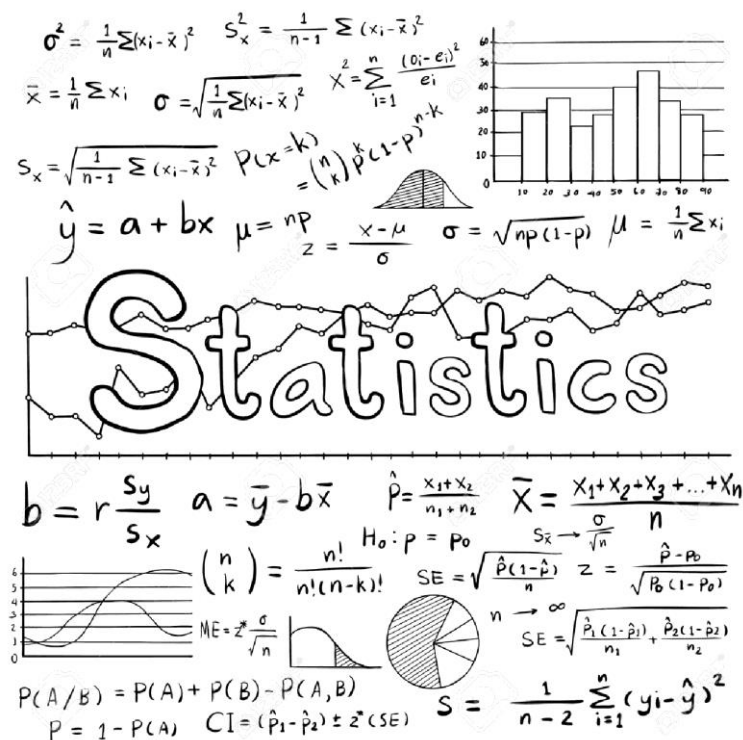


4. ПЕРВИЧНЫЙ АНАЛИЗ И ОБРАБОТКА ДАННЫХ

1. Основные параметры описательной статистики
2. Сравнение выборочных средних



1. Основные параметры описательной статистики

X
3,8
5,1
6,7
3,4
2,9
5,8
3,7
4,2
6,3
1,2

Переменная (Variance , Var)

Значения переменной (Cases)

Генеральная совокупность – это множество всех объектов, относительно которых предполагается делать выводы в рамках конкретного исследования. Генеральную совокупность составляют все объекты, которые отвечают всем заранее заданным параметрам.



Определенная часть генеральной совокупности, по которой обобщают результаты на всю генеральную совокупность - **выборка**

Закон больших чисел

С увеличением числа случайных величин их среднее арифметическое стремится к среднему арифметическому математических ожиданий и перестает быть случайным.

Необходимый объем выборки:

$$n = \left(\frac{z * \sigma}{H} \right)^2$$

где n – необходимый объем выборки, z – доверительный уровень (1,96 при $p=0,05$) σ – стандартное отклонение, H – допустимая ошибка (чаще всего – 3)

Соотношение объёма выборки и фактической ошибки

№ п/п	Показатели	Значения							
1	Объём выборки, если генеральная совокупность ≥ 5000	25	45	100	123	156	204	400	625...
2	Фактическая ошибка при данном объёме, %	20	15	10	9	8	7	5	4

Центральная предельная теорема

Распределение случайной величины, которая получена в результате сложения большого числа независимых случайных величин (ни одно из которых не доминирует, не вносит в сумму определяющего вклада и имеет дисперсию значительно меньшую по сравнению с дисперсией суммы) имеет распределение, близкое к нормальному.

Услуга	Доля	Выборка 1000 чел.		Выборка 1800 чел.		Выборка 1800 чел. + добор	
		Объём подвыборки	Ошибка, %	Объём подвыборки	Ошибка, %	Объём подвыборки	Ошибка, %
Кинотеатр	38%	380	5,0	684	3,7	684	3,7
Бары / кафе	32%	320	5,5	576	4,1	576	4,1
Развлекательный центр	25%	250	6,2	450	4,6	450	4,6
Боулинг / бильярд	24%	240	6,3	432	4,7	432	4,7
Аквапарк	23%	230	6,5	414	4,8	414	4,8
Парк аттракционов	22%	220	6,6	396	4,9	396	4,9
Мастер-классы	18%	180	7,3	324	5,4	324+68	4,9
Караоке	14%	140	8,3	252	6,2	252+140	4,9
Картинг	13%	130	8,6	234	6,4	234+158	4,9
Командные игры	12%	120	8,9	216	6,7	216+176	4,9
Ночной клуб	10%	100	9,8	180	7,3	180+212	4,9

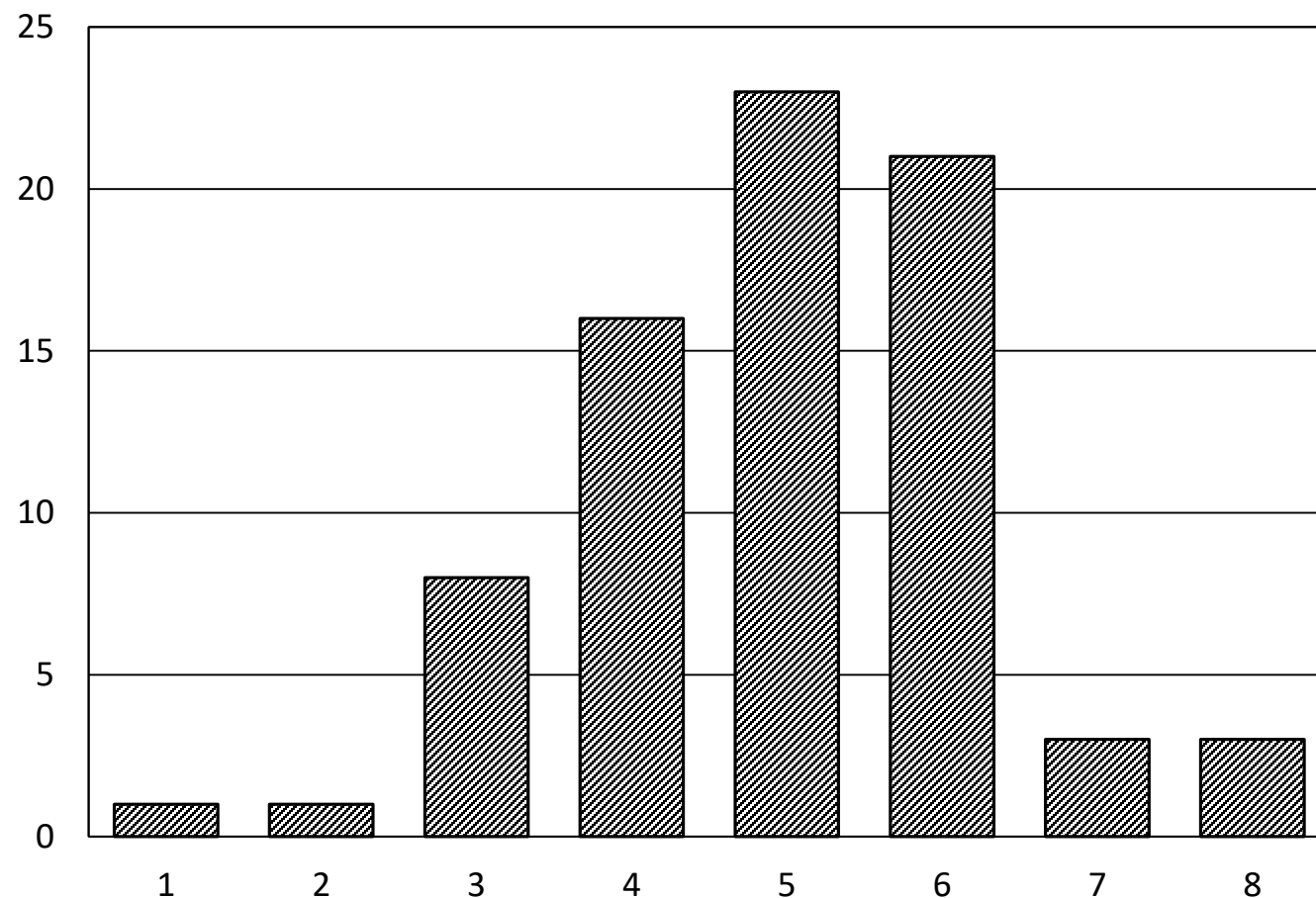
Распределение – это соотношение между значениями случайной величины (изучаемой переменной, признака) и частотой их встречаемости.

