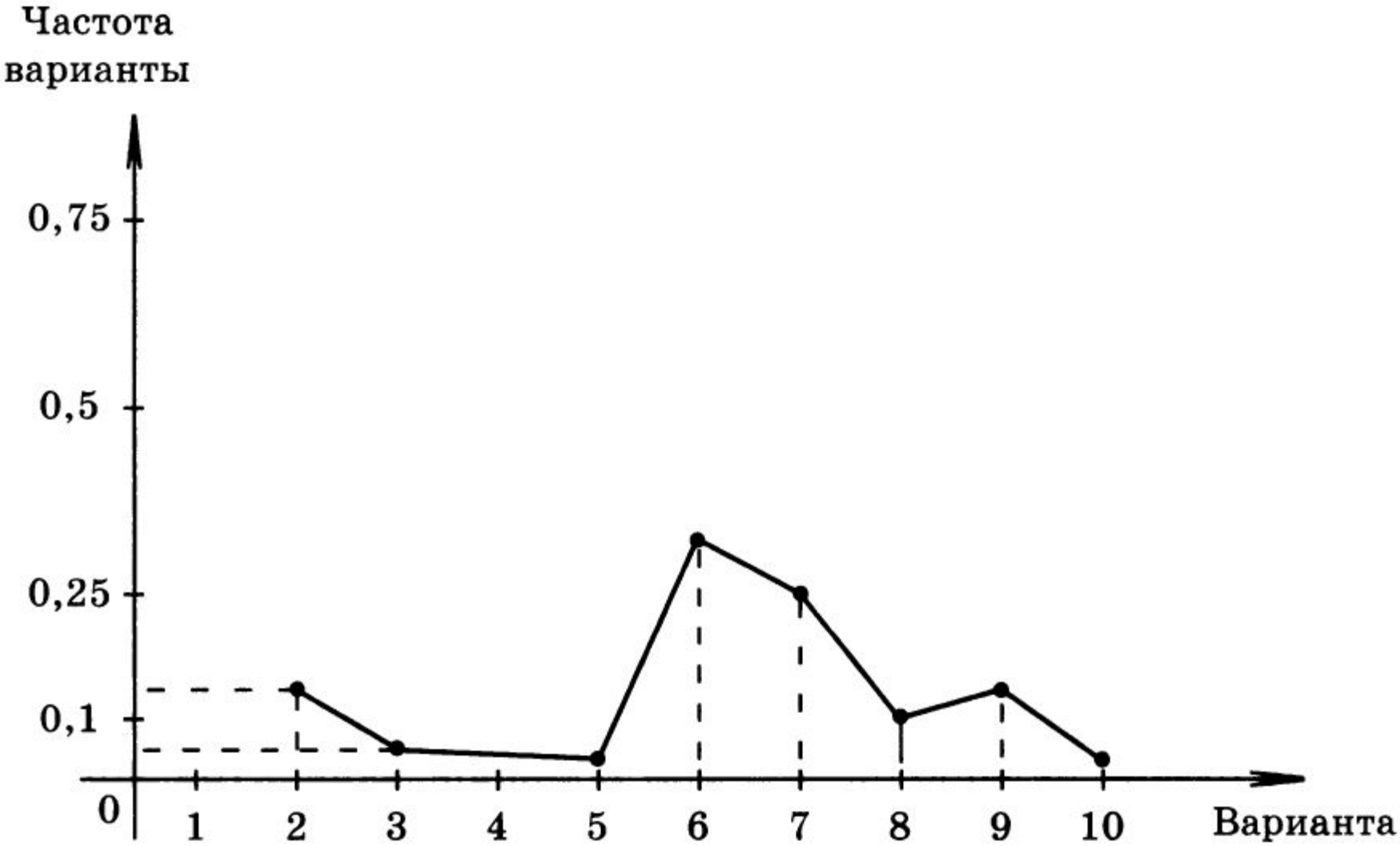
Решение:
Выборка объема 40.
Ряд данных — **2; 3; 5; 6; 7; 8; 9; 10**
Составим таблицу и построим график

Варианта	**2**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**Всего 8 вариант**
Кратность варианты	**5**	**3**	**2**	**11**	**9**	**4**	**5**	**1**	**Сумма = 40**
Частота варианты	**0,125**	**0,075**	**0,05**	**0,275**	**0,225**	**0,1**	**0,125**	**0,025**	**Сумма = 1**
Частота (%) варианты	**12,5**	**7,5**	**5**	**27,5**	**22,5**	**10**	**12,5**	**2,5**	**Сумма = 100%**

