Лекция 2: Характеристика данных выборки и генеральной совокупности

1. Принципы подбора выборки
2. Гистограмма и полигон частот как приближение кривой распределения случайной величины
3. Параметры распределения и их влияние на вид кривой распределения

1 Принципы подбора выборки

Результат эксперимента - некоторая совокупность измерений, которую можно рассматривать как случайный вектор (вектор значений случайной величины).

Однократные измерения допускаются только в виде исключения!

Генеральная совокупность *– полный набор всех возможных значений, которые может принимать случайная величина*.

У исследователя никогда нет генеральной совокупности, а есть выборка ограниченного объема, по которой необходимо определить характеристики генеральной совокупности.

Выборка *– набор значений величины* $\{x_i\}$*, полученный из генеральной совокупности в результате конечного числа испытаний* N*. Количество данных в выборке – ее* ***объем****.*

Для проведения исследований необходимо, чтобы характер поведения данных в выборке как можно более точно повторял характер поведения данных в генеральной совокупности.

При отборе элементов выборки возможны ***ошибки репрезентативности***. Классический пример: «Литрери Дайджест», выборы президента США в 1936 г. выборка: подписчики + абоненты телефонного справочника + автовладельцы. Вернулось 2,5 млн бюллетеней

57% республиканец Альф Лэндон
40% демократ Франклин Рузвельт

выиграл Рузвельт
(более 60% голосов)

Репрезентативность выборки достигается ***рандомизацией*** или случайным отбором членов из генеральной совокупности. Это обеспечивает равную возможность для всех членов генеральной совокупности попасть в состав выборки. На практике применяются принципы частичной рандомизации.

Статистический анализ выборочных данных позволяет:
- дать для больших выборок общие характеристики, отражающие центральную тенденцию ($M(x)$, $D(x)$);
- сравнивать выборки, оценивать их общие характеристики, определять вероятность того, что различия вызваны случайными причинами;
- получить сведения о взаимосвязях элементов в выборке;
- применить результаты анализа для предсказания и описания.

2 Гистограмма и полигон частот как приближение кривой распределения случайной величины

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: $x=\{x_1,\ldots x_N\}$.

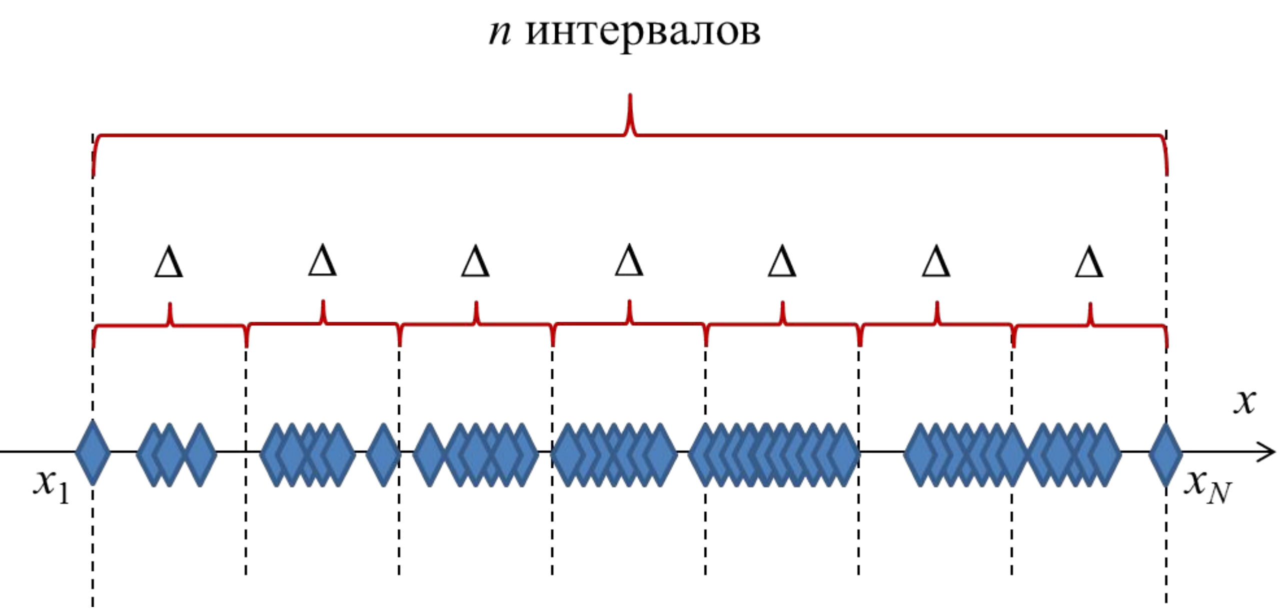
Алгоритм построения гистограммы и полигона частот

