# Лекция 1: Экспериментальные данные. Характеристики выборки и генеральной совокупности

1. Классификация видов экспериментальных исследований.
2. Типы погрешностей измерений и их оценки.
3. Гистограмма и полигон частот.
4. Параметры распределения и их влияние на вид кривой распределения.

# 1 Классификация видов экспериментальных исследований

способы получения данных
через познание окружающего мира

**наблюдение**
воздействие наблюдателя на объект минимально

**эксперимент**
наблюдение с воздействием на объект

Результат эксперимента: выводы и рекомендации. Информация может быть выражена в виде графиков, чертежей, таблиц, статистических данных или словесных описаний.

Эксперимент предполагает проведение опытов.

**Опыт** – *воспроизведение исследуемого явления в определенных условиях проведения эксперимента при возможности регистрации его результатов.*

По цели проведения и форме представления полученных результатов

**эксперимент**

**качественный**

устанавливает факт существования явления, не дает количественных характеристик объекта (словесное описание результатов эксперимента)

**количественный**

фиксирует существование явления, устанавливает соотношение между количественными характеристиками явления и внешнего воздействия на объект (количественное описание факторов)

**Фактор** – *переменная величина, по предположению влияющая на результаты эксперимента*.

**Уровень фактора** - *фиксированное значение фактора относительно начала отсчета*.

В отдельном конкретном опыте каждый фактор может принимать одно из значений уровня. Фиксированный набор уровней всех факторов в опыте определяет одно из возможных состояний объекта исследований.

| Вид фактора | Регистрация уровня | Установка уровня |
|---|---|---|
| Контролируемый, управляемый | + | + |
| Контролируемый, неуправляемый | + | - |
| Неконтролируемый | - | - |

**количественный эксперимент**

**пассивный** — нет управляемых факторов

**активный** — есть управляемые факторы

**Отклик** – *наблюдаемая случайная величина, по предположению зависящая от факторов,* т.е. некое исследуемое свойство объекта.

В количественном эксперименте необходимо:
-регистрировать уровни всех контролируемых факторов;
-количественно описать отклик – установить **функцию отклика** (*зависимость между факторами и откликом*).

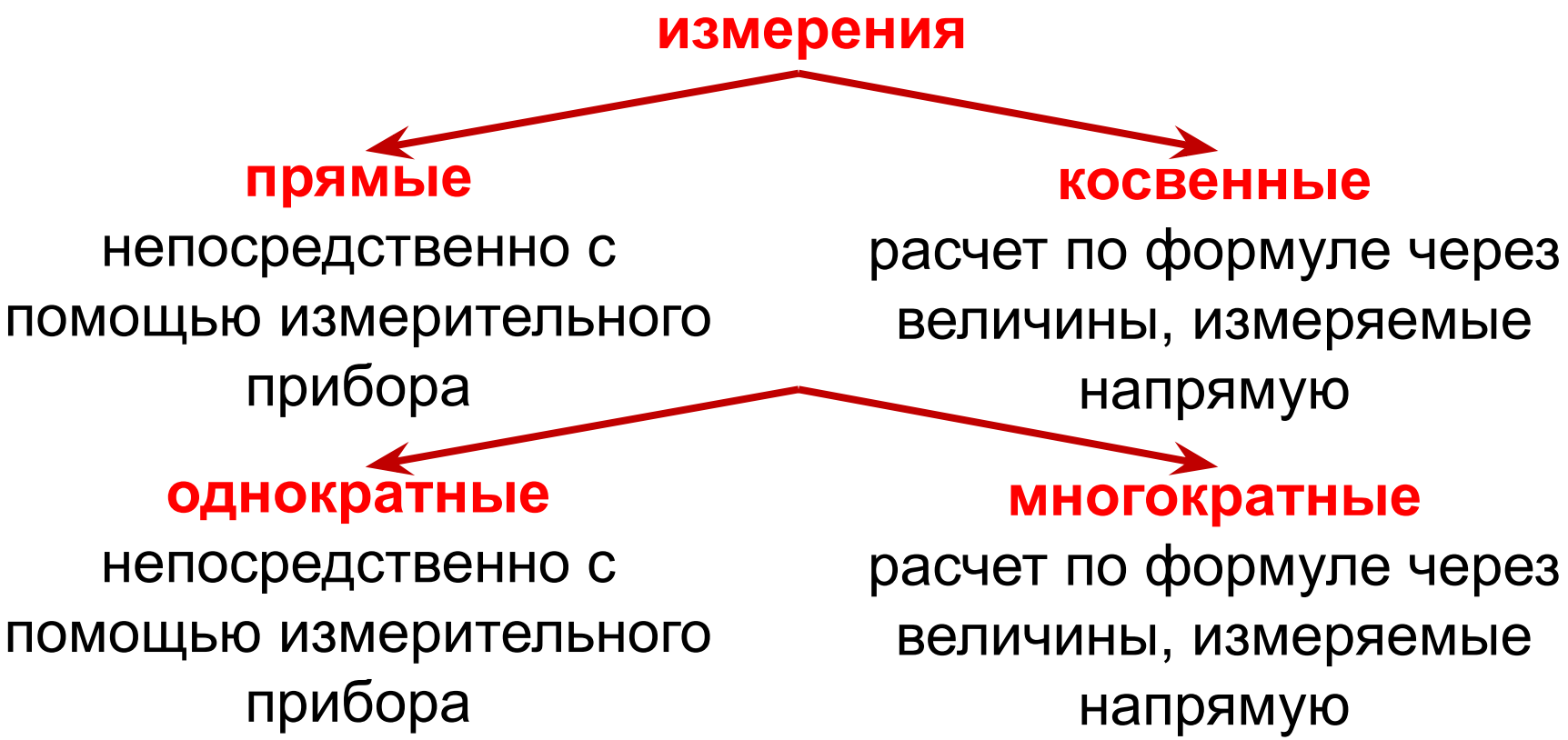
Вид функции отклика в общем случае:

$$f(x_i, h_j) + \varepsilon_\delta$$

$x_i$ – контролируемые и управляемые факторы;
$h_j$ – контролируемые и неуправляемые факторы;
$\varepsilon_\delta$ - ошибка эксперимента (влияние неконтролируемых факторов)

# 2 Типы погрешностей измерений и их оценки

Предмет количественного эксперимента - количественные величины. Для определения абсолютного значения величины ее сравнивают с единицей величины - эталоном.



***Погрешность*** – количественная характеристика неоднозначности результата измерений.

Нельзя получить экспериментальные данные с абсолютной точностью => необходимо

- оценить значение измеряемой величины
- указать, насколько оценка близка к истинному значению
- оценить качество измерений

Пусть результат многократных измерений – случайный вектор $X = \{x_1, x_2, \ldots, x_N\}$, тогда оценка значения -

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

***Абсолютная погрешность*** $\Delta X$ - разность между истинным значением измеряемой величины $X$ и его оценкой.

Абсолютная погрешность

- выражается в единицах измеряемой величины $X$;
- не отражает качества измерений.

***Относительная погрешность*** – критерий качества измерений. Безразмерная величина.

$$\varepsilon = \frac{\Delta X}{\overline{X}} \qquad \text{или} \qquad \varepsilon = \frac{\Delta X}{\overline{X}} \cdot 100\%$$

Высокой точности измерения соответствует малое значение относительной погрешности.

**погрешность**

**промах**
результат с аномальным числовым значением

**систематическая ошибка**
постоянная составляющая, изменяется закономерно

**методологическая ошибка**
неправильный выбор метода измерения. Максимально учитывается введением поправок.

**инструментальная погрешность**
погрешность прибора измерения.
Средняя инструментальная погрешность:

$$\delta = \frac{X_m}{100} K$$ односторонняя шкала;

$$\delta = \frac{2X_m}{100} K$$ двусторонняя шкала

# Алгоритм обработки данных прямых многократных измерений:

1 оценить истинное значение величины $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$

2 оценить СКО и среднеквадратическую ошибку среднего

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N-1}} \quad \text{и} \quad \sigma_x = \frac{\sigma}{\sqrt{N}}$$

3 вычислить среднюю инструментальную погрешность

$$\delta = \frac{X_m}{100} \quad \text{или} \quad \delta = \frac{2X_m}{100} K$$

4 при априорно известной доверительной вероятности $P$ найти коэффициенты Стьюдента $t(P, N-1)$ и $t(P, \infty)$

# Алгоритм обработки данных прямых многократных измерений:

5 вычислить абсолютное значение случайной погрешности и абсолютное значение инструментальной погрешности

$$\Delta X_{сл} = t(P, N-1) \cdot \sigma_x \qquad \Delta X_{пр} = t(P, \infty) \cdot \frac{\delta}{3}$$

6 рассчитать абсолютную погрешность измерений

$$\Delta X = \sqrt{\Delta X_{сл}^2 + \Delta X_{пр}^2} = \sqrt{(t(P, N-1) \cdot \sigma_x)^2 + \left(t(P, \infty) \cdot \frac{\delta}{3}\right)^2}$$

7 вычислить относительную погрешность

$$\varepsilon = \frac{\Delta X}{\overline{X}} \quad \text{или} \quad \varepsilon = \frac{\Delta X}{\overline{X}} \cdot 100\%$$

8 записать результат в виде: $X = \overline{X} \pm \Delta X$ при доверительной вероятности $P$ и погрешности измерений ε.