Основой для построения гистограммы служит **вариационный ряд** – ряд значений изучаемого признака, расположенных в порядке возрастания с соответствующими им частотами их встречаемости в выборке представленный в виде таблицы

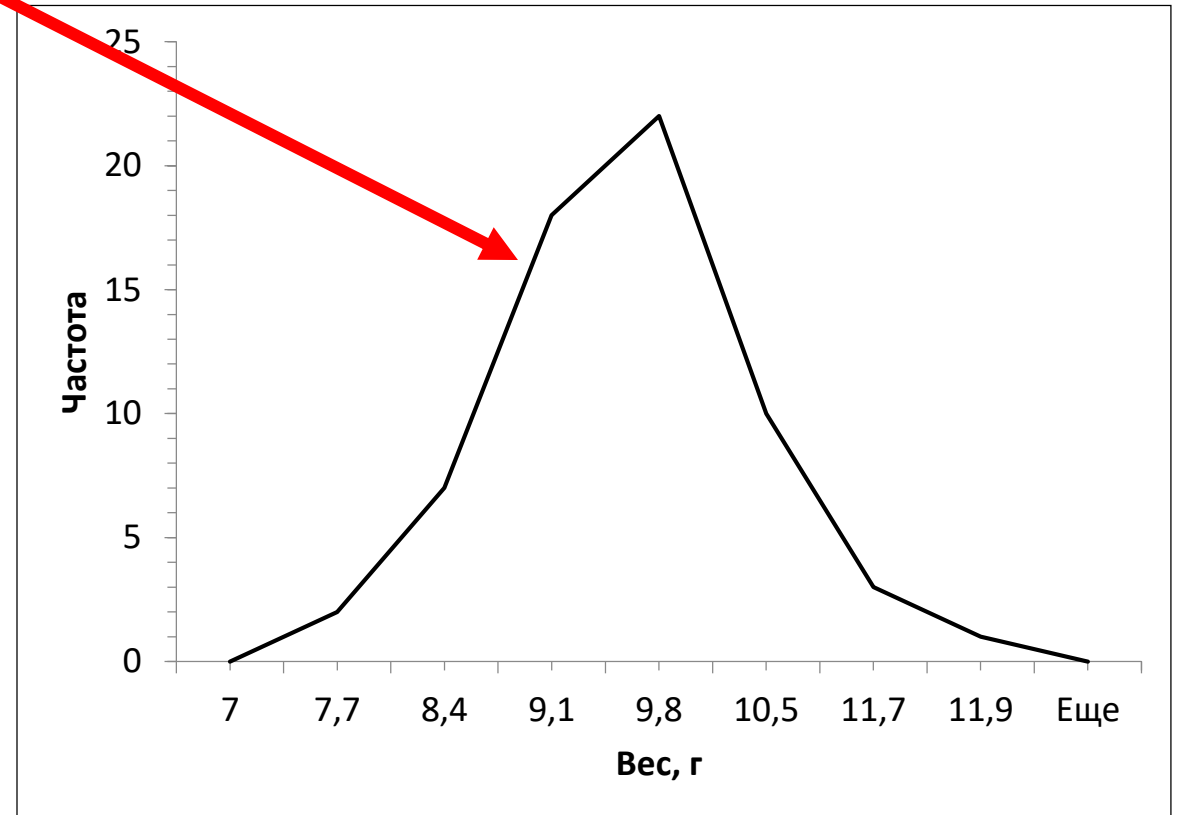
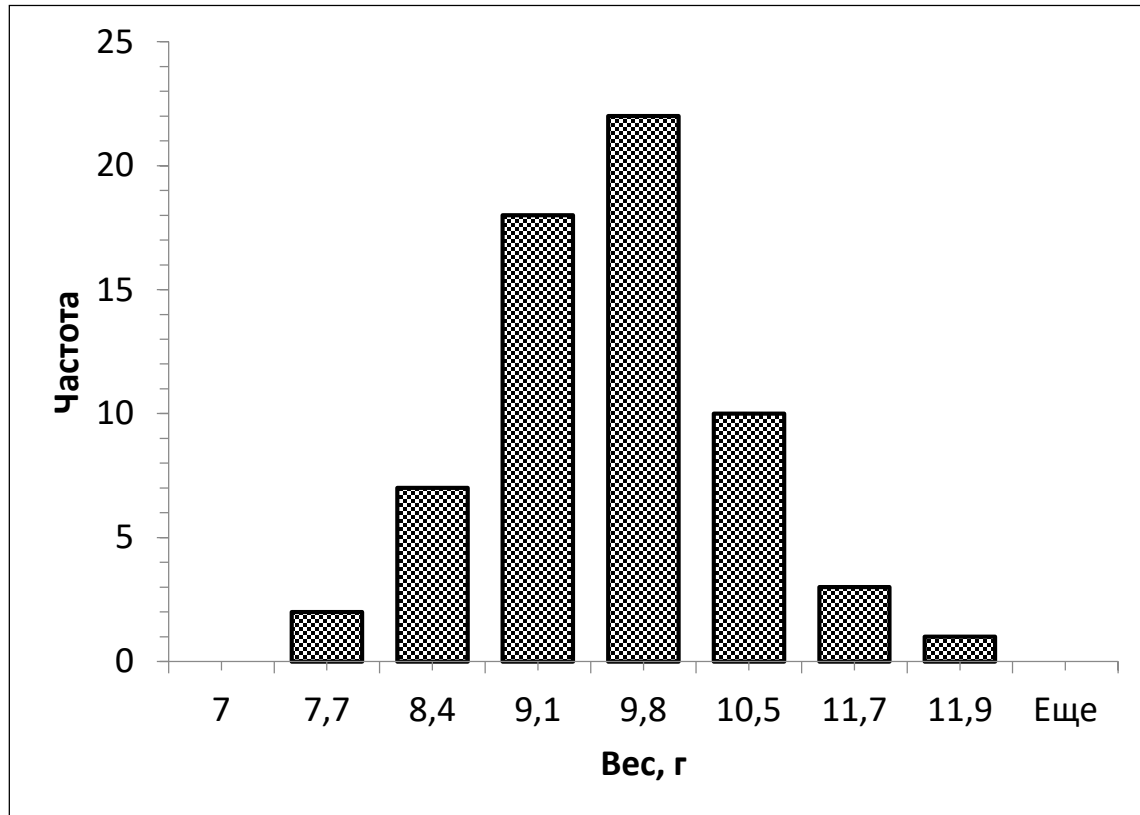
Таблица частот

Плодовитость, x	Частота, a
1	1
2	1
3	8
4	16
5	23
6	21
7	3
8	3

Гистограмма



Такая гистограмма строится в том случае, если признак дискретен, то есть имеет определенные предельные значения. Если же изучаемый признак непрерывен (таковы, например, размерно-весовые характеристики), то для построения вариационного ряда используется **полигон распределения**



Виды распределений

1) *Дискретные распределения:*

- а) биномиальное;
- б) геометрическое;
- в) гипергеометрическое;
- г) полиномиальное;
- д) Пуассона;
- е) Кокса;
- ж) альтернативное