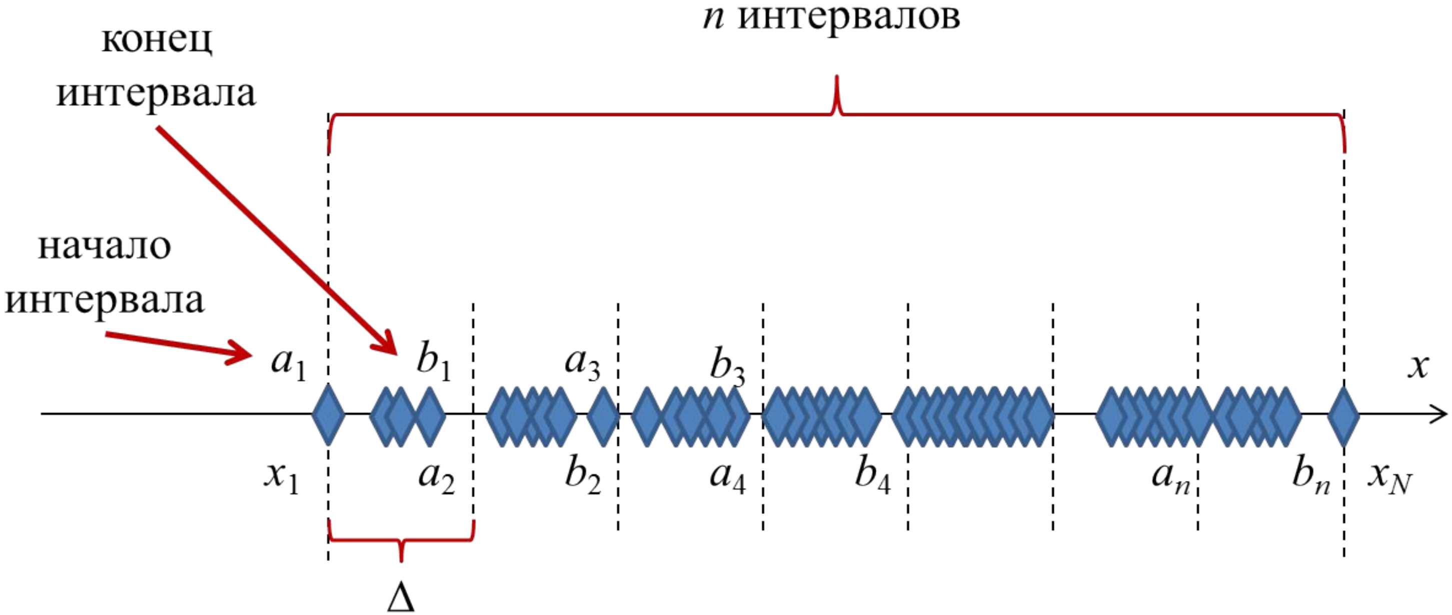
1. Построение вариационного ряда $x_1 \le x_2 \le \ldots \le x_N$
2. Группировка данных: разбиение отрезка $[x_1, x_N]$ на «карманы». Как и на сколько «карманов» разбивать?

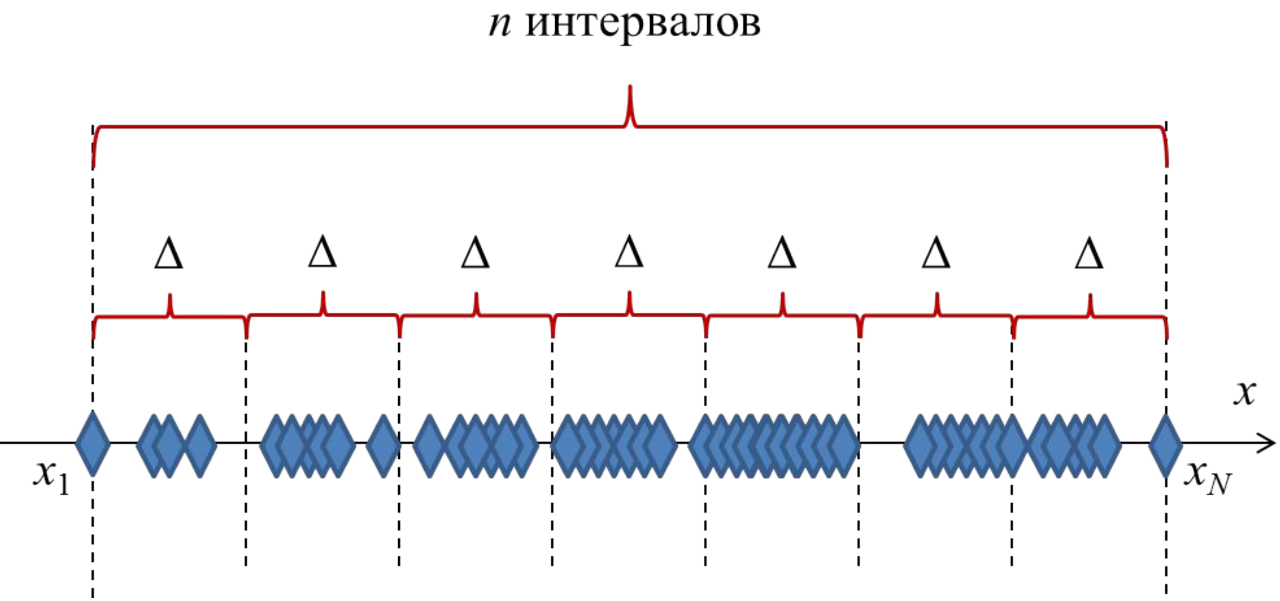
   Рассмотрим разбиение на «карманы» равной длины.