# Алгоритм обработки данных косвенных многократных измерений:

Пусть $Z = f(A, B, C, \ldots)$ - измеряемая величина; величины $A, B, C, \ldots$ - измерены прямыми многократными измерениями.

1 результаты измерения величин $A, B, C, \ldots$ обработать по алгоритму обработки прямых многократных измерений

2 оценить истинное значение измеряемой величины $Z$

$$\overline{Z} = f\left(\overline{A}, \overline{B}, \overline{C}, \ldots\right)$$

3 найти абсолютную погрешность измеряемой величины $Z$

$$\Delta Z = \sqrt{\left(\frac{\partial f}{\partial A}\right)^2_{A=\overline{A},B=\overline{B}\ldots} (\Delta A)^2 + \left(\frac{\partial f}{\partial B}\right)^2_{A=\overline{A},\ldots} (\Delta B)^2 + \left(\frac{\partial f}{\partial C}\right)^2_{A=\overline{A},\ldots} (\Delta C)^2 + \ldots}$$

4 найти относительную погрешность $\varepsilon = \frac{\Delta Z}{\overline{Z}}$ или $\varepsilon = \frac{\Delta Z}{\overline{Z}} \cdot 100\%$

# Соотношения для расчета погрешностей косвенных измерений для простейших функций

| $f$ | $\Delta f$ | $\varepsilon = \Delta f / f$ |
|---|---|---|
| A+B | $\sqrt{(\Delta A)^2 + (\Delta B)^2}$ | $\frac{\sqrt{(\Delta A)^2 + (\Delta B)^2}}{\overline{A} + \overline{B}}$ |
| A-B | $\sqrt{(\Delta A)^2 + (\Delta B)^2}$ | $\frac{\sqrt{(\Delta A)^2 + (\Delta B)^2}}{\overline{A} - \overline{B}}$ |
| A·B | $\sqrt{\overline{B}^2(\Delta A)^2 + \overline{A}^2(\Delta B)^2}$ | $\sqrt{\left(\frac{\Delta A}{\overline{A}}\right)^2 + \left(\frac{\Delta B}{\overline{B}}\right)^2}$ |
| $\frac{A}{B}$ | $\frac{\sqrt{\overline{B}^2(\Delta A)^2 + \overline{A}^2(\Delta B)^2}}{\overline{B}^2}$ | $\sqrt{\left(\frac{\Delta A}{\overline{A}}\right)^2 + \left(\frac{\Delta B}{\overline{B}}\right)^2}$ |

# Соотношения для расчета погрешностей косвенных измерений для простейших функций

| $f$ | $\Delta f$ | $\varepsilon = \Delta f / f$ |
|---|---|---|
| $\sin A$ | $\cos \overline{A} \cdot \Delta A$ | $\operatorname{ctg} \overline{A} \cdot \Delta A$ |
| $\cos A$ | $\sin \overline{A} \cdot \Delta A$ | $\operatorname{tg} \overline{A} \cdot \Delta A$ |
| $\operatorname{tg} A$ | $\frac{1}{\cos^2 \overline{A}} \Delta A$ | $\frac{2}{\sin 2\overline{A}} \Delta A$ |
| $\ln A$ | $\frac{\Delta A}{\overline{A}}$ | $\frac{1}{f} \frac{\Delta A}{\overline{A}}$ |

Результат косвенных измерений записывается в виде:

$Z = \overline{Z} \pm \Delta Z$ при доверительной вероятности $P$ и погрешности измерений ε

# 3 Гистограмма и полигон частот.

Предварительная обработка данных начинается с определения того, какими типами переменных представлены данные.

***Типы переменных (признаков) представления данных:***

- ***непрерывные*** – представлены действительными числами (например, длина или вес);
- ***дискретные*** – представлены целыми, как правило, положительными числами;
- ***категориальные*** (например, марка кабеля, тип материала, географический регион). Значения категориальных данных не могут быть положены на числовую прямую.

Построение ***гистограммы*** или ***полигона частот*** - самый простой способ наглядного представления о распределении вероятности выпадения того или иного значения случайной величины по выборке.