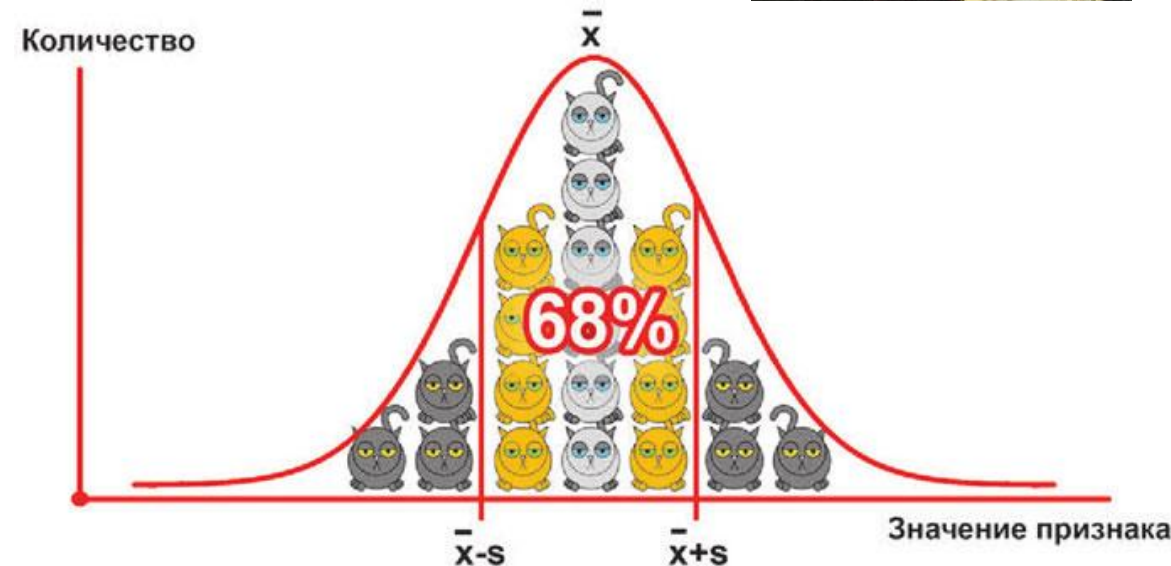
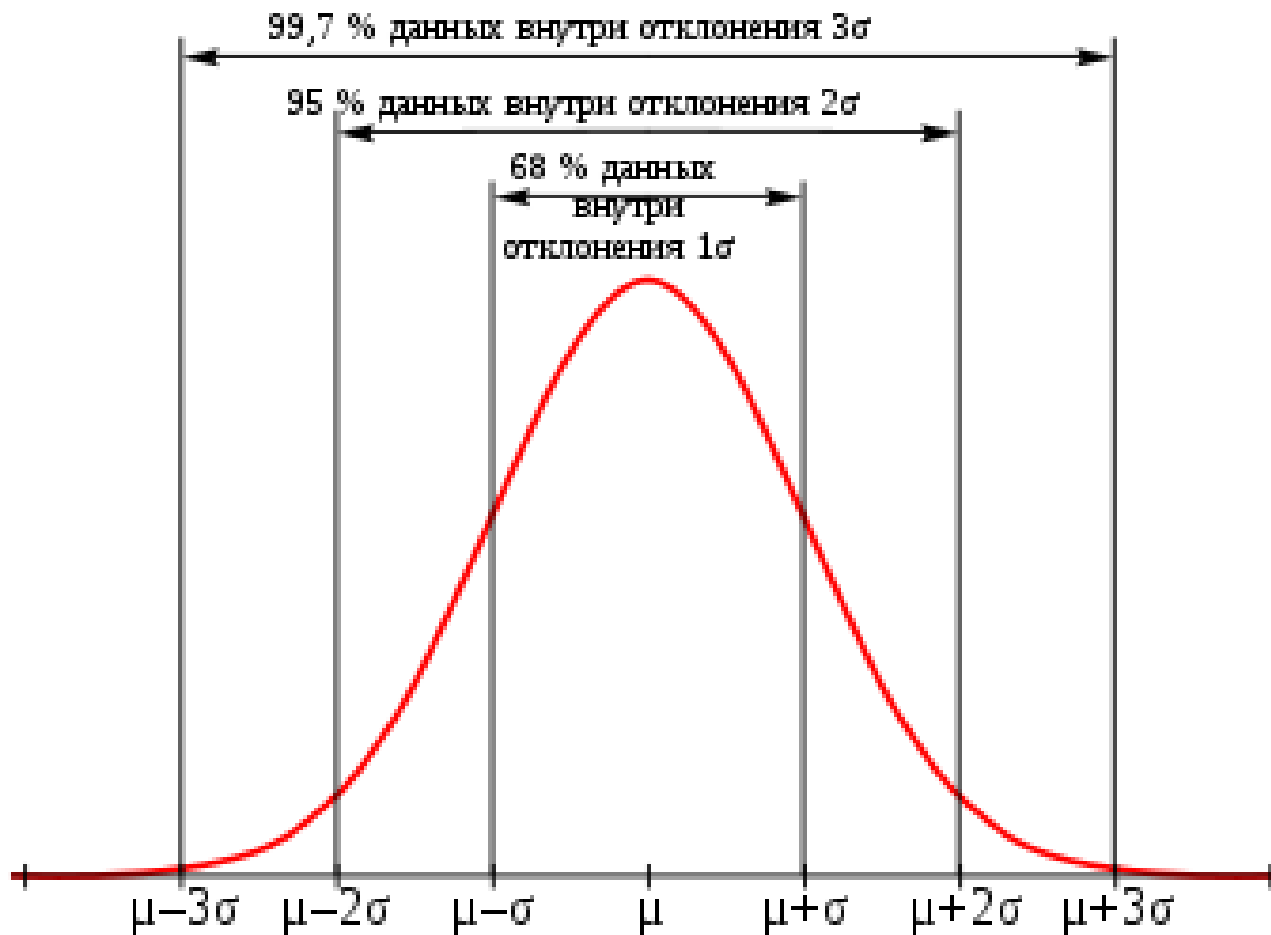
2) *Непрерывные распределения:*

- а) равномерное;
- б) нормальное, или Гауссово;
- в) многомерное нормальное;
- г) логарифмически нормальное;
- д) экспоненциальное;
- е) гамма;
- ж) Эрланга;
- з) Стьюдента (Уильяма Госсета);
- и) Фишера

Нормальное, или Гауссово распределение

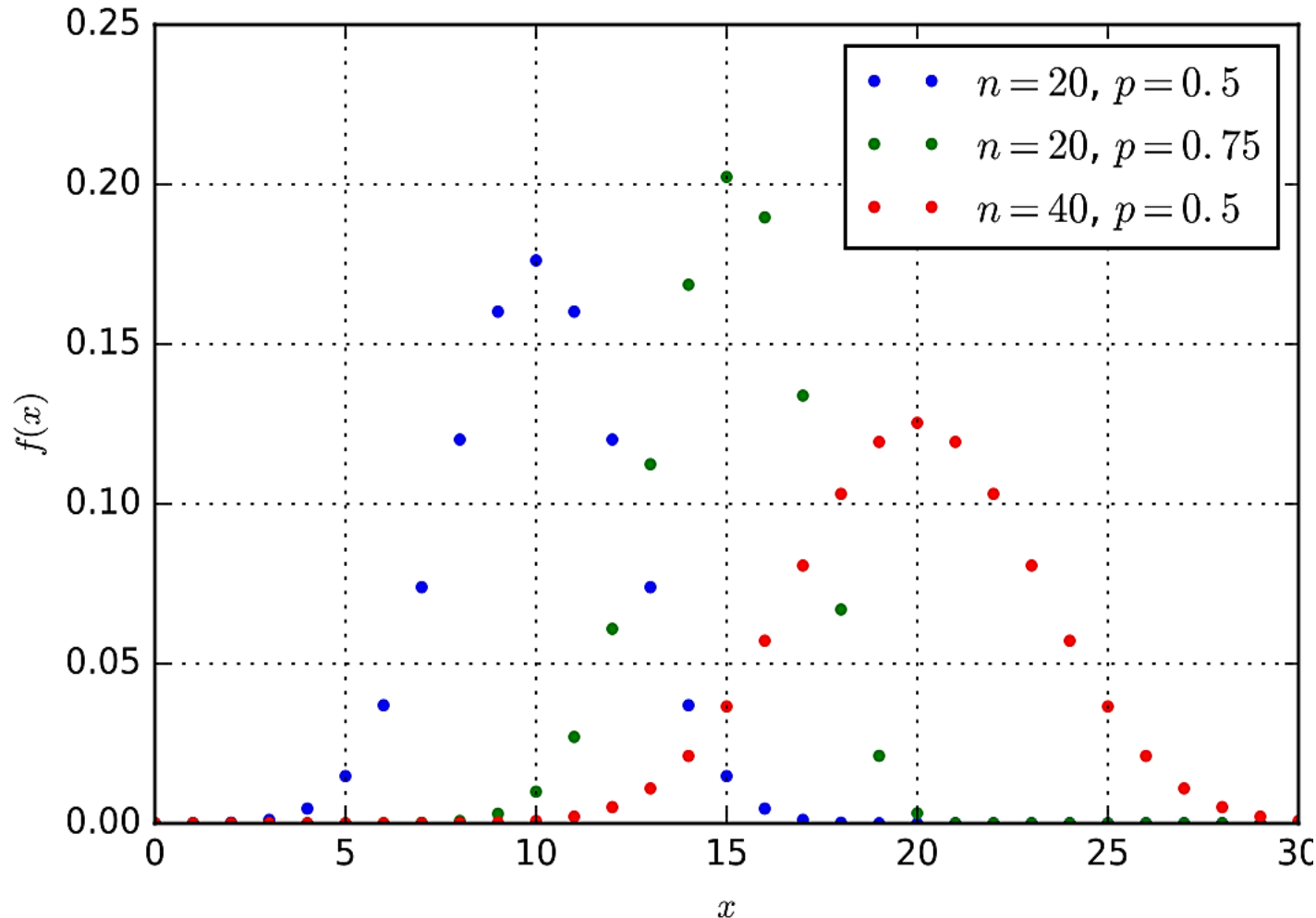
Это наиболее характерный тип распределения непрерывных случайных величин, из него можно вывести (к нему сводятся) все остальные. Распределение симметрично, причем крайние значения (наибольшие и наименьшие) появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем оно чаще встречается