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: $n = \sqrt{N}$

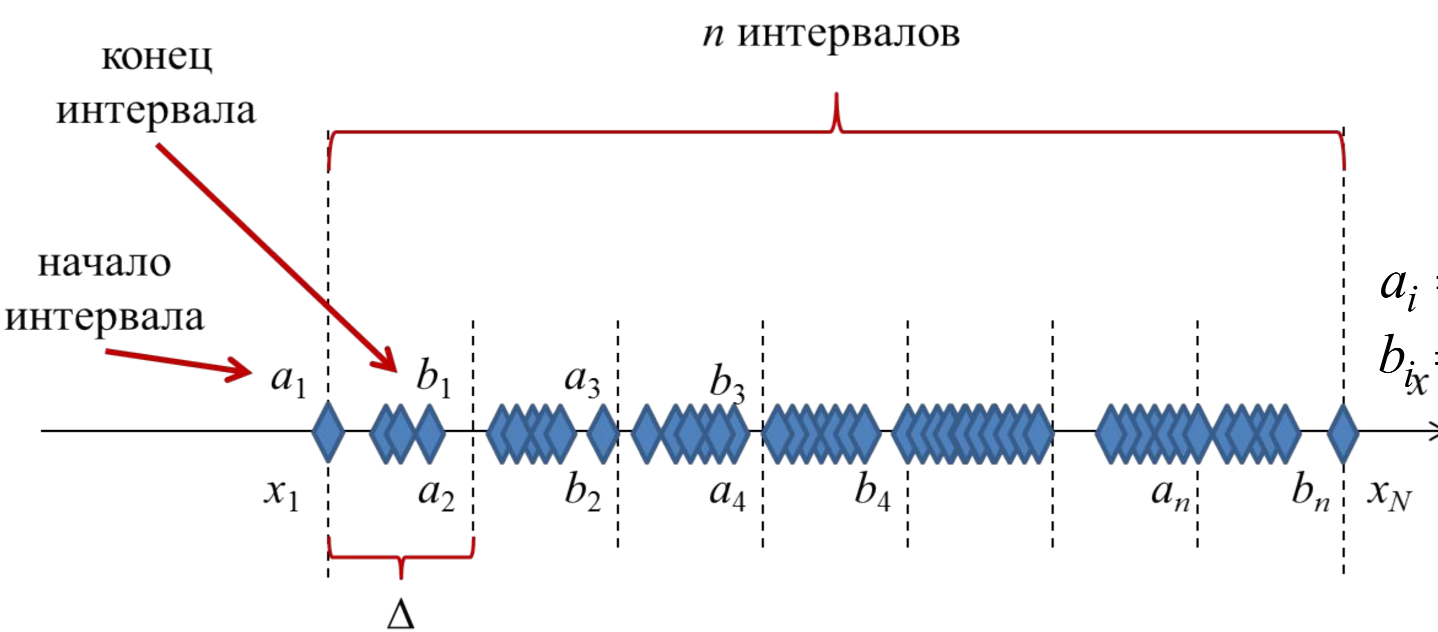

n интервалов
Δ Δ Δ Δ Δ Δ Δ
x
x_1
x_N


n интервалов
конец интервала
начало интервала
a_1 b_1 a_3 b_3
x
x_1 a_2 b_2 a_4 b_4 a_n b_n x_N
Δ



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

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



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

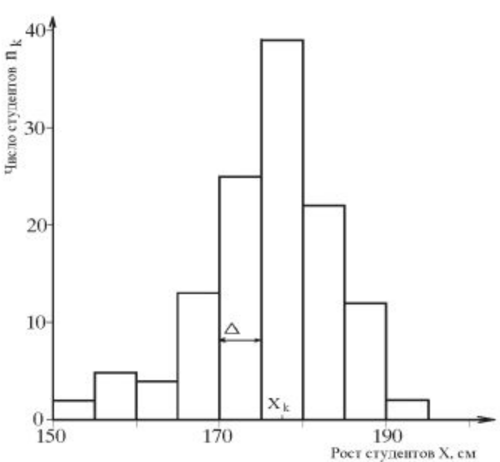
$$b_i = x_1 + i\Delta$$

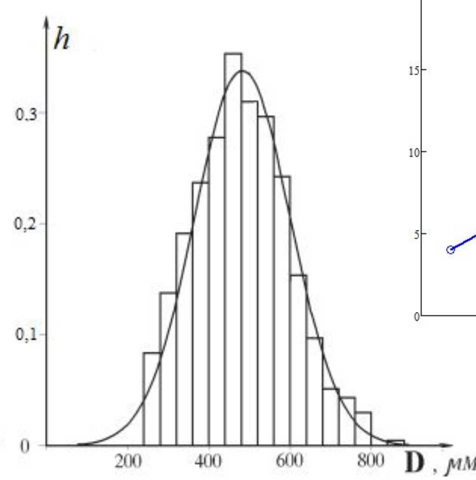
3. Вычисление числа значений, попавших в каждый интервал и построение (нормированной) ***гистограммы***

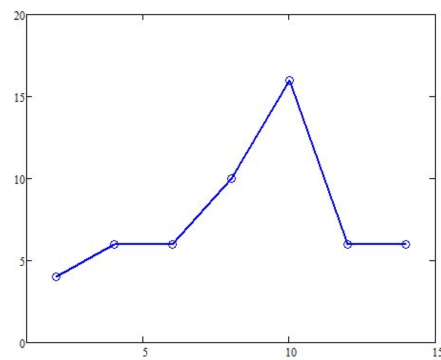
$$T_i = \sum_{j=1}^{N} t_{j,i}, \quad t_{j,i} = \begin{cases} 1, \text{ если } x_j \in [a_i, b_i], \\ 0, \text{ если } x_j \notin [a_i, b_i]. \end{cases} \qquad h_i = \frac{T_i}{N \cdot \Delta} \ \text{- нормировка } T_i$$