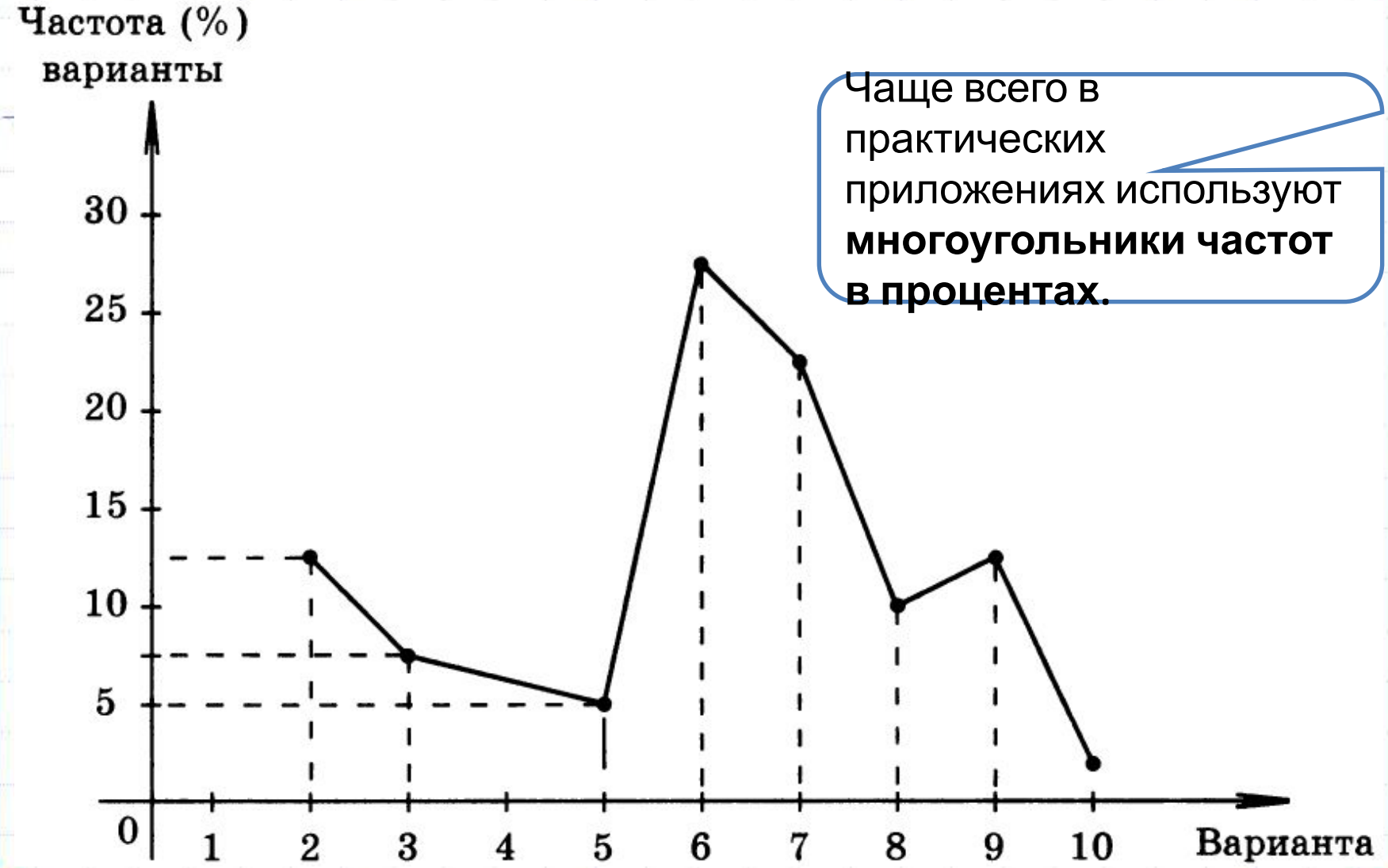
Многоугольник распределения кратностей



Многоугольник распределения частот



Многоугольник распределения частот (%)



Чаще всего в практических приложениях используют **многоугольники частот в процентах.**

Пример 3.

Постройте график распределения и многоугольник частот для следующих результатов письменного экзамена по математике:

6	7	7	8	9	2	10	6	5	6
7	3	7	9	9	2	3	2	6	6
6	7	8	8	2	6	7	9	7	5
9	8	2	6	6	3	7	7	6	6

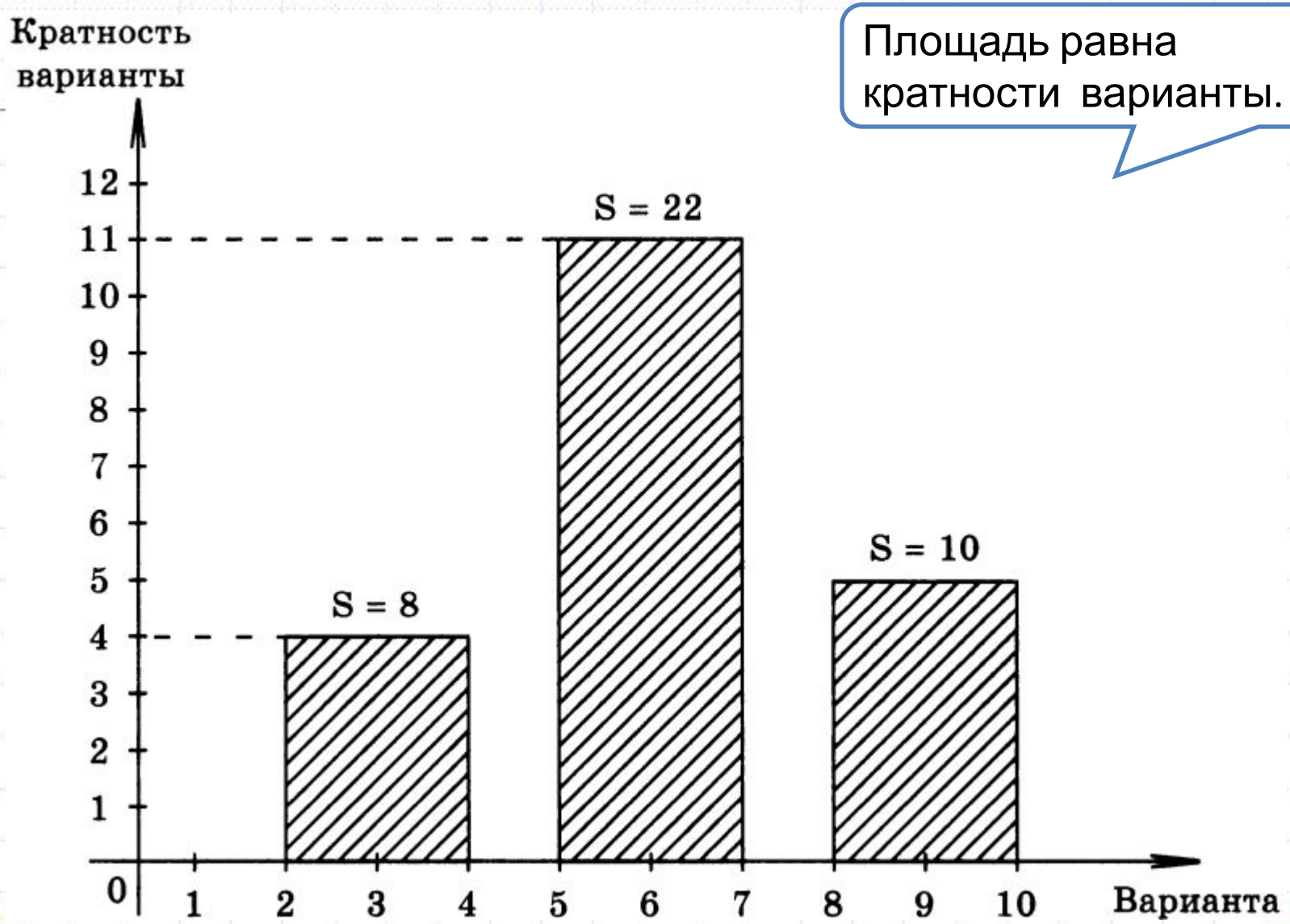
Построение гистограмм (столбчатых диаграмм) распределения:

Разбиваем промежуток между самой маленькой и самой большой вариантой на участки:

- «Плохие» оценки $\in [2; 4]$
- «Средние» оценки $\in [5; 7]$
- «Хорошие» оценки $\in [8; 10]$