Иоганн Карл Фридрих Гаусс (1777-1855) – немецкий математик, механик, физик, астроном и геодезист



Биноминальное распределение

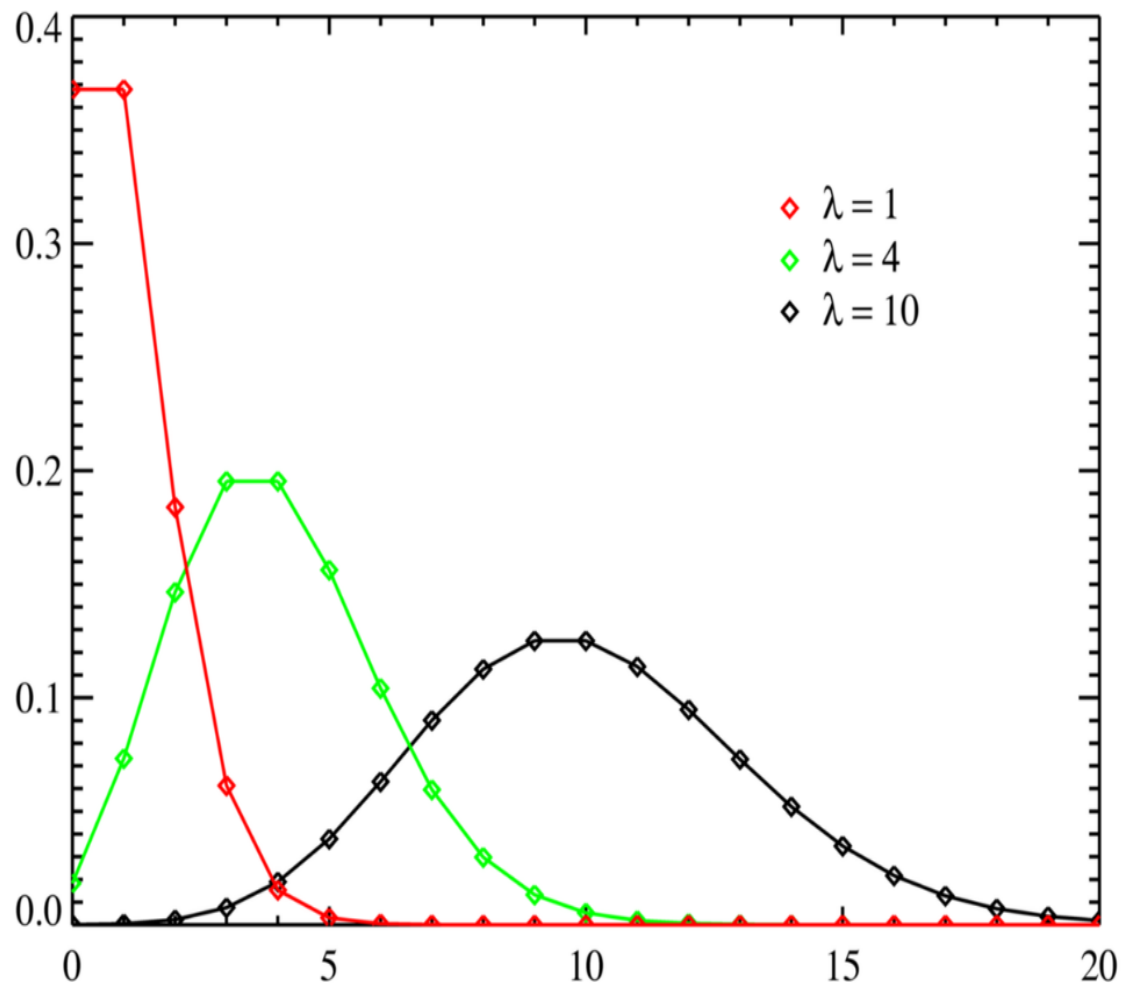
Во многом близко к нормальному. Отличие состоит лишь в том, что оно характеризует поведение дискретных признаков, выраженных целыми числами



Примерами описания признаков с помощью биномиального распределения могут служить число поврежденных участков на листьях, число волосков на единице площади шкурки, количество лучей в плавниках рыб, число хвостовых щитков у рептилий, плодовитость (размер выводка) самок и т. п. В основе биномиального распределения лежит альтернативное проявление качественного признака: он может присутствовать у единичного объекта или отсутствовать, проявиться или нет

Распределение Пуассона

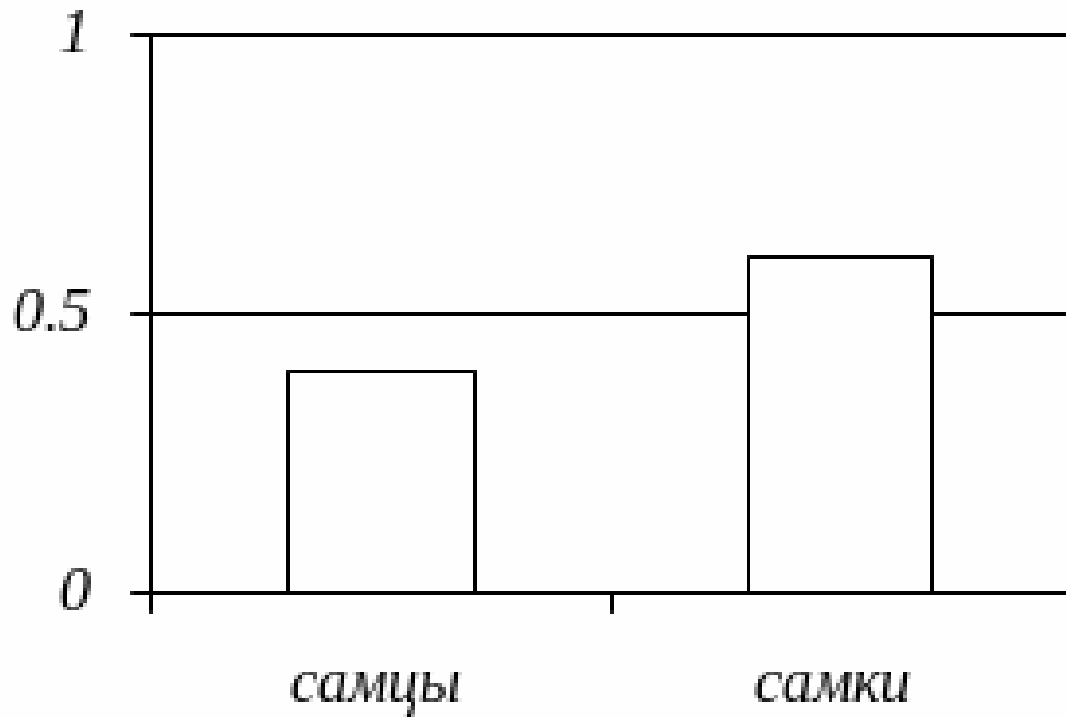
Является вариантом описания поведения дискретных количественных признаков для случаев, когда вероятность элементарных альтернативных событий неодинакова, одно из них наблюдается заметно чаще другого ($p \ll q$).



Поведение биологических объектов, соответствующее закону Пуассона, наблюдается в том случае, когда по пробам случайно распределены редкие объекты. Примеры таких явлений – частота нарушений хромосомного аппарата на каждую тысячу митозов, встречаемость семян сорняка в большой серии навесок семян культурного растения, число повторных попаданий животных в ловушки, встречаемость животных на отрезках длинных маршрутов (или на пробных площадках обширной территории), отловы животных в отдельные промежутки времени при длительных наблюдениях.