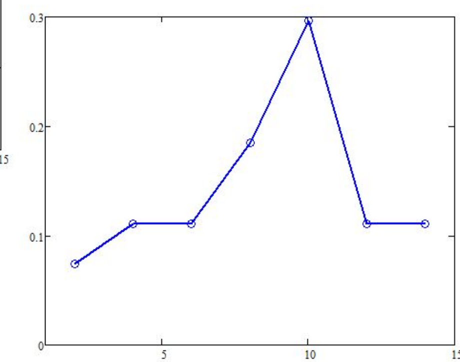
или

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,

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

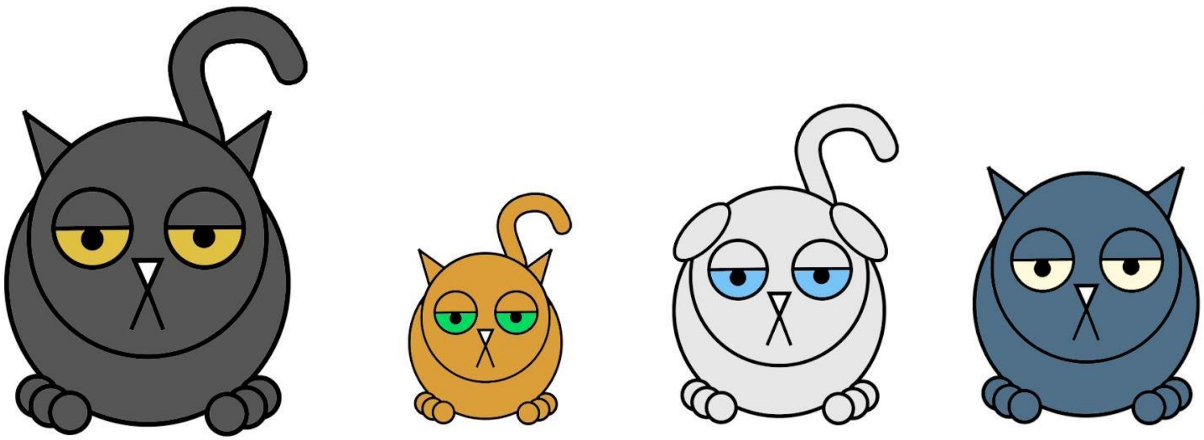
Вероятность попадания измеряемой величины в интервал (-∞, x] называют ***функцией распределения*** или ***интегральной функцией распределения***:

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

Исходя из определения,

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

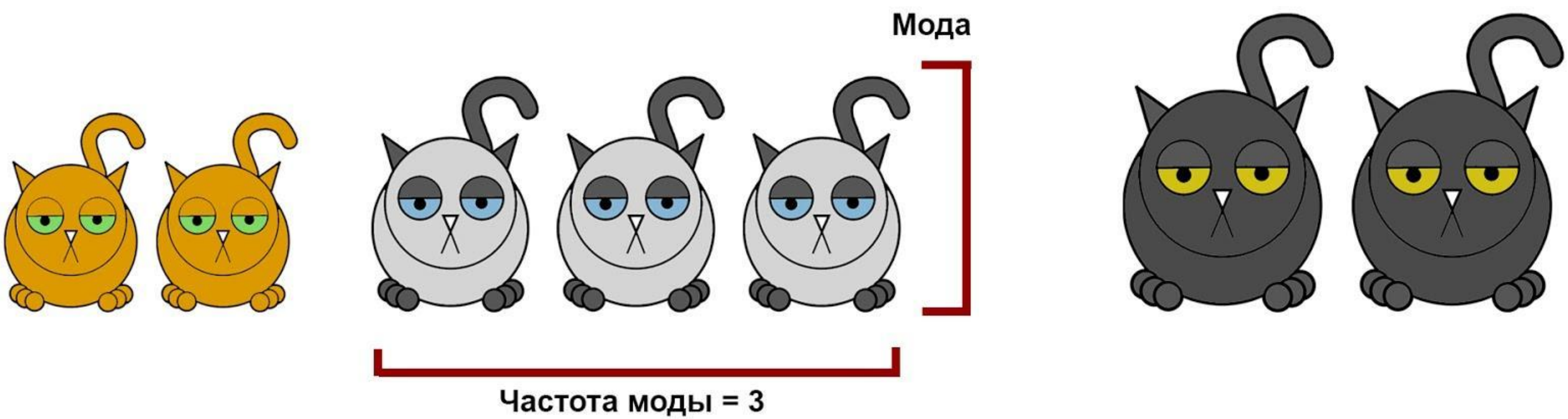
3 Параметры распределения и их влияние на вид кривой распределения (котики)



Котики бывают разные. Как же выглядит типичный котик?

Для простоты рассмотрим одно свойство котиков: ***размер***.

1 способ: какой размер котиков встречается чаще всего? Этот показатель называется **МОДА**



2 способ: упорядочить всех котиков по размеру и найти середину этого ряда. Как правило, там находится котик, который обладает самым типичным размером. И этот размер называется **МЕДИАНОЙ**.

Если по середине два котика (общее число котиков, *N* – четное)
МЕДИАНА = сложить размеры двух средних котов и поделить пополам

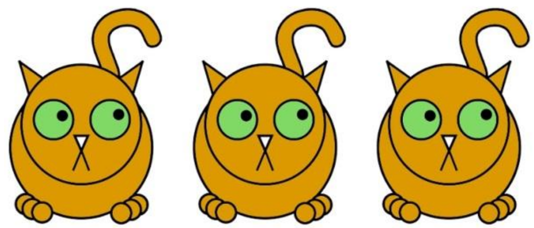
3 способ: сложить размер всех котиков, поделить на их количество – найти **СРЕДНЕЕ ЗНАЧЕНИЕ**.

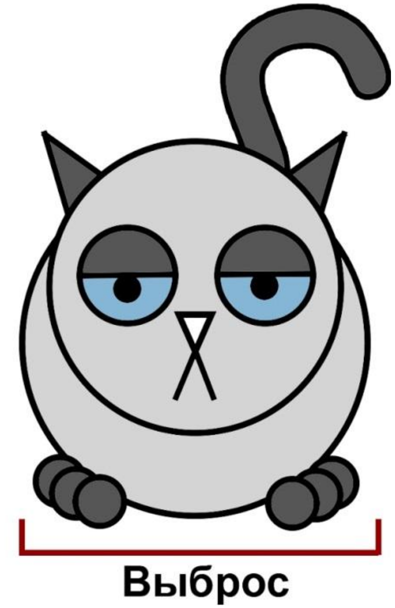
/ 3

Среднее значение $\bar{x}$

НО!