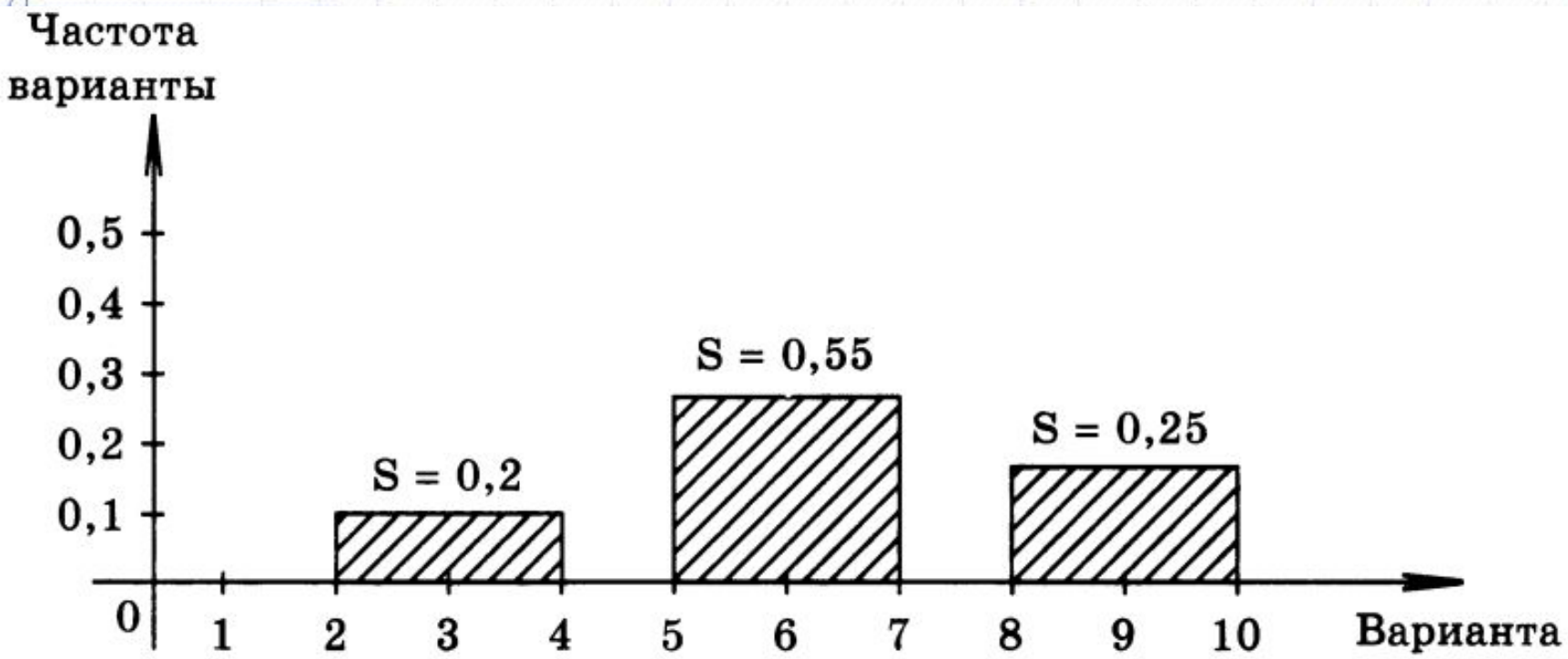
Варианта	«Плохие»	«Средние»	«Хорошие»
Кратность варианты	8	22	10
Частота варианты	0,2	0,55	0,25
Частота (%) варианты	20	55	25

Гистограмма распределения кратностей

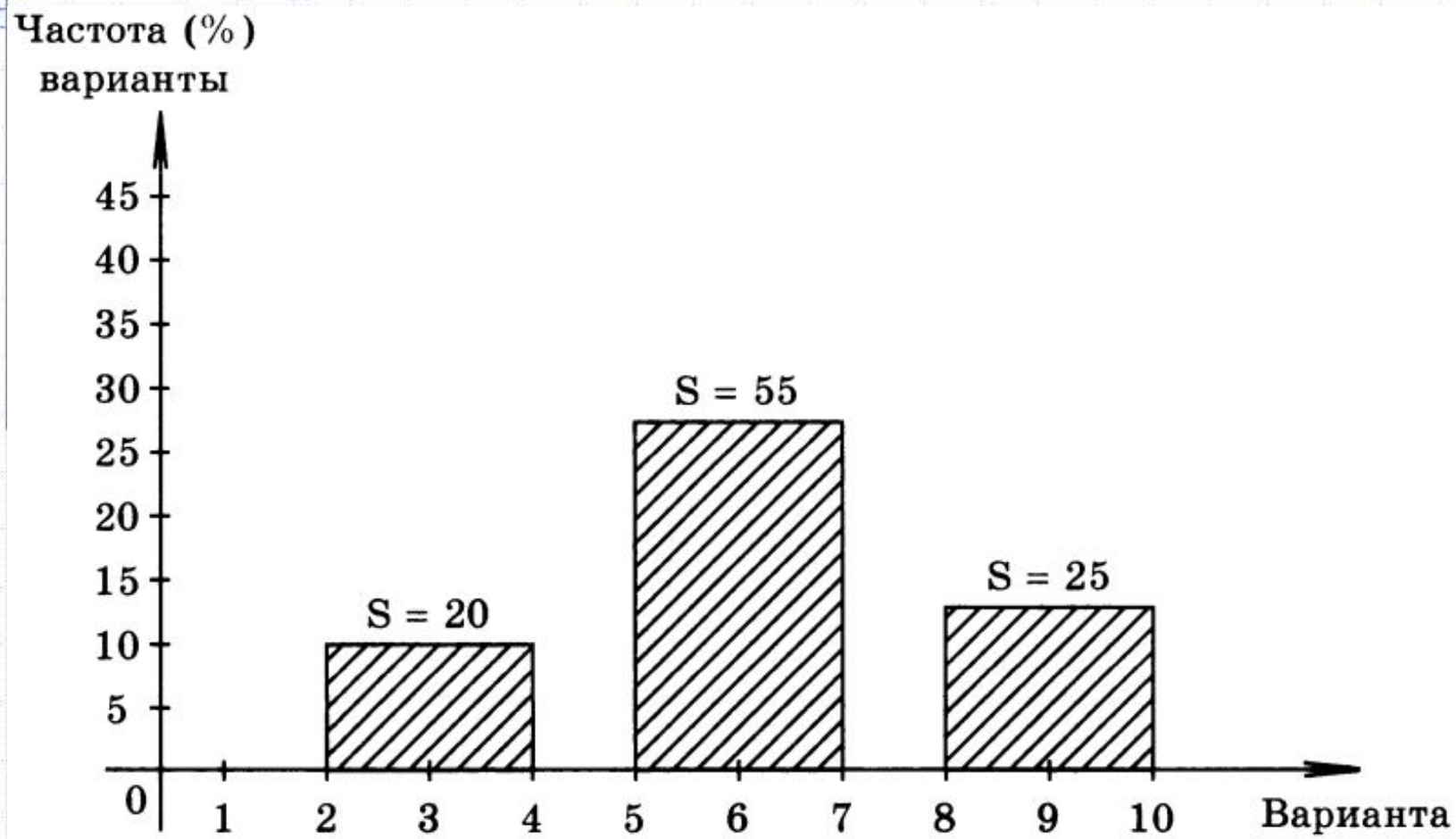


Площадь равна кратности варианты.