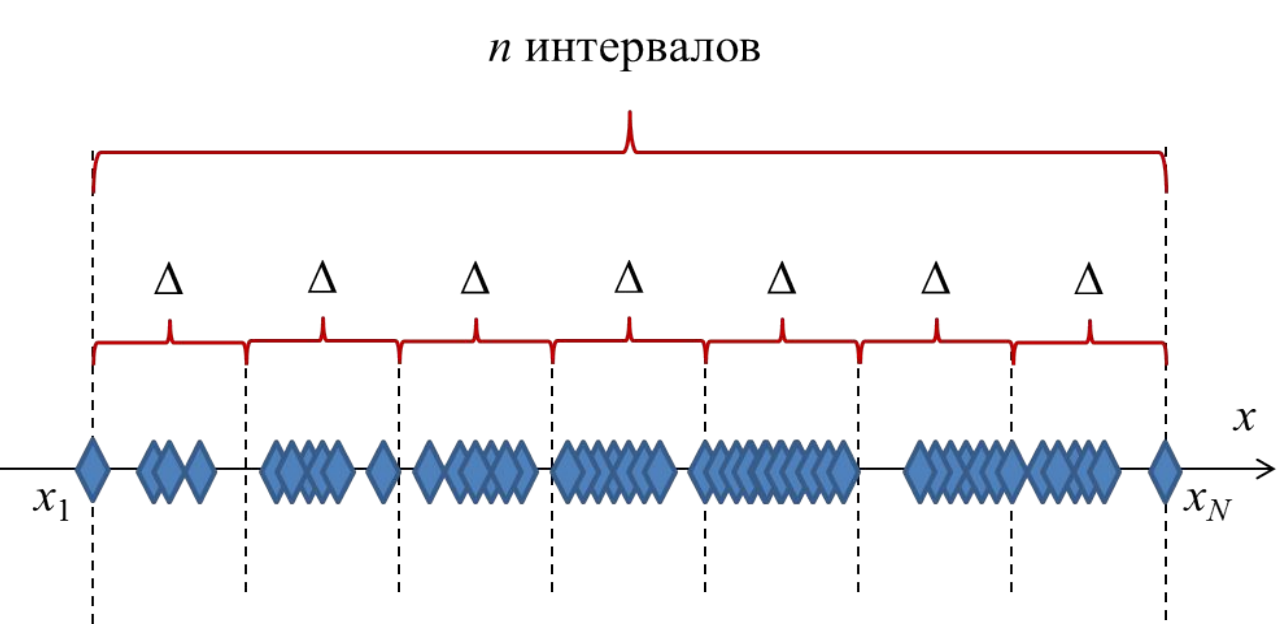
Пусть выборка из экспериментальных данных: $x=\{x_1,\dots x_N\}$.

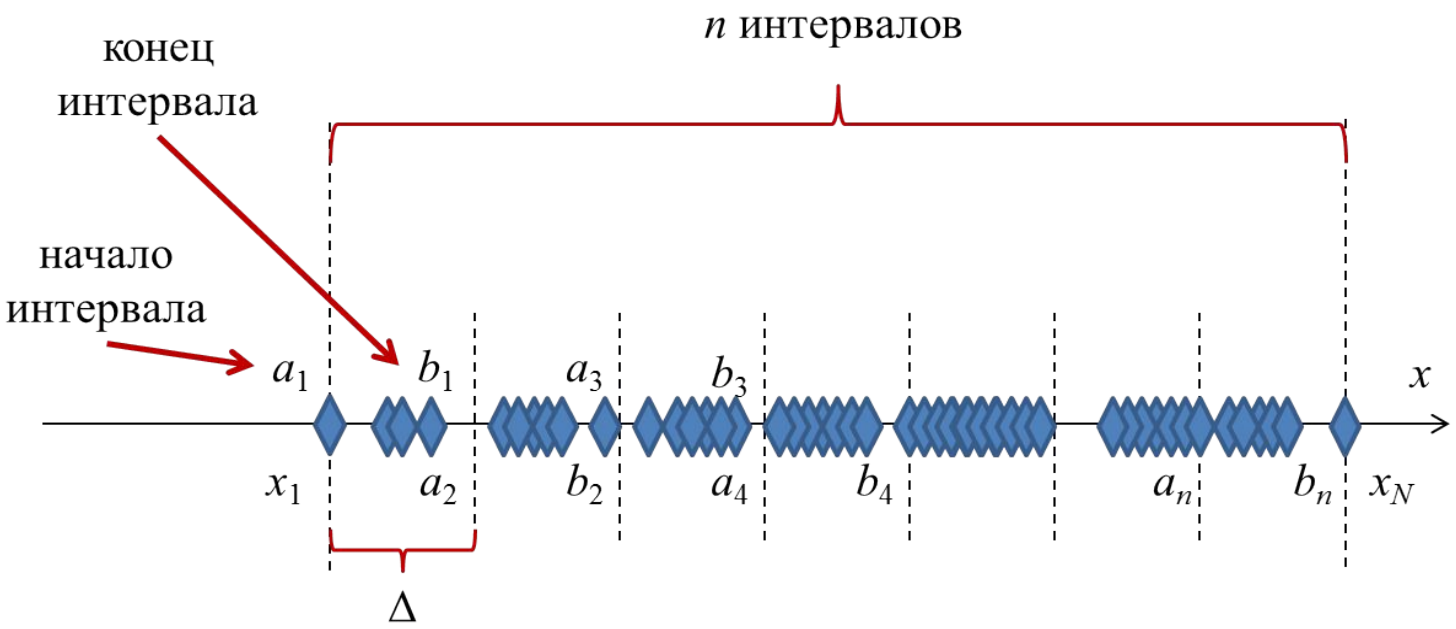
## *Алгоритм построения гистограммы и полигона частот*