СРЕДНЕЕ ЗНАЧЕНИЕ чувствительно к **ВЫБРОСАМ** (при их наличии перестает отражать типичный котиковый размер)





Чтобы избавиться от **ВЫБРОСОВ**

а) либо убирают по 5—10% самых больших и самых маленьких котиков и уже от оставшихся считают среднее - **УСЕЧЕННОЕ (ИЛИ УРЕЗАННОЕ) СРЕДНЕЕ**;

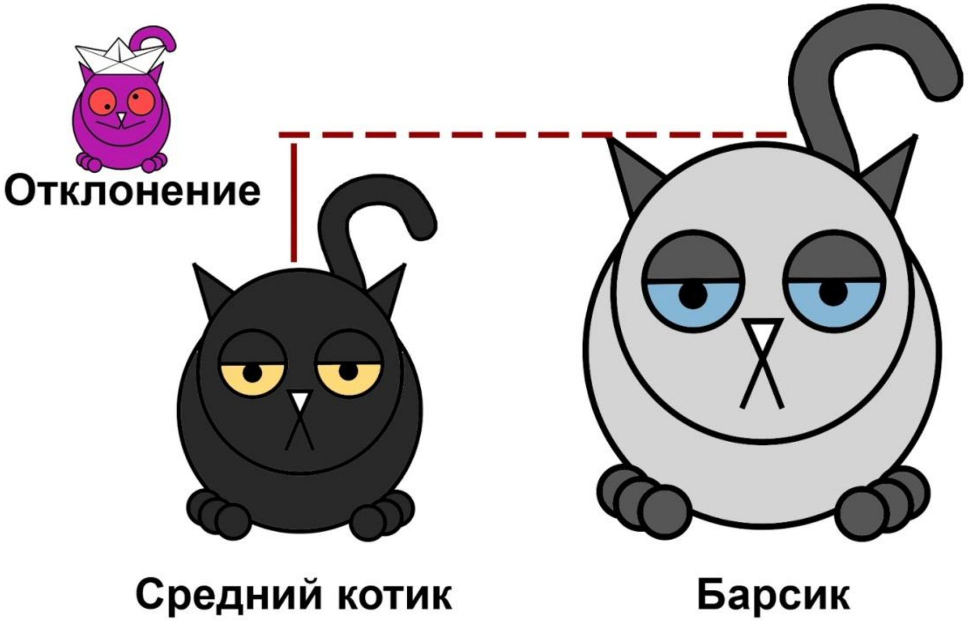
б) вместо **СРЕДНЕГО** используют **МЕДИАНУ**

МОДА, МЕДИАНА, СРЕДНЕЕ ЗНАЧЕНИЕ - это основные методы нахождения типичного размера котиков.
Все вместе они называются **МЕРАМИ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ**.

Кроме типичности нас часто интересует, насколько разнообразными могут быть котики по размеру. И в этом нам помогают **МЕРЫ ИЗМЕНЧИВОСТИ**:

1) **РАЗМАХ** - разность между самым большим и самым маленьким котиком. Эта мера очень чувствительна к выбросам.
Чтобы избежать искажений применяют **МЕЖКВАРТИЛЬНЫЙ РАЗМАХ** - отсеивают 25% самых больших и 25% самых маленьких котиков и найти размах для оставшихся.

2) **ОТКЛОНЕНИЕ** - разность между размером нашего конкретного котика (Барсика) и средним котиковым размером

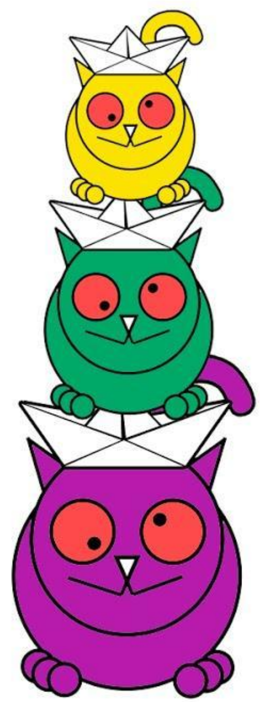


Чем крупнее (мельче) Барсик, тем больше **ОТКЛОНЕНИЕ**.

Чем больше котиков с **ОТКЛОНЕНИЕМ**, тем более разнообразны котики по размеру.

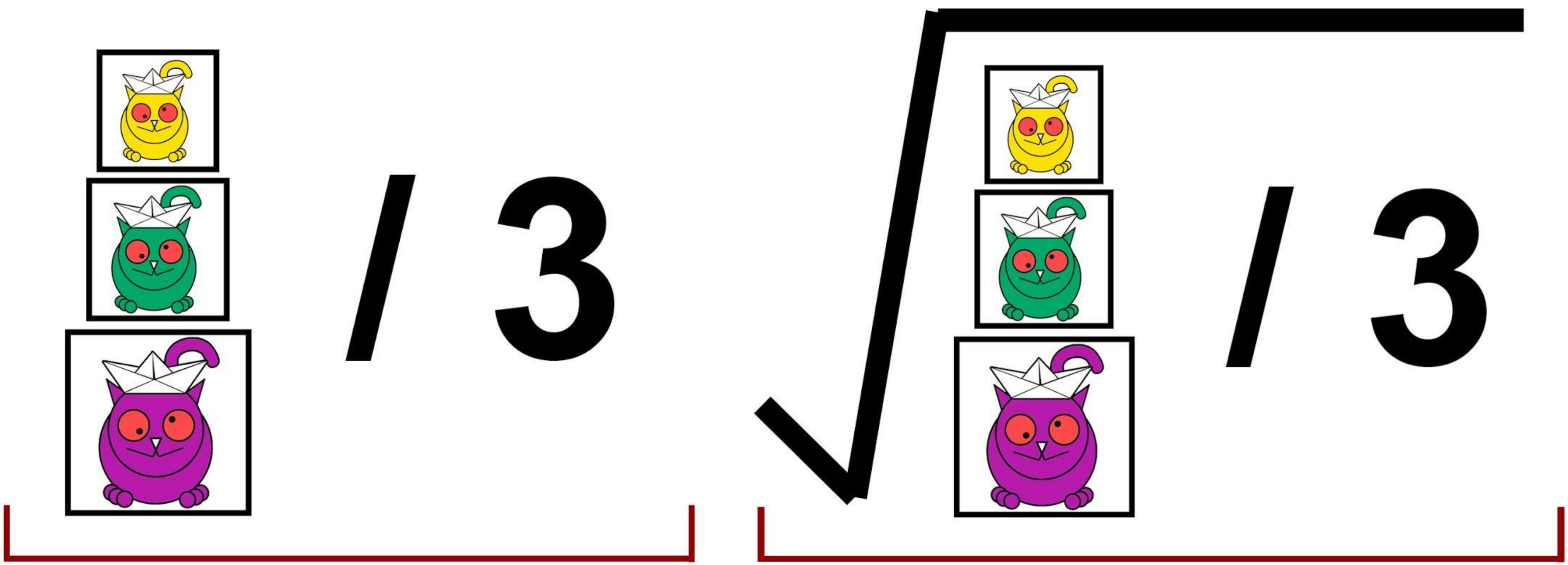
Какое **ОТКЛОНЕНИЕ** наиболее типично для котиков? Можно найти его **СРЕДНЕЕ ЗНАЧЕНИЕ!**