Гистограмма распределения частот



Гистограмма распределения частот (%)



«-» представления информации в виде гистограмм

- Теряется первоначальная точная информация

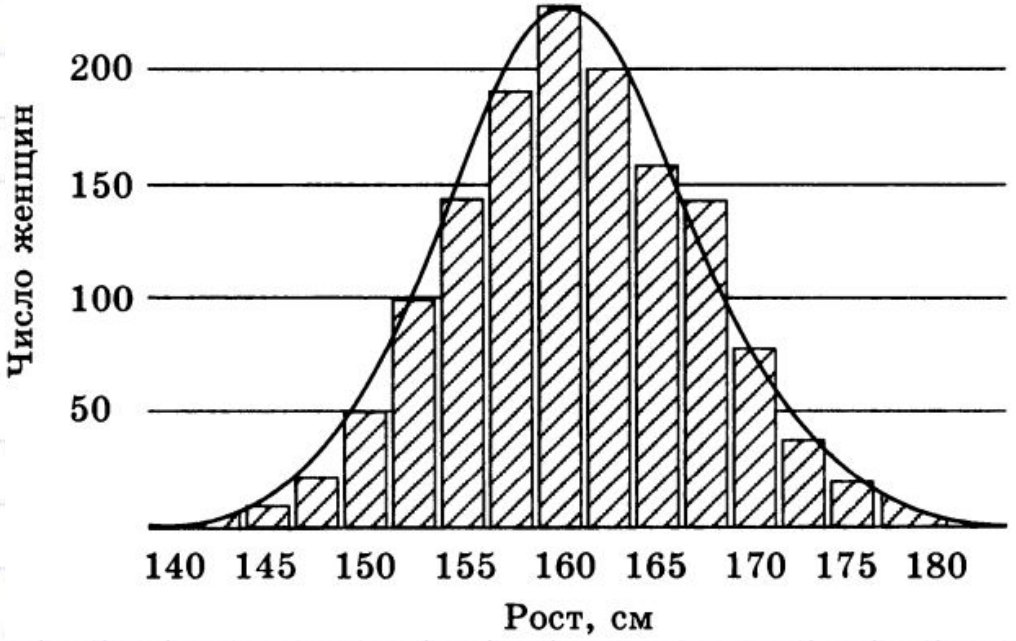
«+»

- Ответ получается более быстро
- Наглядно видна качественная оценка распределения данных

III. Гистограммы распределения большого объёма информации

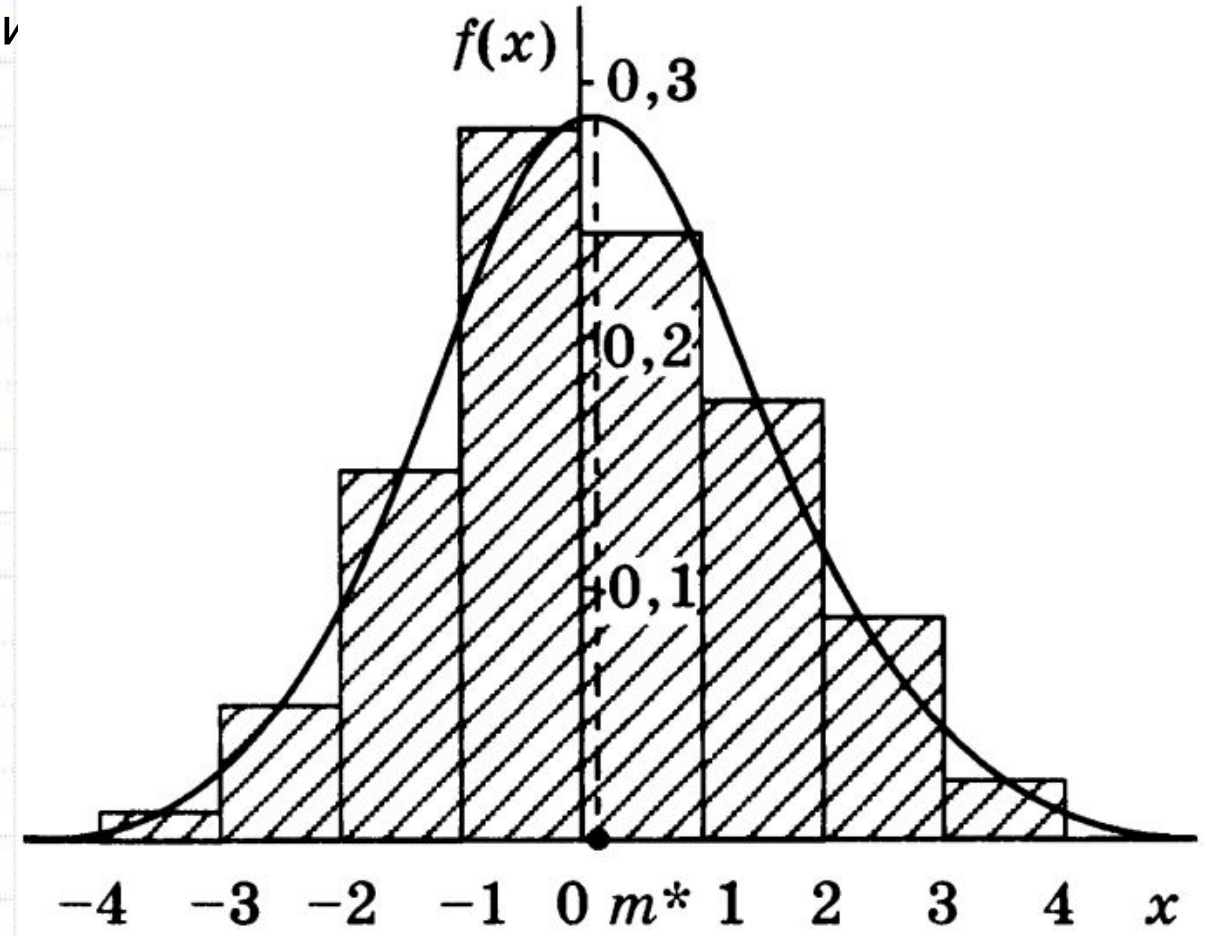
Гистограммы незаменимы, когда ряд данных состоит из большого количества чисел (сотни, тысячи и т. п.).
Если ширина столбцов гистограммы мала, а основания столбцов в объединении дают некоторый промежуток, то сама гистограмма похожа на график непрерывной функции.
Такую функцию называют ***выравнивающей* функцией**.

Пример 4.
Гистограмма роста женщин, построенная по выборке, в которой было 1375 женщин.