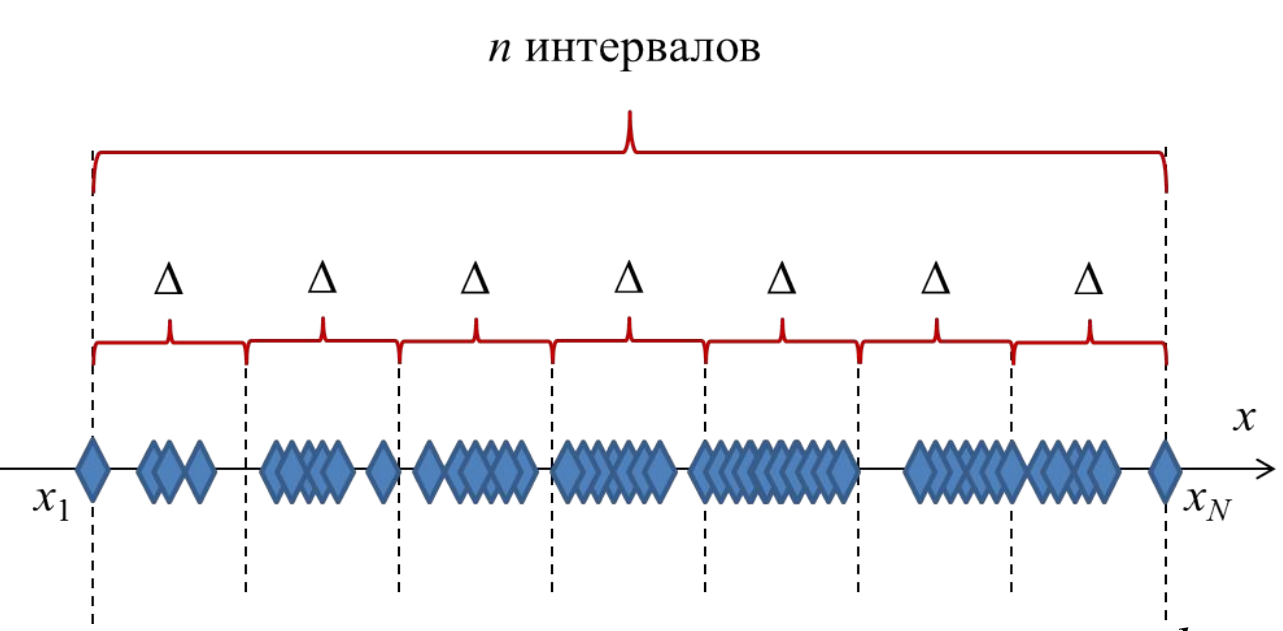
1. Построение вариационного ряда $x_1 \le x_2 \le \dots \le x_N$
2. Группировка данных: разбиение отрезка $[x_1, x_N]$ на «карманы». Как и на сколько «краманов» разбивать? Рассмотрим рзбиение на «карманы» равной длины.