НО! СРЕДНЕЕ ЗНАЧЕНИЕ ОТКЛОНЕНИЙ = 0 (из-за знаков **ОТКЛОНЕНИЙ**)



/ 3

Избавиться от знака в математике можно двумя способами:
а) возвести в квадрат. Среднее от квадратов отклонений называется **ДИСПЕРСИЕЙ** (для оценки не сильно удобна, т.к. единицы измерения в квадрате)
б) взять корень квадратный из дисперсии и получить **СРЕДНЕКВАДРАТИЧЕСКОЕ ОТКЛОНЕНИЕ**



Дисперсия D **Среднеквадратическое отклонение S**

Обе меры чувствительны к **ВЫБРОСАМ.**

МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ и **МЕРЫ ИЗМЕНЧИВОСТИ** очень часто совместно используются для описания той или иной группы котиков, т.к. как правило большинство (около 68%) котиков находятся в пределе **СРЕДНЕКВАДРАТИЧЕСКОГО ОТКЛОНЕНИЯ** от **СРЕДНЕГО ЗНАЧЕНИЯ**.

Оставшиеся 32% либо очень большие, либо очень маленькие.

Для большинства котиковых признаков имеет место такая картина:

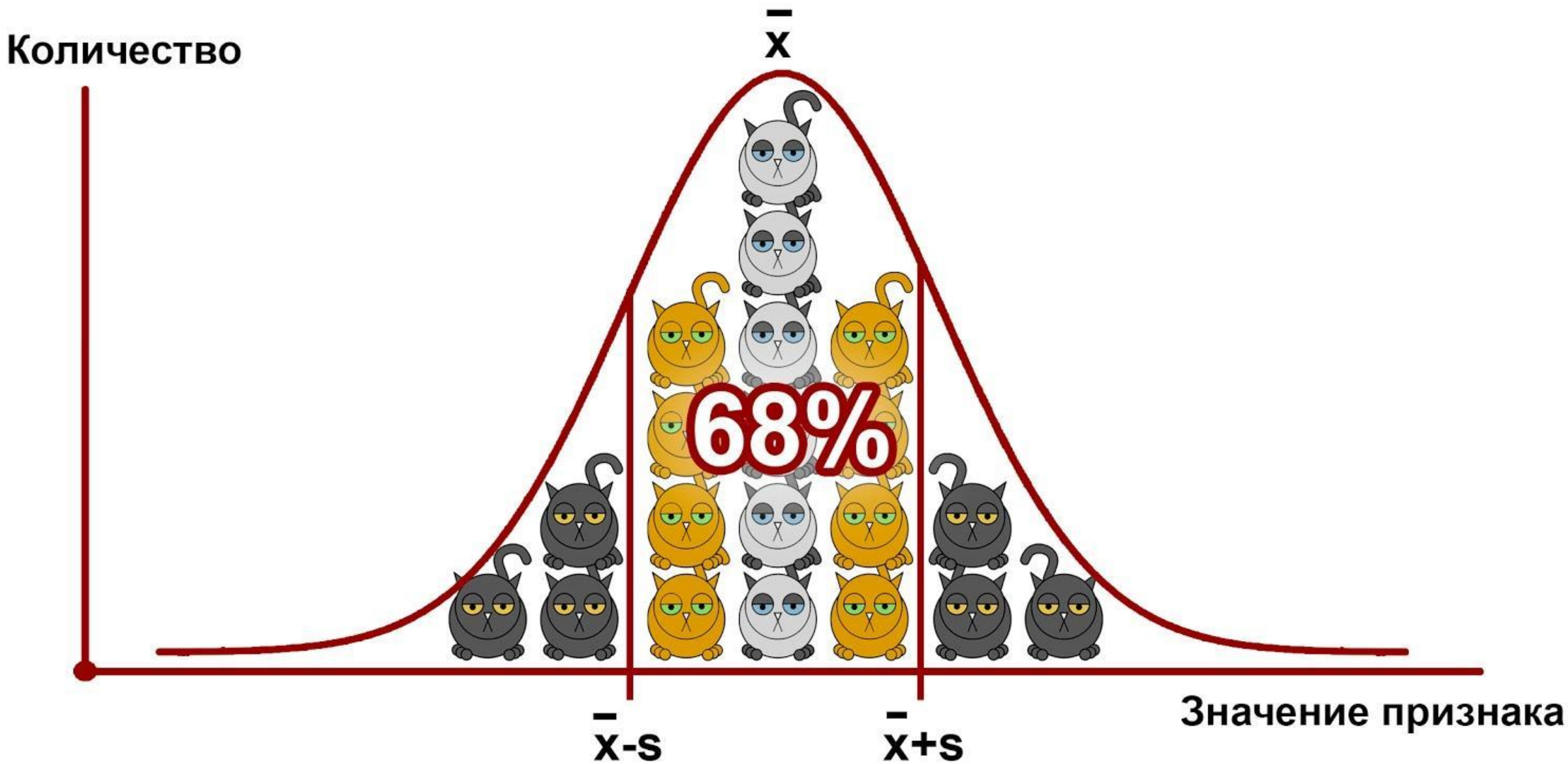


График называется **НОРМАЛЬНЫМ РАСПРЕДЕЛЕНИЕМ ПРИЗНАКА**.

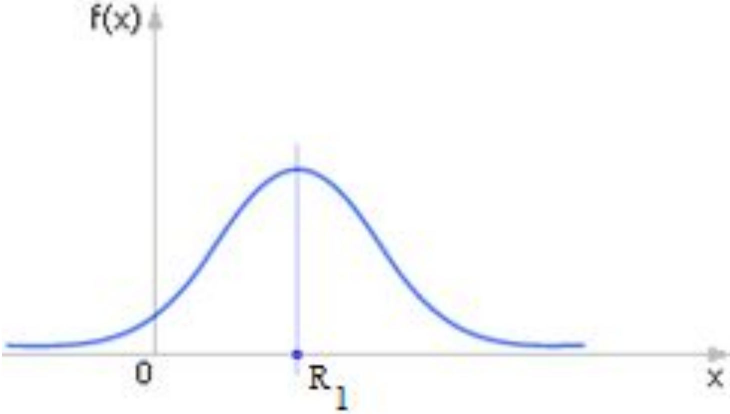
<u>Математически:</u>

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.



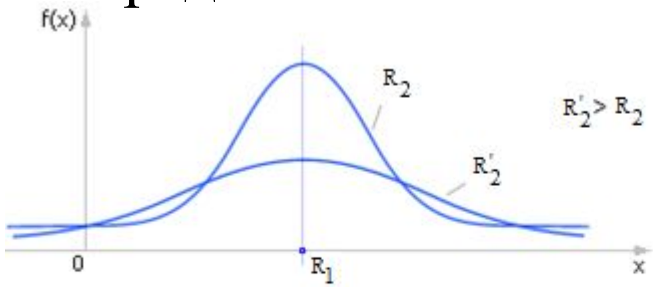
Медиана делит площадь, ограниченную функциеи плотности вероятности, на две равные части $P(X \le Me) = F(Me) = 0{,}5$

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:

$$\sigma = \sqrt{D} = \sqrt{\frac{\sum_{i=1}^{N}(x - M_x)^2}{N}}$$

Стандартное отклонение:

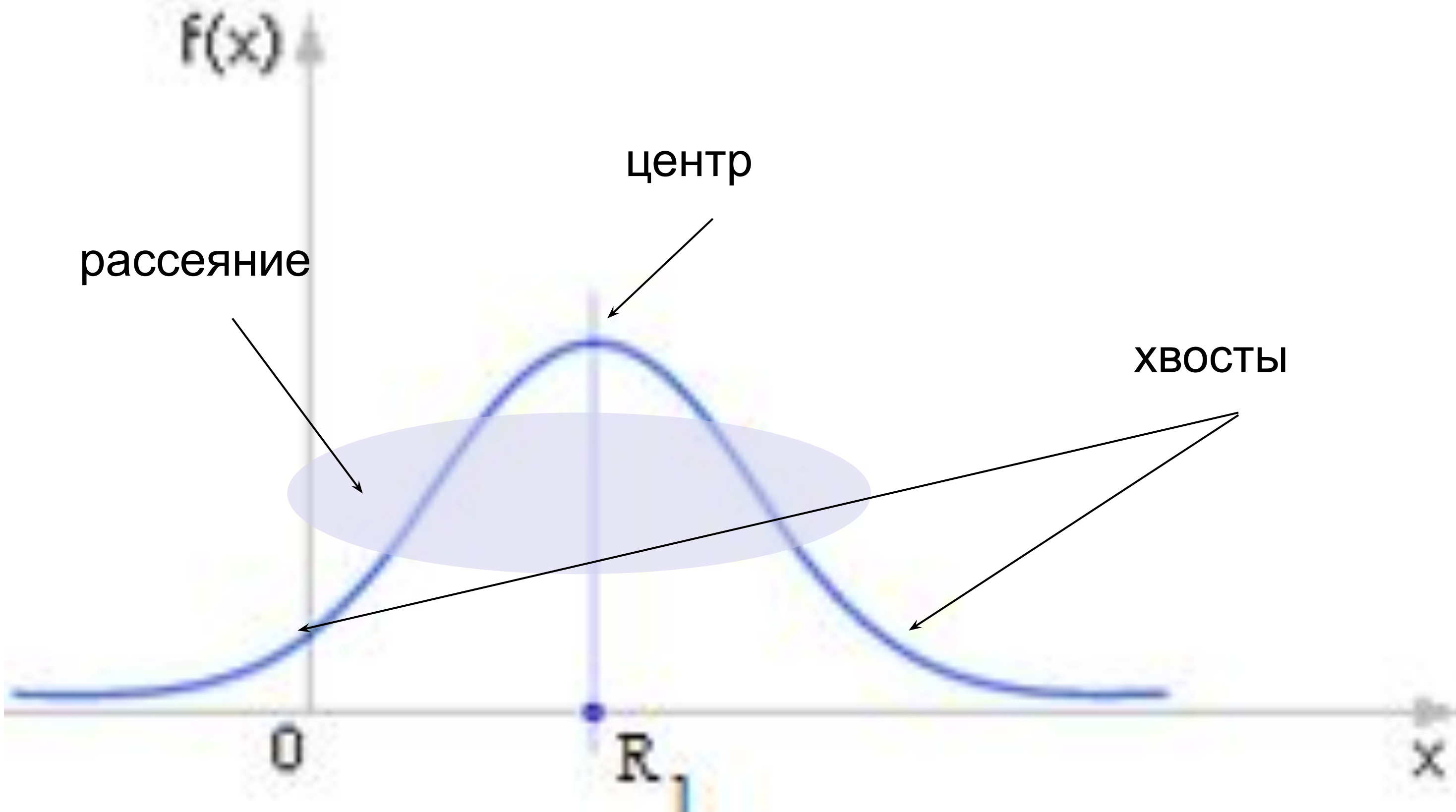
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x - M_x)^2}{N-1}}$$