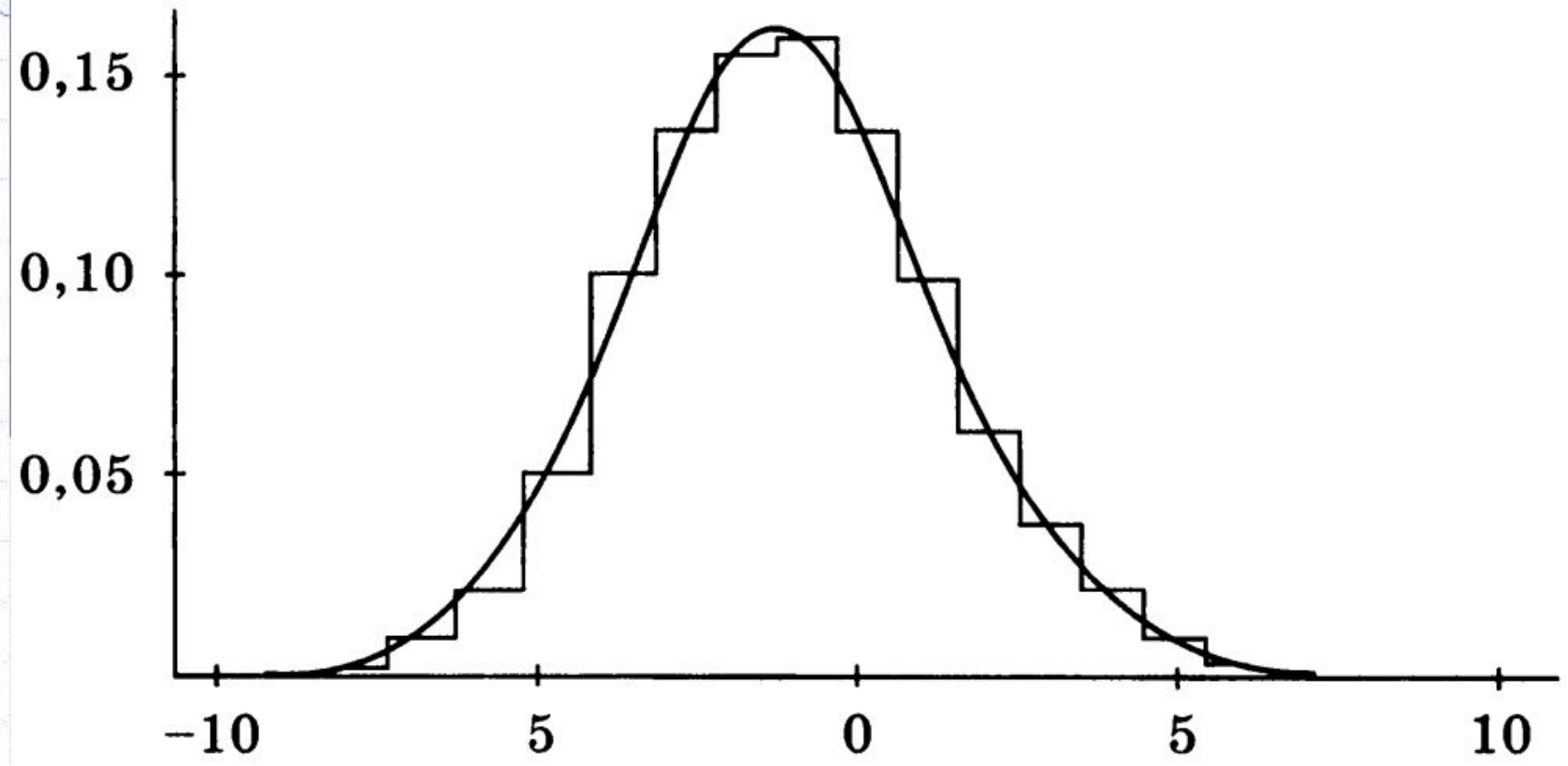


Пример 5. Произвели 500 измерений боковой ошибки при стрельбе с самолета.

На графике по оси абсцисс отложены величины ошибок («левее или правее» цели), а по оси ординат отложены частоты этих оши



Пример 6. Измерялся размер 12000 бобов.
По оси абсцисс откладывались величины отклонений от среднего размера бобов, а по оси ординат соответствующие частоты



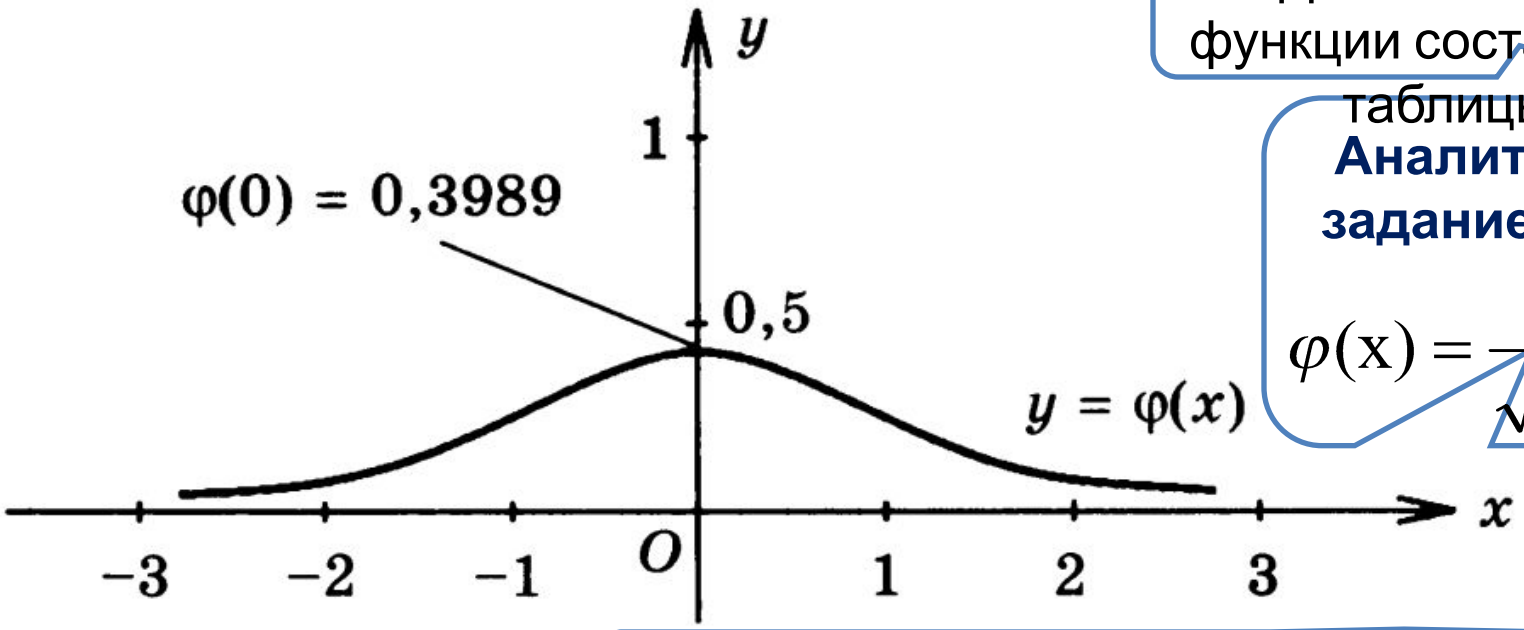
Примеры взяты из различных областей, а **графики функций, выравнивающих гистограммы, похожи** друг на друга.