Альтернативное распределение

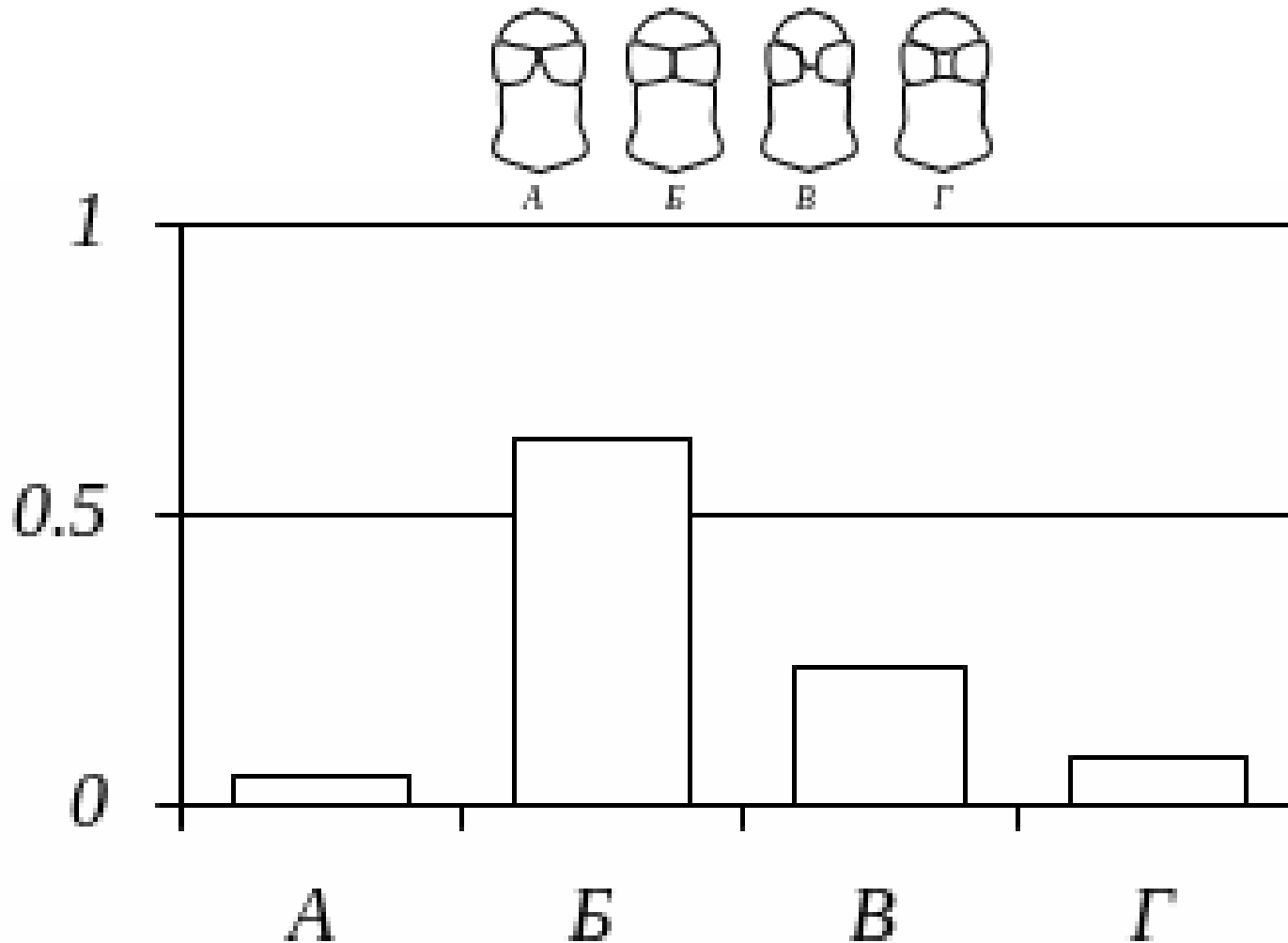
Это распределение дискретной случайной величины, имеющей лишь два противоположных (разнокачественных) значения (два класса, $k=2$). В одной пробе (в одном наблюдении) содержится одна варианта ($m=1$), одно из двух возможных значений. Вероятности каждого из них могут быть равны ($p=q$) либо не равны ($p < q$; $p > q$).



Примеры: самцы и самки, больные и здоровые организмы, сработавшие и пустые ловушки на одной учетной линии, два варианта аллельных признаков, вакцинированные и невакцинированные пациенты среди заболевших и др. Вычисления констант достаточно просты и не требуют построения вариационного ряда.

Полиномиальное распределение

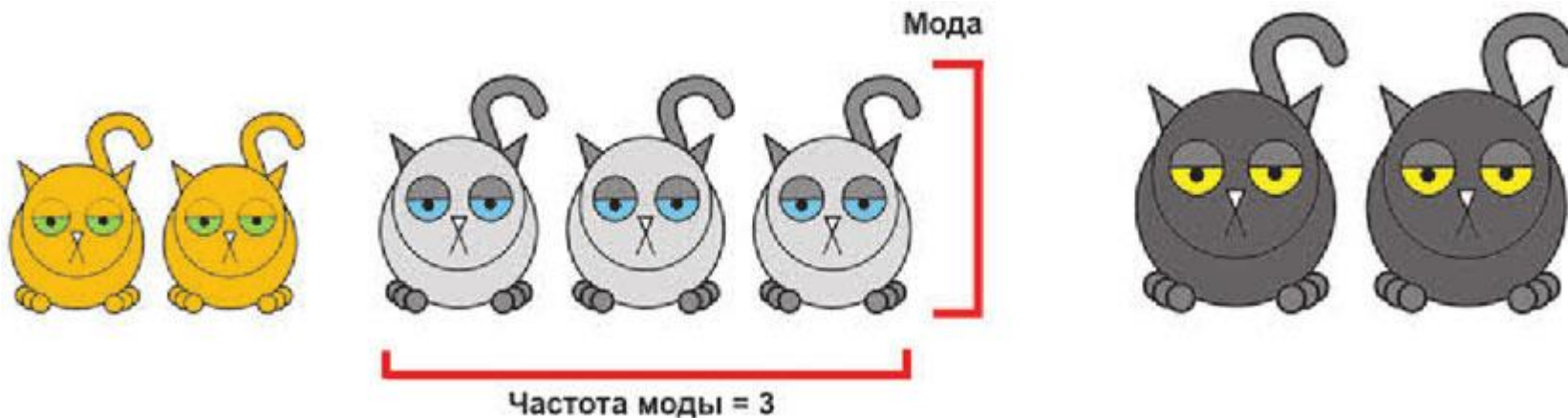
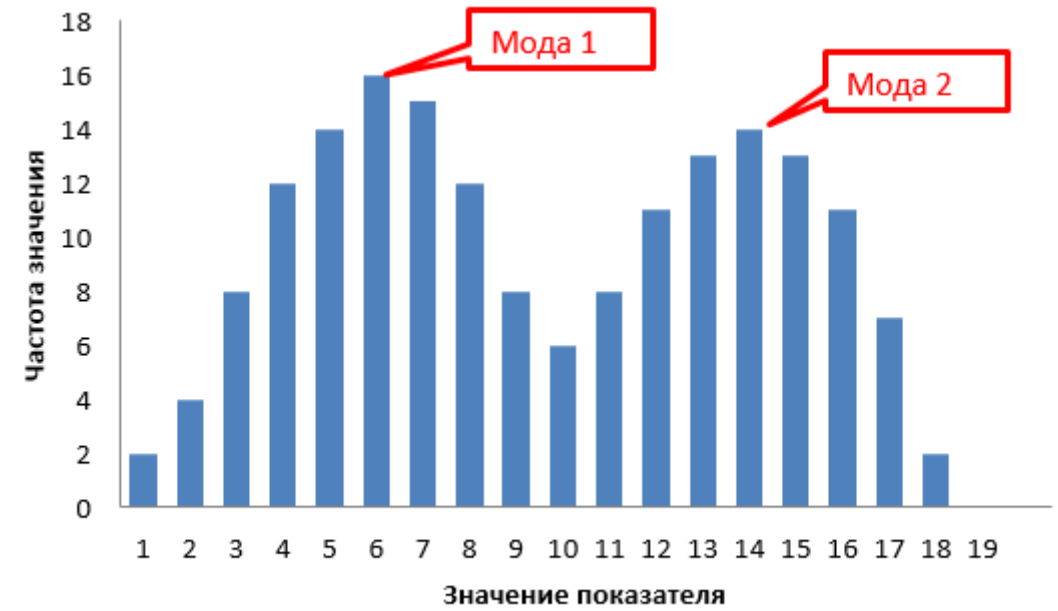
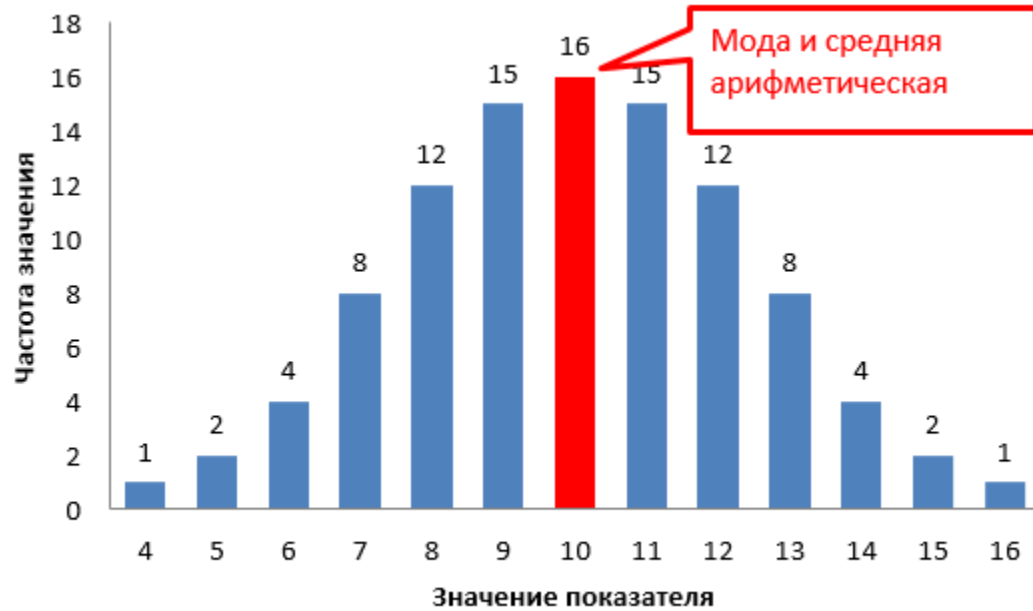
Наблюдается для качественных признаков, имеющих не два альтернативных свойства, но несколько возможных проявлений качества.