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

Размах $w = x_{max} - x_{min}$

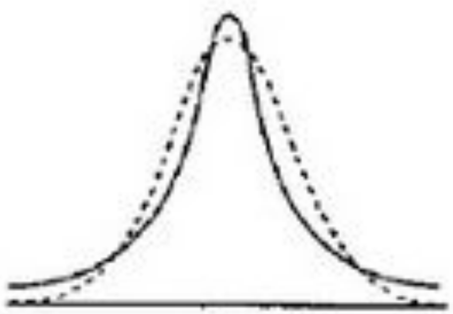
Другие меры для описания характера кривой распределения признака (распределения в обоих случаях сравниваются с нормальным):

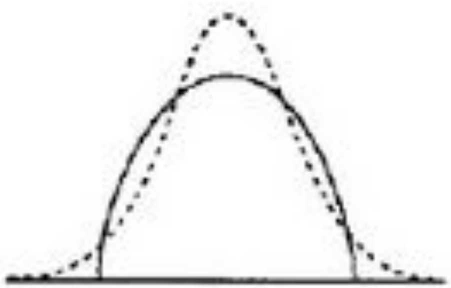
- симметричность распределения (к-т асимметрии);
- вес хвостов распределения (тяжелые или лёгкие – к-т эксцесса).



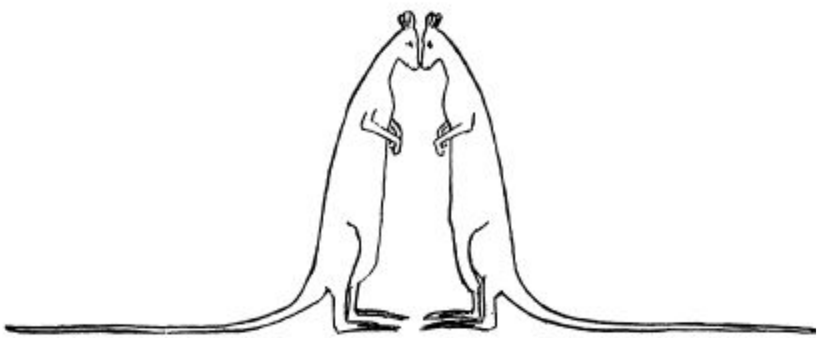
Вес хвоста распределения

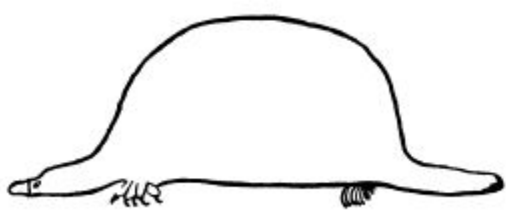
- «легкие» хвосты содержат лишь несколько значений. На графике плотности вероятности тонкие и длинные;
- «тяжелые» хвосты содержат довольно много значений. На графике выглядят толстыми.





Мнемоническое правило:

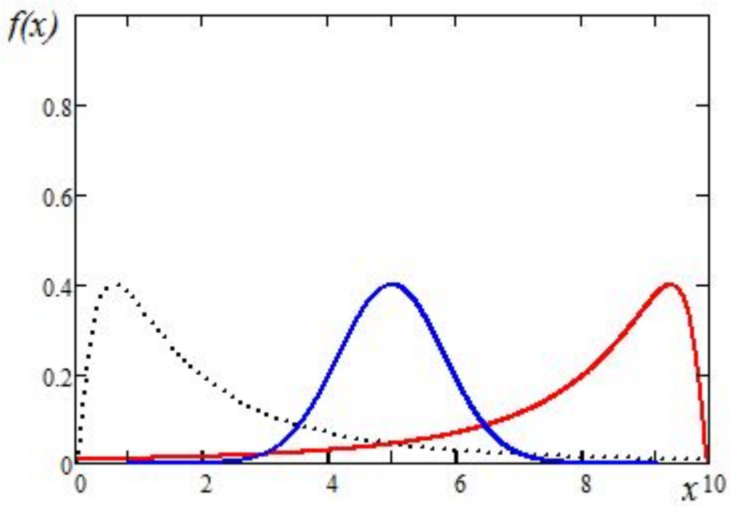




Скошенность распределения, когда один хвост кривой распределения крутой, а другой - пологий, характеризует ***коэффициент асимметрии***, a_3.

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

Скошенность нормального распределения = 0.



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

Вес хвостов распределения описывается ***коэффициентом эксцесса (куртозиса)*** a_4.

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

«-3» в формуле для того, чтобы облегчить сравнение с нормальным распределением.

У нормального распределения a_3=0;

у распределения с «легкими» хвостами a_3>0;

у распределения с «тяжелыми» хвостами 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)*.