Такому же закону распределения подчиняется:

- Распределение горошин по размеру
- Распределение новорожденных младенцев по весу
- Распределение частиц газа по скоростям движения
- …

Все эти кривые получаются из одной кривой.
Её называют ***кривой нормального распределения*** или, в честь Карла Гаусса, ***гауссовой кривой.***

Гауссова кривая
(кривая нормального распределения)



Для значений функции составлены таблицы

Аналитическое задание кривой:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

e (число Эйлера) = 2,7182818284590452353602874713527…

Свойства:

1)Симметрична относительно оси **Oy**

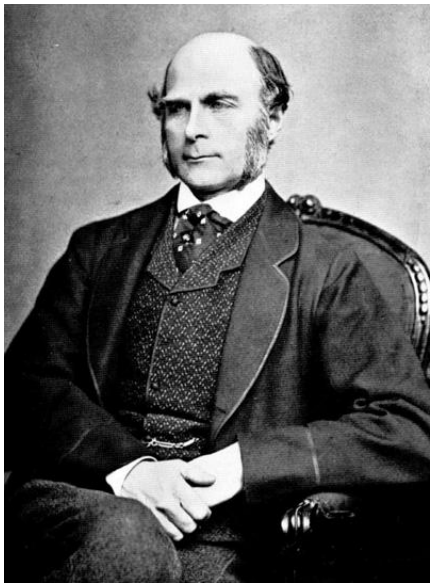
2)Единственный максимум (**ϕ(0) = 0,3989**)

3)Площадь части плоскости, ограниченной кривой и осью Ox равна **1**.

4)«Ветви» очень быстро приближаются к оси абсцисс: площадь «под гауссовой кривой» на [-3; 3] равна **0,99**

Доска Гальтона *(квинкункс, **1873 г.**)*

Устройство для наглядной демонстрации нормального (гауссова) закона распределения

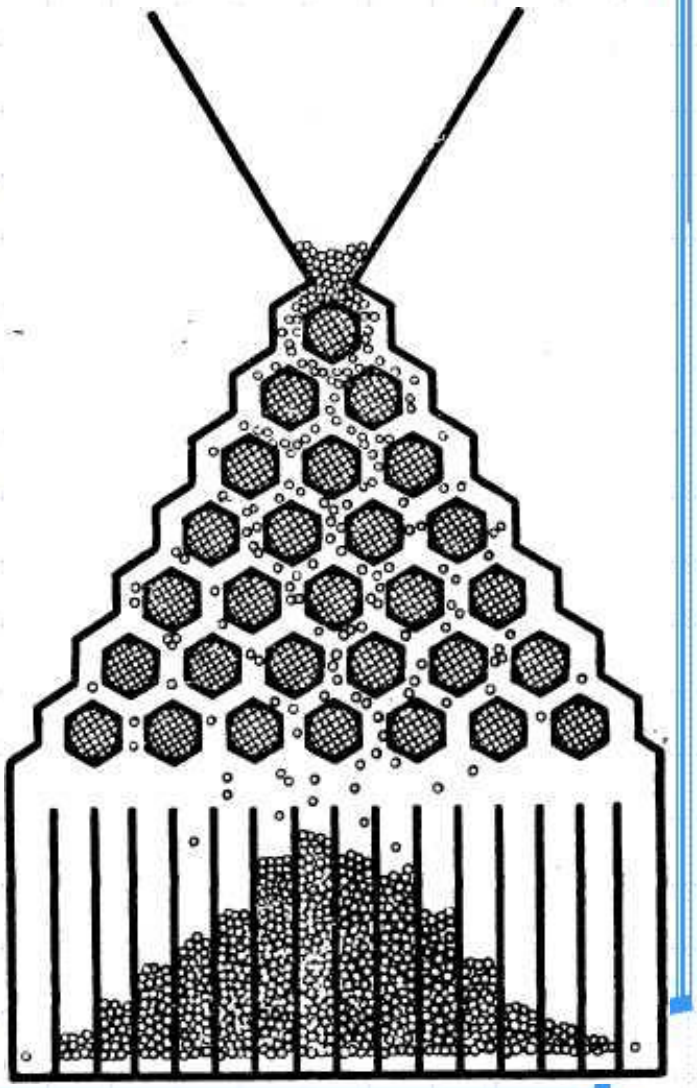


Ф. Гальтон
(1822 — 1911)

- География
- Антропология
- Статистика
- **Дифференциальная психология**
- **Психометрика**

Принцип действия:

- Падающие сверху шарики распределяются между правильными шестиугольниками
- В результате попадают на горизонтальную поверхность
- Образуют картинку, похожую на «подграфик» гауссовой кривой.



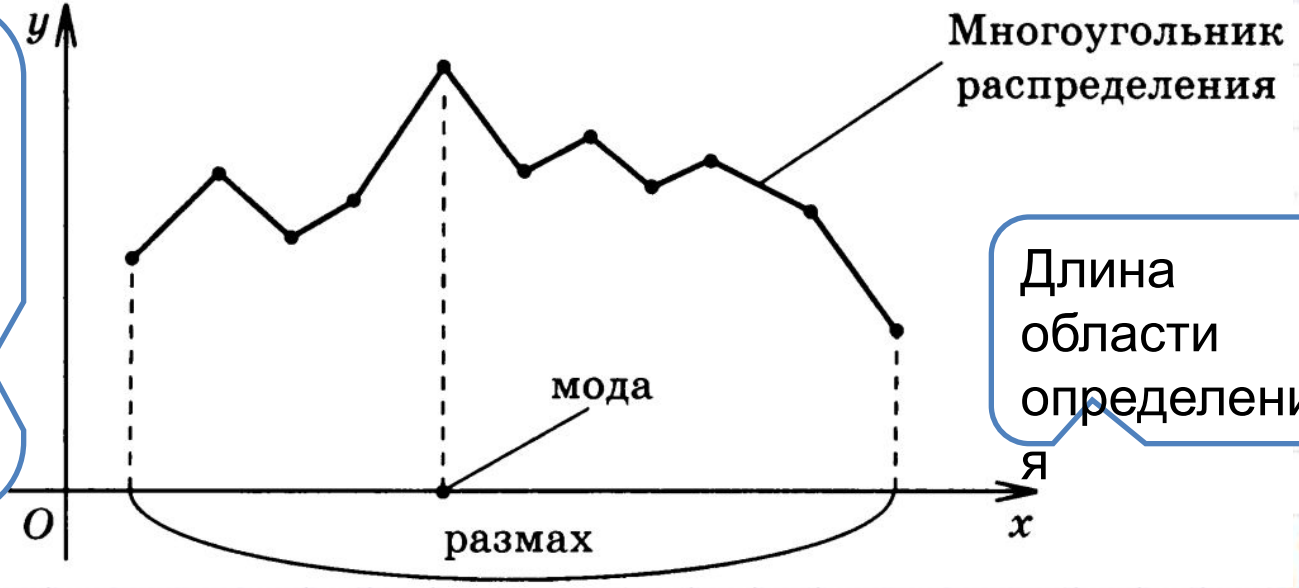
IV. Числовые характеристики выборки

Объемы выборок данных велики ⇒
Приходится иметь дело с числовыми характеристиками

1) **Размах (R)**
— это разница между наибольшей и наименьшей вариантой
(**$R = X_{max} - X_{min}$**)