Полиморфизм популяций – типичный пример полиномиального распределения. В их числе варианты окраски покровов и волос, типы рисунков в определенных областях тела, способы жилкования листьев растений или крыльев насекомых, варианты расположения и формы щитков рептилий и другие проявления множественности фенотипов особей

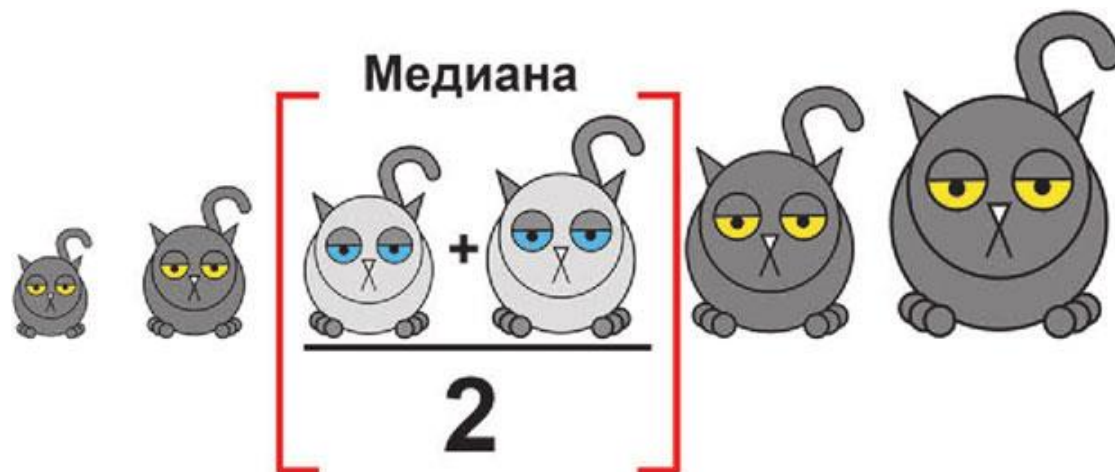
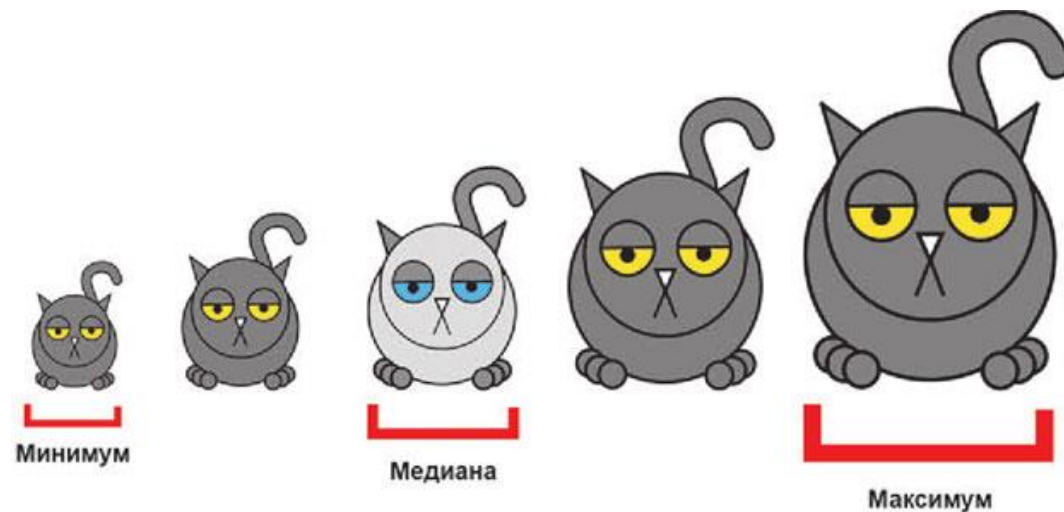
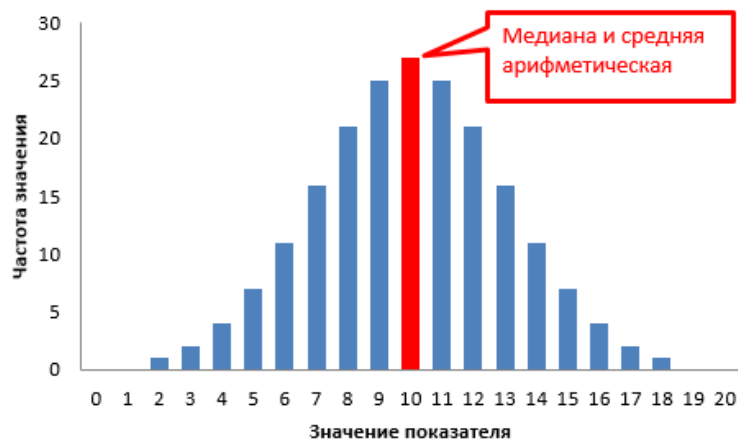
Мода

Мода, или **модус** – это варианта или класс распределения вариантов, который в исследуемой группе особей встречается наиболее часто.