Определение числа «карманов»

- по правилу Стерджесса: $n = 1 + 3{,}322 \cdot \lg N$,
- по формуле Брукса и Каррузера: $n = 5 \cdot \lg N$
- по формуле: $n = \sqrt{N}$

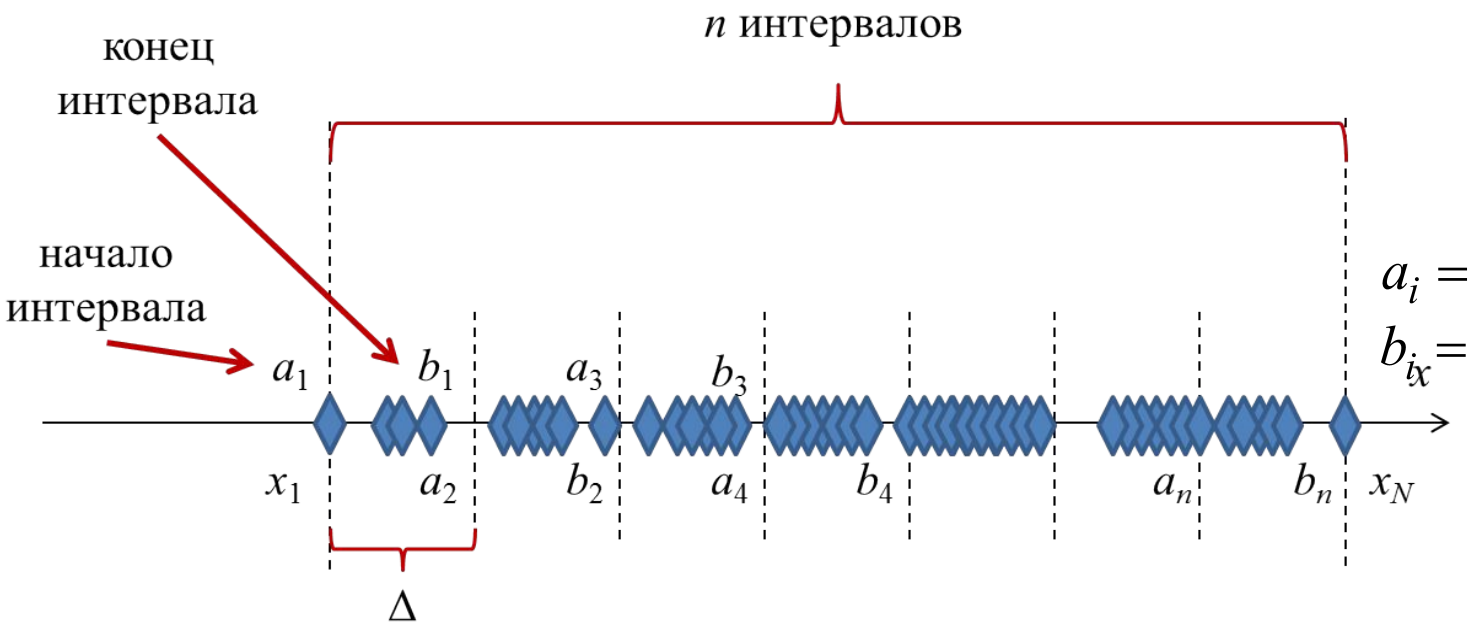

n интервалов
Δ
Δ
Δ
Δ
Δ
Δ
Δ
x
x1
xN


n интервалов
конец интервала
начало интервала
a1
b1
a3
b3
x
x1
a2
b2
a4
b4
an
bn
xN
Δ




$$\Delta = \frac{x_N - x_1}{n},$$

$$a_1 = x_1,\ b_n = x_N,\ a_i = b_{i-1},\ \text{для}\ i = 2...n$$



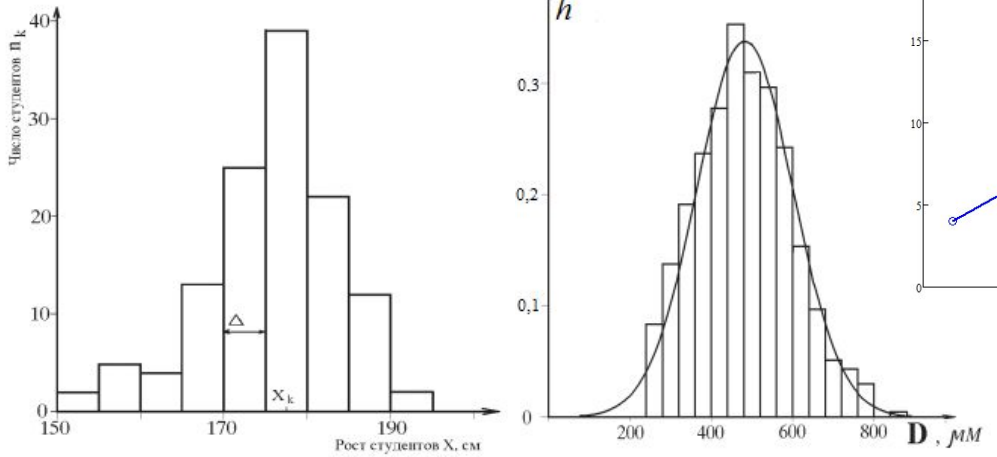

$$a_i = x_1 + (i-1)\Delta,$$
$$b_i = x_1 + i\Delta$$

3. Вычисление числа значений, попавших в каждый интервал и построение (нормированной) ***гистограммы***

$$T_i = \sum_{j=1}^{N} t_{j,i}, \quad t_{j,i} = \begin{cases} 1, \text{ если } x_j \in [a_i, b_i], \\ 0, \text{ если } x_j \notin [a_i, b_i]. \end{cases} \qquad h_i = \frac{T_i}{N \cdot \Delta} \text{ - нормировка } T_i$$