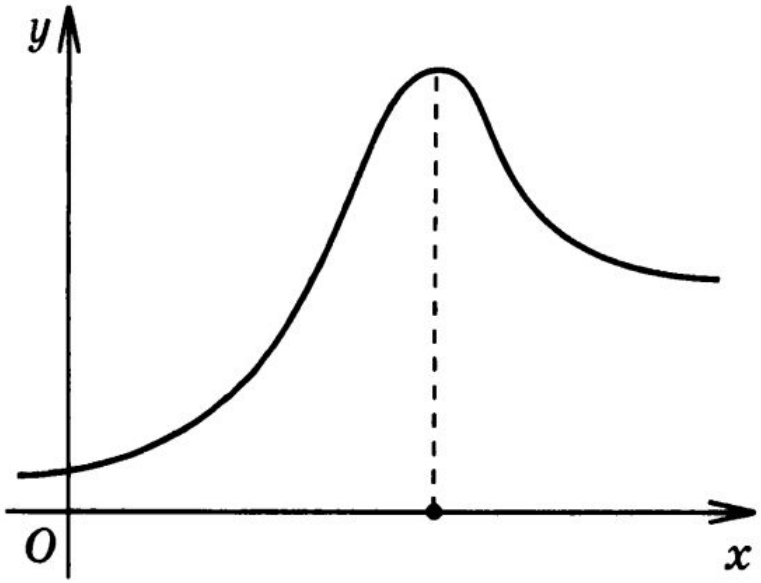
2) **Мода (Mo)**
— это наиболее часто встречающаяся ее варианта

Точка, в которой достигается максимум (Если одна, то выборка – **унимодальная**)

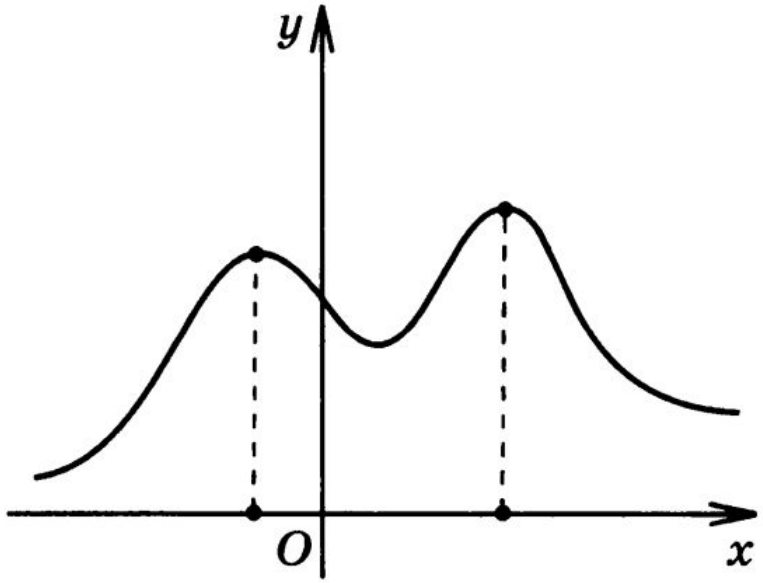


Длина области определения

Унимодальная кривая



Бимодальная кривая



3) **Медиана (Me)**

(от лат. mediana – «среднее»)

- **Медианой** выборки с **нечетным** числом вариант называется варианта, записанная **посередине** в упорядоченной выборке
- **Медианой** выборки с **четным** числом вариант называется **среднее арифметическое** двух вариант, записанных посередине в упорядоченной выборке

4) **Среднее значение (среднее арифметическое значение, $\bar{x}$)**

- Сумма результатов разделённая на их количество

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Пример 7.

Найдите среднее значение, размах и моду выборки:

а) 32; 26; 18; 26; 15; 21; 26

1. $\bar{x} = \frac{32+26+18+26+15+21+26}{7} = \frac{164}{7} = 23\frac{3}{7}$

2. X_{max}: 32

X_{min}: 15

$R = X_{max} - X_{min} = 32 - 15 = \mathbf{17}$

3. Mo = **26**

б) 21; 18,5; 25,3; 18,5; 17,9

1. $\bar{x} = \frac{21+18{,}5+25{,}3+18{,}5+17{,}9}{7} = \frac{101{,}2}{5} = 20{,}24$

2. X_{max}: 25,3

X_{min}: 17,9

$R = X_{max} - X_{min} = 25{,}3 - 17{,}9 = \mathbf{7{,}4}$

3. Mo = **18,5**

Пример 8.

В выборке 2, 7, 10, _, 18, 19, 27 одно число оказалось стертым.

Восстановите его, зная, что среднее значение этих чисел равно 14.

Решение:

Пусть искомое число **X**

$$\bar{x} = \frac{2+7+10+X+18+19+27}{7} = 14 \quad \Rightarrow$$

$$\frac{83+X}{7} = 14 \quad \Rightarrow \quad 83+X = 98 \quad \Rightarrow \quad X = 15$$

Ответ: 15

Пример 9.
Найдите медиану выборки:
30, 32, 37, 40, 41, 42, 45, 49, 52;

Решение:
1) Упорядочить выборку: **30, 32, 37, 40, 41, 42, 45, 49, 52**
2) Число членов ряда: n = 9
3) Серединный элемент (5-ый): 41
4) Me = **41**

Пример 10.
Зная, что в упорядоченном ряду содержится m чисел, где m — нечетное число, укажите номер члена, являющегося медианой, если m равно: 5

Решение:
Номер члена, являющегося медианой: **3**

Пример 11.

В ряду данных, состоящем из 12 чисел, наибольшее число увеличили на 6. Изменятся ли при этом и как:

а) среднее значение;

Увеличится на 1/2

б) размах;

Увеличится на 6

в) мода;

Не изменится (?)

г) медиана?

Не изменится (?)

5) **Среднее отклонение ($\overline{d}$)**

Среднее арифметическое отклонений (в абсолютных показателях) всех вариант выборки от их среднего значения.

$$\overline{d} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| x_i - \overline{x} \right|$$

6) **Дисперсия (D)**

Величина колебания вариант около их среднего значения

$$D = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2$$

7) **Среднее квадратичное отклонение (σ** - сигма**)**

$$\sigma = \sqrt{D} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}$$

8) **Коэффициент вариации (CV)**

$$CV = \frac{\sigma}{\overline{x}} \cdot 100\%$$

0 ≤ CV ≤ 10% - выборка однородна

11 ≤ CV ≤ 20% - средняя степень однородность

21 ≤ CV – низкая степень однородности

Пример 12.

Вычислите среднее отклонение, дисперсию, среднее квадратичное отклонение и коэффициент выборки:

46; 50; 59; 60; 55; 49

№	x_i	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
1	46	7,2	51,4
2	50	3,2	10,0
3	59	5,8	34,0
4	60	6,8	46,7
5	55	1,8	3,4
6	49	4,2	17,4
Σ	319	29	162,9

$$\bar{x} = \frac{319}{6} = 53{,}2$$

0 ≤ CV ≤ 10% - выборка однородна

$$\bar{d} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| x_i - \bar{x} \right| = \frac{29}{6} = 4{,}8$$

$$\sigma = \sqrt{D} = 5{,}2$$

$$D = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{162{,}9}{6} = 27{,}2$$

$$CV = \frac{5{,}2}{53{,}2} \cdot 100\% = 9{,}8\%$$

V. Экспериментальные данные и вероятности событий

Пример 13. *Бросание монеты*

Запишем **О** или **Р** в зависимости от того, выпал «орел» или «решка».