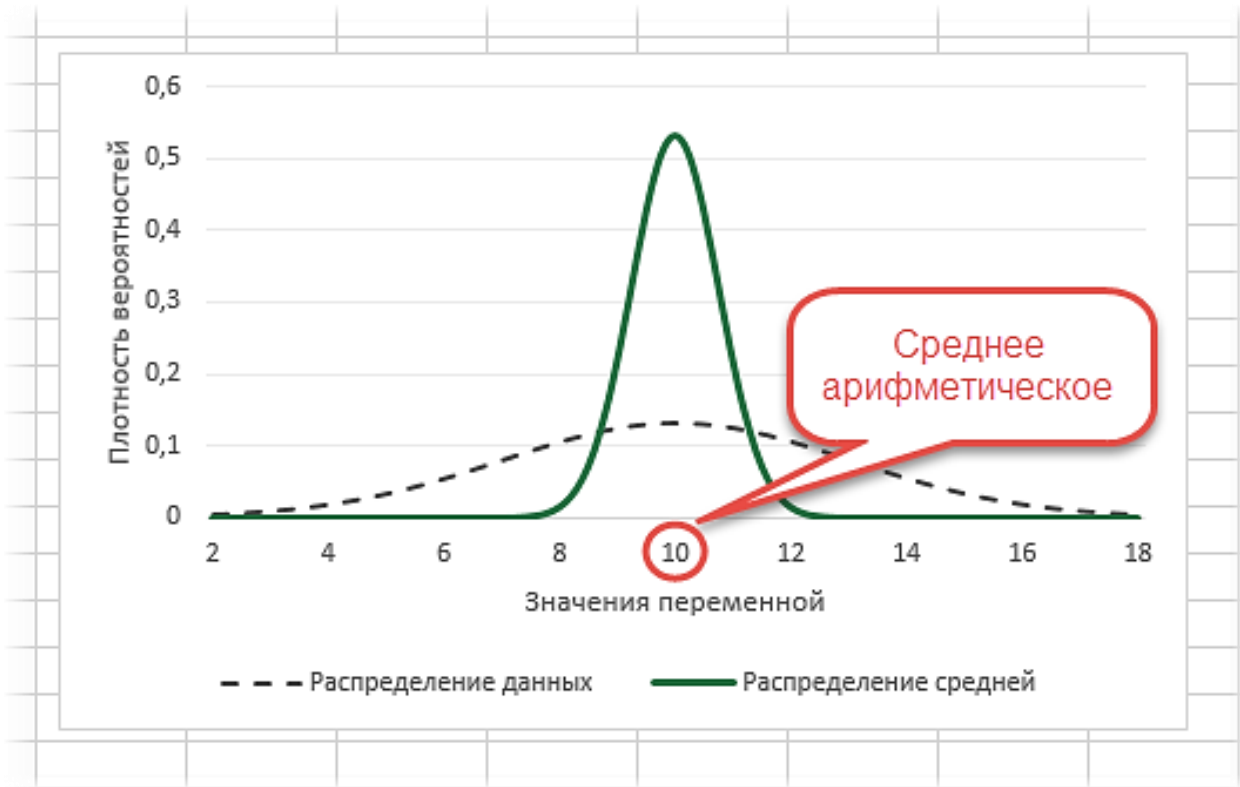
Медиана

Медиана – это такое значение признака, которое разделяет всю группу на две равные части: одна часть имеет значения признака меньшее, чем медиана, а другая – большее.



Среднее арифметическое выборки (выборочное среднее)

Среднее арифметическое – это сумма всех известных вариантов, разделённых на общее количество вариантов



$$\mu = \sum \frac{Xi}{n}$$



/ 3

$$T = \frac{\dot{Xi} - \mu}{\sigma} \geq Tst$$

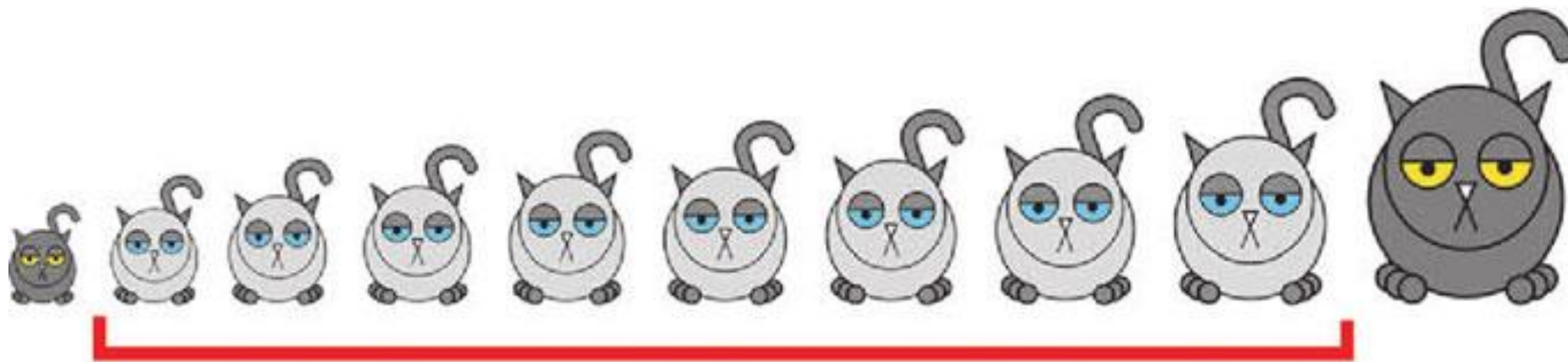


Выброс

Среднее значение $\bar{x}$

Усеченная средняя арифметическая

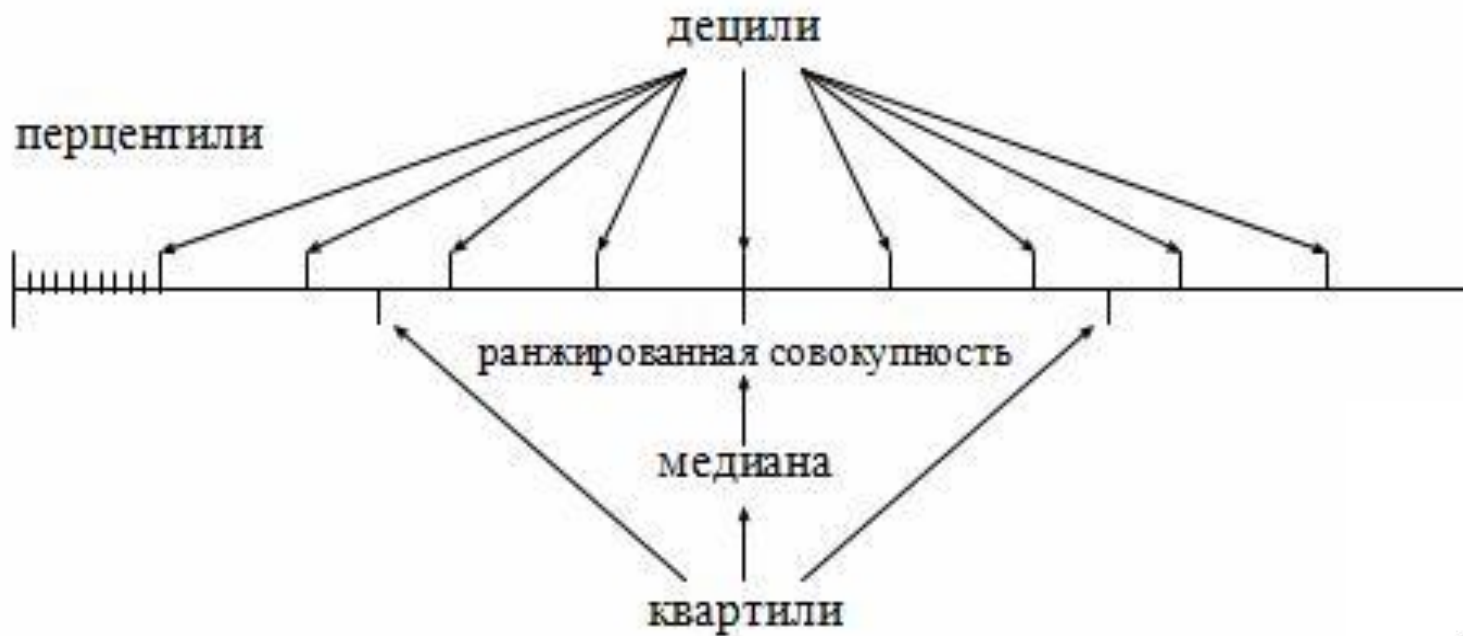
Для того, чтобы заранее подстраховаться от таких выбросов из выборки убирают 5-10% самых маленьких и самых больших значений, а затем от оставшихся значений считают среднее.



Котики для усеченного среднего

Квантили (квартили, децили, перцентили)

Квантиль – это конкретная варианта совокупности, отсекающая в пределах вариационного ряда определенную часть его членов. Медиана делит распределение на 2 равные части. Таким же образом распределение можно разделить тремя квартилями, чтобы получить 4 области равного размера или вероятности, или девятью децилями, чтобы получить 10 областей равного размера или вероятности, или 99 перцентилими чтобы получить 100 областей равного размера или вероятности. При этом 50-й перцентиль является медианой и вместе с 25-м и 75-м перцентилими даст нам квартили



На практике обычно используют перцентили. Причем перцентили P_{25} и P_{75} соответствуют первому и третьему квартилям, между которыми содержится 50% элементов выборки

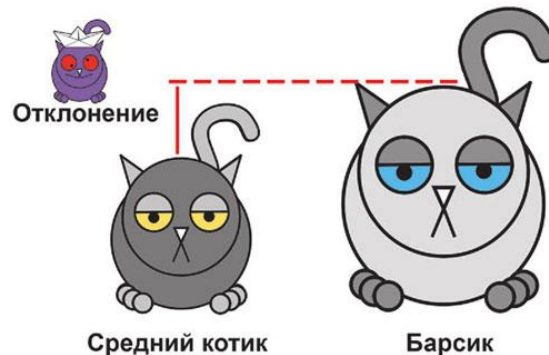


Дисперсия

Отклонения конкретного значения в выборке от среднего значения дают определенное разнообразие выборке. Чем больше будет в выборке значений с сильным отклонением, тем более разнообразными будет выборка и признак. Чтобы понять, какое отклонение является для выборки наиболее типичным, можно просто найти среднее значение по этим отклонениям (т. е. сложить все отклонения и поделить их на количество значений). Однако в этом случае результат будет равен 0, так как одни отклонения являются положительными (варианта больше среднего), а другие – отрицательными (варианта меньше среднего). Поэтому избавляются от знака либо взяв модуль отклонений, либо возведя их в квадрат. Среднее от квадратов отклонений и есть *дисперсия*.