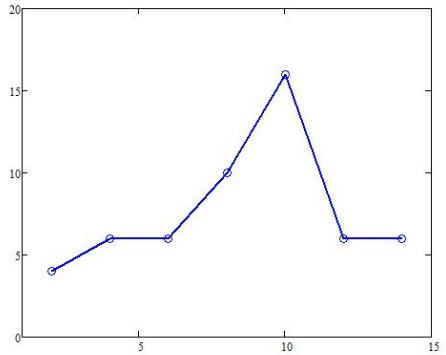
или

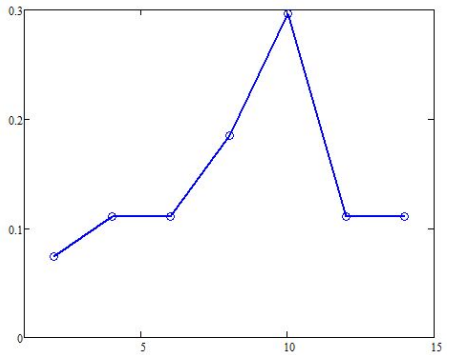
4. Определение координат центров отрезков $c_i$ и построение ***полигона (относительных) частот*** – ломанной по точкам $(c_i, T_i)$ или $(c_i, h_i)$











$h_i \cdot \Delta$ - вероятность попадания результата отдельно измерения в данный интервал. Полная вероятность равна 1, значит

$$\sum_{i=1}^{N} h_i \Delta = 1$$

При увеличении числа измерений в пределе получаем вместо гистограммы ***кривую распределения*** – график ***функции плотности вероятности f(x).***

Следовательно,

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Вероятность попадания измеряемой величины в интервал $(-\infty, x]$ называют ***функцией распределения*** или ***интегральной функцией распределения***:

$$F(x) = \int_{-\infty}^{x} f(z)dz$$

Исходя из определения,

$$F(-\infty) = 0 \quad F(+\infty) = 1 \qquad \mathrm{P}(x_1 < x < x_2) \equiv \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1)$$

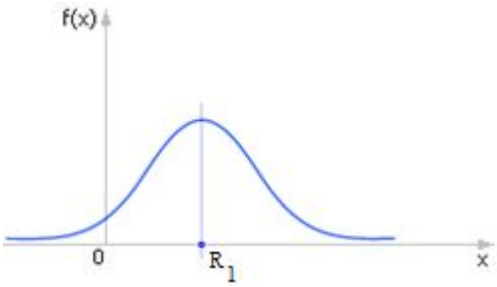
# 4 Параметры распределения и их влияние на вид кривой распределения

Характер кривой распределения описывается специальными мерами.

***Центр распределения*** характеризуется ***средним значением μ***, ***медианой Me*** и ***модой Mo***.

***Среднее значение (первый начальный момент)*** равно математическому ожиданию случайной величины:

$$\mu = R_1 = \frac{1}{N}\sum_{i=1}^{N} x_i = \int_{-\infty}^{+\infty} x\, f(x)dx$$




$R_1$ - центр тяжести в геометрии распределения.

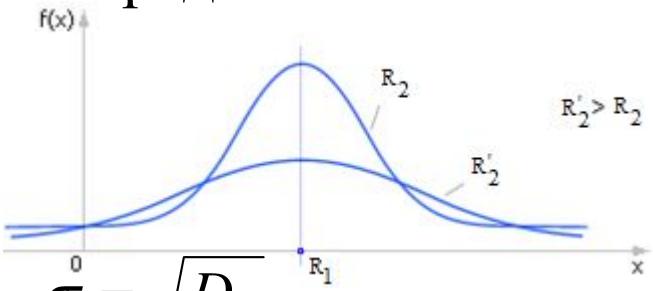
***Медиана*** делит площадь, ограниченную функцией плотности вероятности, на две равные части $P(X \le Me) = F(Me) = 0{,}5$

***Мода*** является наиболее вероятным значением случайной величины, то есть соответствует значению $x$, для которого $f(x)$=max

***Рассеяние случайных величин вокруг центра группирования*** оценивается ***дисперсией***, ***стандартным отклонением***, ***коэффициентом вариации*** и ***размахом***.