После **n** бросаний при неизменных условиях этого испытания, получится случайная последовательность.

Например: **О, О, Р, О, Р, Р, О, Р, Р, Р, О, О, Р, О, Р, О, О, Р, Р, О, О, Р…**

Т[illegible] бросаний [illegible] О и Р.

Рассмотрим расчеты для указанной последовательности:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Частота Р	0	0	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{2}{5}$	$\frac{3}{6}$	$\frac{3}{7}$	$\frac{4}{8}$	$\frac{5}{9}$	$\frac{6}{10}$	$\frac{6}{11}$	$\frac{6}{12}$	$\frac{7}{13}$	$\frac{7}{14}$
Частота О	1	1	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{3}{5}$	$\frac{3}{6}$	$\frac{4}{7}$	$\frac{4}{8}$	$\frac{4}{9}$	$\frac{4}{10}$	$\frac{5}{11}$	$\frac{6}{12}$	$\frac{6}{13}$	$\frac{7}{14}$

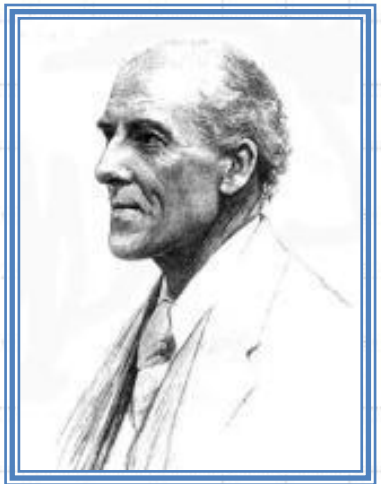
При достаточно большом числе бросаний частота приближается к некоторому постоянному числу.

В данном случае к **0,5**.

Ж. Бюффон
(1707 — 1788)

Бросил монету **4040** раз, и при этом герб выпал в **2048** случаях.

$$Частота\ O = \frac{2048}{4040} = 0{,}50693\ldots$$



К. Пирсон
(1857-1936)

Бросил монету **24000** раз, и при этом герб выпал в **12012** случаях.

$$Частота\ O = \frac{12012}{24000} = 0{,}50005\ldots$$

Статистическая устойчивость (СУ)

При ***большом*** числе независимых повторений одного и того же опыта в неизменных условиях частота появления определенного случайного события практически ***совпадает*** с некоторым ***постоянным числом***. Такое число называют **статистической вероятностью этого события.**

СУ имеет место при:

- Выпадении определенного числа очков на игральных кубиках
- Рождении мальчиков
- Времени восхода солнца
- …

СУ соединяет реально проводимые испытания с теоретическими моделями этих испытаний.

Пример 14.

Статистические исследования над литературными текстами показали, что частоты появления той или иной буквы (или пробела между словами) стремятся при увеличении объема текста к некоторым константам.

Таблицы, в которых собраны буквы того или иного языка и соответствующие константы, называют ***частотными таблицами*** языка.

Таблица для букв русского алфавита и пробелов (частоты приведены в процентах)

Буква	А	Б	В	Г	Д	Е	Ж	З	И	Й
Частота	6,2	1,4	3,8	1,1	2,5	7,2	0,7	1,6	6,2	1,0

Буква	К	Л	М	Н	О	П	Р	С	Т	У	Ф
Частота	2,8	3,5	2,6	5,3	9,0	2,3	4,0	4,5	5,3	2,1	0,2

Буква	Х	Ц	Ч	Ш	Щ	Ы	Ь	Э	Ю	Я	–
Частота	0,9	0,4	0,4	0,6	0,3	1,6	1,4	0,3	0,6	1,8	17,5

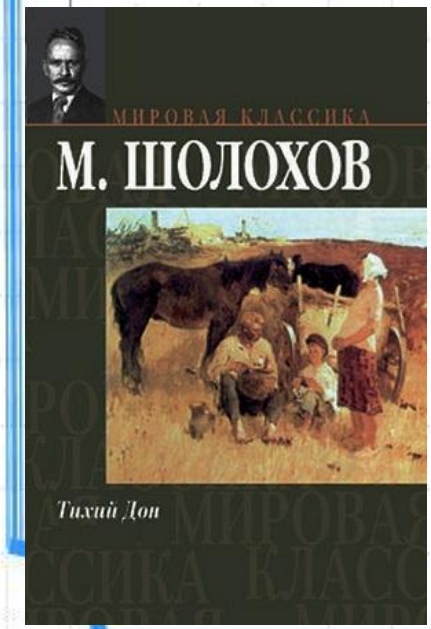
М.А. Шолохов
(1905 — 1984)

Пример 15.

До сегодняшнего дня не утихают споры об авторстве «Тихого Дона».

Многие считают, что в **23** года М. А. Шолохов такую глубокую и поистине великую книгу написать не мог.

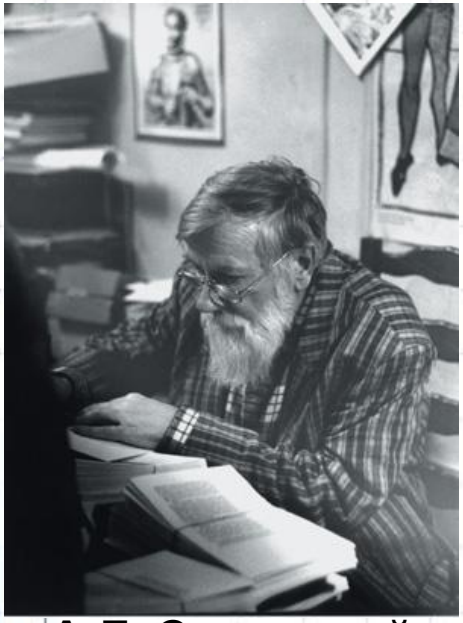
Особенно жаркими были споры в момент при суждения М. А. Шолохову Нобелевской премии в области литературы (**1965 г.**).



Статистический анализ романа и сличение его с текстами, в авторстве которых не было сомнений, подтвердил гипотезу о М. А. Шолохове, как об ***исти на».***

Шведский король Густав Адольф поздравляет М. А. Шолохова с присуждением ему Нобелевской премии (Стокгольм, 1965)

Пример 16.



А.Д. Синявский (1925 — 1997)

В середине 60-х годов в одной из стран Западной Европы были опубликованы «***очерняющие прогрессивный характер социалистической системы***» литературные произведения.

Автором был **А. Терц**, но это псевдоним.

Был проведен сравнительный анализ опубликованных «вредительских» текстов и результаты были сличены с произведениями ряда возможных канди

Ответ оказался однозначным: настоящим автором был литературовед **А.Д. Синявский**.

В **1967** году («***Процесс Синявского и Даниэля***») получил 5 лет тюрьмы и 7 лет ссылки.

А. Д. Синявский и Ю. М. Даниэль в зале суда

*

Домашнее задание

1. Конспект