$$\sigma^2 = \frac{\sum (Xi - \mu)^2}{n - 1}$$

Нужно запомнить, что существует 2 формулы дисперсии: одна для генеральной совокупности, другая – для выборки. В знаменателе первой всегда указывается точное количество значений признака, а у второй – ровно на одно значение меньше

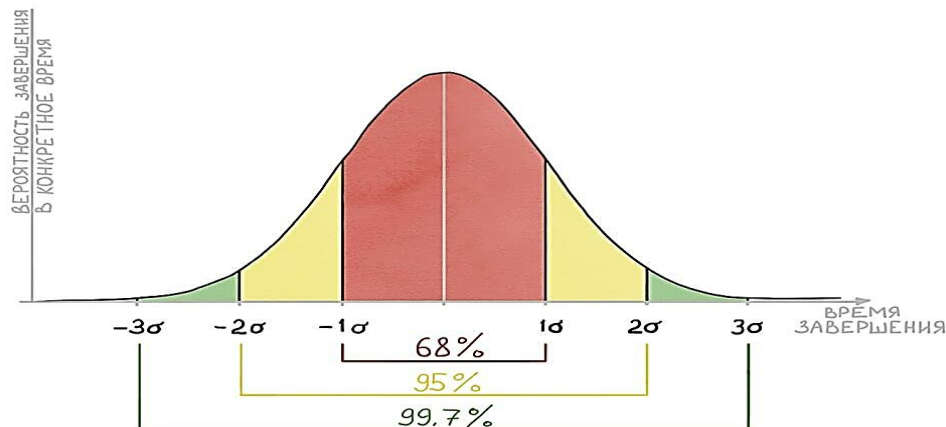


Стандартное отклонение

В связи с возведением в квадрат пользоваться дисперсией не всегда удобно, поэтому используется либо *стандартное отклонение* (если речь идет о выборке), либо *среднеквадратичное отклонение* (в случае генеральной совокупности)

$$\sigma = \sqrt{\frac{\sum (Xi - \mu)^2}{n-1}} - \text{стандартное отклонение}$$

$$\sigma = \sqrt{\frac{\sum (Xi - \mu)^2}{N}} - \text{среднеквадратичное отклонение}$$



A stylized representation of the formula for Root Mean Square Deviation (RMSD). It features a large square root symbol $\sqrt{\quad}$. Inside the radical, there are three cartoon cat icons (yellow, green, and purple) stacked vertically, followed by a large division sign $/$ and the number 3 . Below the radical, a red bracket spans the width of the radical, with the text 'Среднеквадратическое отклонение σ ' (Root Mean Square Deviation σ) underneath.

2. Сравнение выборочных средних

2.1 Параметрический критерий Стьюдента

Поскольку выборочные средние имеют нормальное распределение, критерий отличия двух выборочных средних также базируется на свойствах нормального распределения: в границах $\mu_{\text{общ.}} \pm 1.96 \cdot m$ (или приблизительно $\mu_{\text{общ.}} \pm 2 \cdot m$) выборочные средние арифметические отличаются от общей (генеральной) средней по случайным причинам. Тогда рабочая формула для t критерия отличия средних для *независимых выборок* будет:

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{m_1^2 + m_2^2}}$$

Парный t -критерий Стьюдента – одна из модификаций метода Стьюдента, используемая для определения статистической значимости различий парных (повторных) измерений, характерных для *зависимых выборок*. Парный t -критерий Стьюдента рассчитывается по следующей формуле:

$$t = \frac{Md}{\sigma d / \sqrt{n}}$$

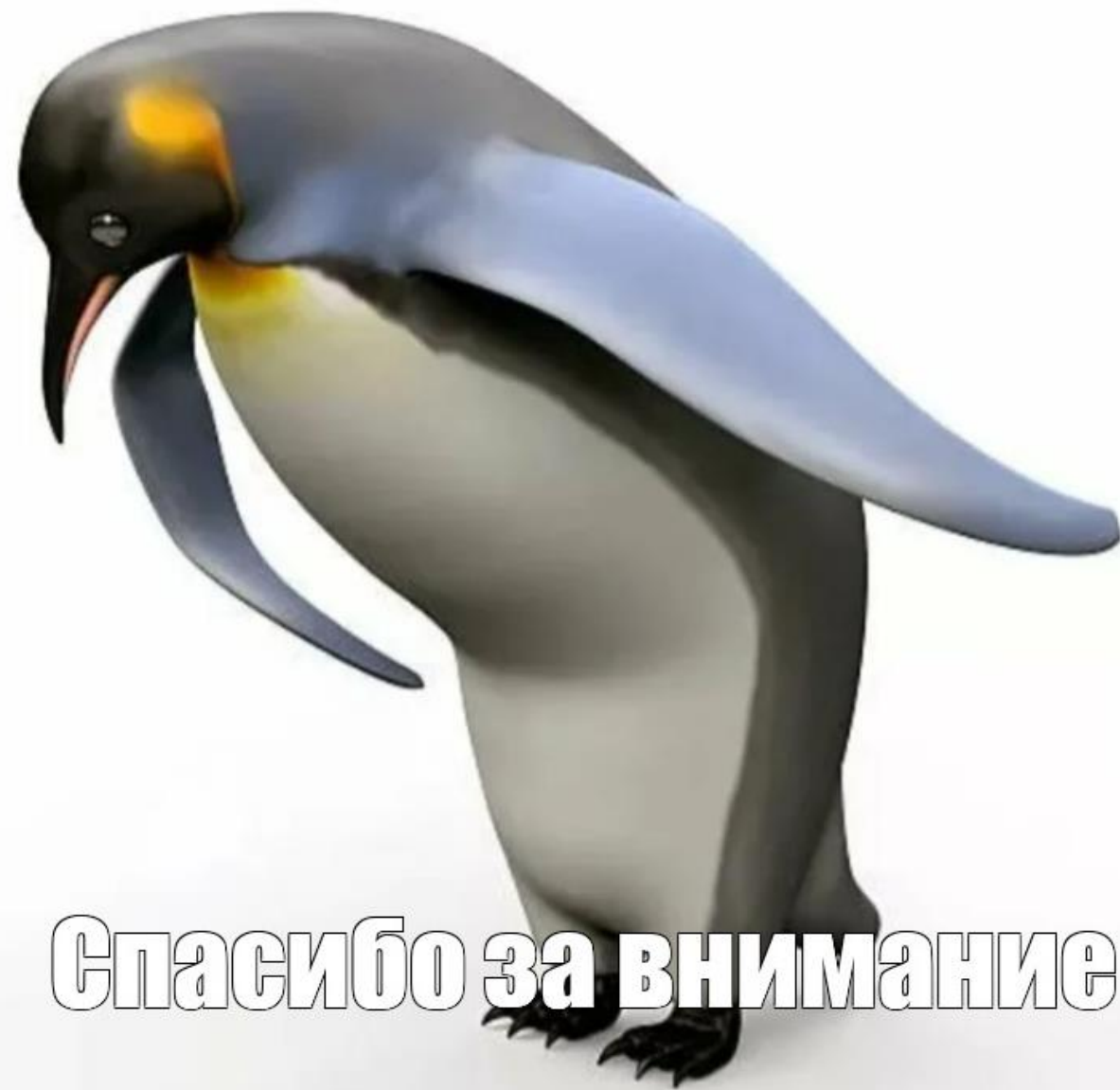
где Md – средняя арифметическая разностей показателей, измеренных до и после, σd – среднее квадратическое отклонение разностей показателей, n – число исследуемых

2.2. Непараметрический критерий U Уилкоксона – Манна – Уитни

Этот метод сравнения двух выборок признается наиболее чувствительным и мощным среди прочих непараметрических критериев. U -критерий Манна-Уитни представляет непараметрическую альтернативу t -критерию Стьюдента для независимых выборок и вычисляется по формуле:

$$U = n_x \cdot n_y + \frac{n(n + 1)}{2} - T$$

где n_x и n_y – объемы выборок; n – объем выборки, имеющей большую ранговую сумму; T – большая сумма рангов из выборок X и Y



Спасибо за внимание