***Дисперсия (второй момент)*** – это математическое ожидание квадрата отклонения случайной величины от их среднего арифметического значения.

$$D_x = R_2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx$$




***Среднее квадратическое отклонение***, СКО: $\sigma = \sqrt{D_x}$

***Стандартное отклонение***:

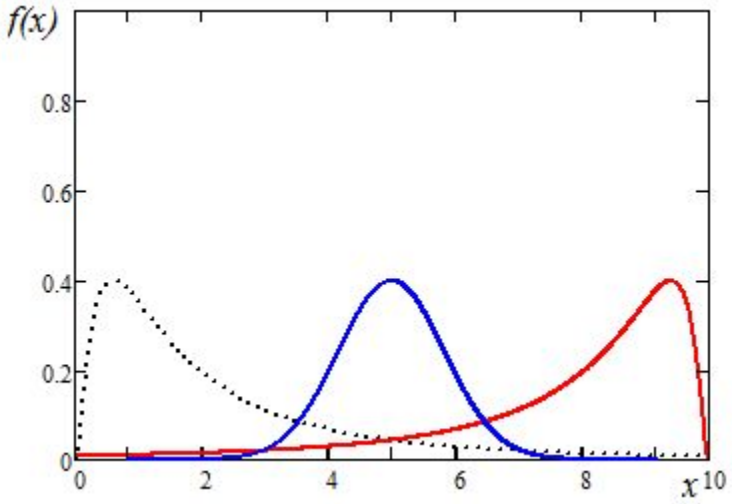
$$\sigma_{cm} = \sigma\begin{cases} 1/\sqrt{N}, & \textit{для выборки} \\ 1/\sqrt{N-1}, & \textit{для генеральной совокупности} \end{cases}$$

***Коэффициент вариации*** – отношение стандартного отклонения к математическому ожиданию случайной величины.

***Размах*** является разностью между большим и меньшим элементом выборки, то есть он равен $w=x_{\max}-x_{\min}$

***Скошенность распределения***, когда один хвост кривой распределения крутой, а другой - пологий, характеризует ***коэффициент асимметрии***, $a_3$.

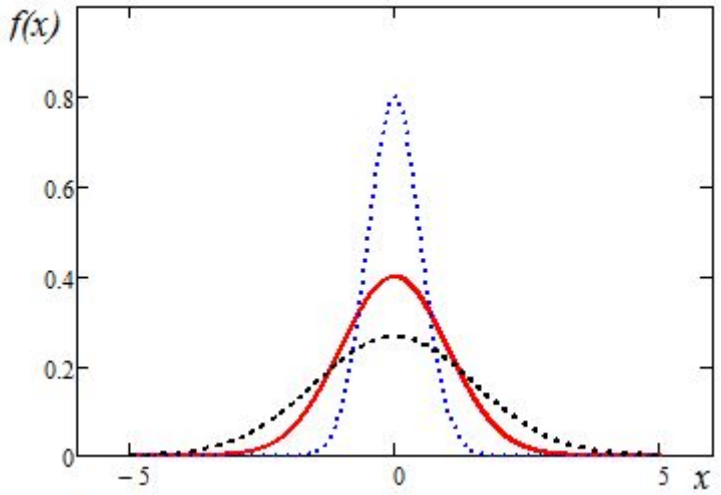
$$a_3 = \frac{R_3}{\sigma_{cm}^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - M_x)^3}{\sigma_{cm}^3} = \frac{1}{\sigma_{cm}^3}\int_{-\infty}^{\infty}(x - M_x)^3 f(x)dx$$




Синим – симметричное ($a_3$=0).
Черным - положительная асимметрия ($a_3$<0).
Красным - отрицательной асимметрия ($a_3$>0).

***Протяженность распределения*** описывается ***коэффициентом эксцесса (куртозиса)*** $a_4$.

$$a_4 = \frac{R_4}{\sigma_{cm}^4} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - M_x)^4}{\sigma_{cm}^4} - 3 = \frac{1}{\sigma_{cm}^4}\int_{-\infty}^{\infty}(x - M_x)^4 f(x)dx - 3$$




Красным – нормальное распределение ($a_3$=0)
Синим – менее протяженное распределение ($a_3$<0).
Черным – более протяженное распределение ($a_3$>0).

***Квантиль*** - значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Т.е. квантиль можно рассматривать как обратную величину функции *F(x)*.