Лекция 1

Основные понятия. Описательная статистика.

Статистические методы анализа в биологии и медицине

I семестр 2022/23 года

Преподаватели (каф. Зоологии и общей биологии ИФМиБ):

1-4 лекция – Александр Федорович Беспалов

5-8 лекция – Александр Николаевич Беляев

Регламент:
1. Работа на практических занятиях в семестре, контрольные работы – до 50 баллов
2. Зачет в виде теста – до 50 баллов

Математическая статистика — это раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей.

Статистические данные - данные, полученные в результате обследования большого числа объектов или явлений (то есть, математическая статистика имеет дело с массовыми явлениями).

Статистический анализ данных

Описательная статистика
descriptive statistics

Индуктивная статистика
inferential statistics

описательная статистика	**аналитическая статистика (теория статистических выводов)**
методы описания статистических данных, представления их в форме таблиц, распределений и пр.	обработка данных, полученных в ходе эксперимента, и формулировка выводов, имеющих прикладное значение для конкретной области человеческой деятельности. Теория статистических выводов тесно связана с другой математической наукой — **теорией вероятностей** и базируется на ее математическом аппарате

Предметом изучения в статистике являются **изменяющиеся (варьирующие) признаки**, которые иногда называют статистическими признаками.

Наличие общего признака является основой для образования статистической совокупности. Таким образом, **статистическая совокупность - э**то результат **описания или измерения общих признаков** объектов исследования.

Измерение – это приписывание числовых форм объектам или событиям в соответствии с определенными правилами.

Признаки и переменные – это измеряемые явления. Значения признака определяются при помощи специальных шкал наблюдения.

В ***1946*** году ***Стенли Стивенсом*** предложена классификация из 4 типов шкал измерения:

1) **номинативная, или номинальная, или шкала наименований**;

2) **порядковая, или ординальная, шкала**;

3) **интервальная, или шкала равных интервалов**;

4) **шкала равных отношений**.

Стэнли Смит Стивенс
(*Stanley Smith Stevens*; 1906-1973)

Данные – результаты некоторого количества измерений какой-либо ПЕРЕМЕННОЙ (признака) или Переменных (признаков) – variable. Например: вес, длина тела, пол, окрас, температура .....

ДАННЫЕ

Категориальные (качественные)

Номинальные
nominal

Категории взаимоисключающие (альтернативные) и неупорядоченные (их нельзя выстроить в последователь-ность)

Порядковые
ordinal

Категории взаимоисключающие (альтернативные) и упорядоченные (могут быть упорядочены; размер интервалов на шкале неодинаковый)

Количественные (числовые)

шкала отношений
ratio scale

интервальная шкала
interval scale

Дискретные
discrete

Целочисленные значения, типичные для счета

Непрерывные
continuous

Любые значения в определенном интервале

Потеря информации и точности

шкала отношений (ratio scale):

•размер интервалов на протяжении всей шкалы одинаковый;

•существует реальное нулевое значение.

Примеры: масса тела, размер выводка, объём

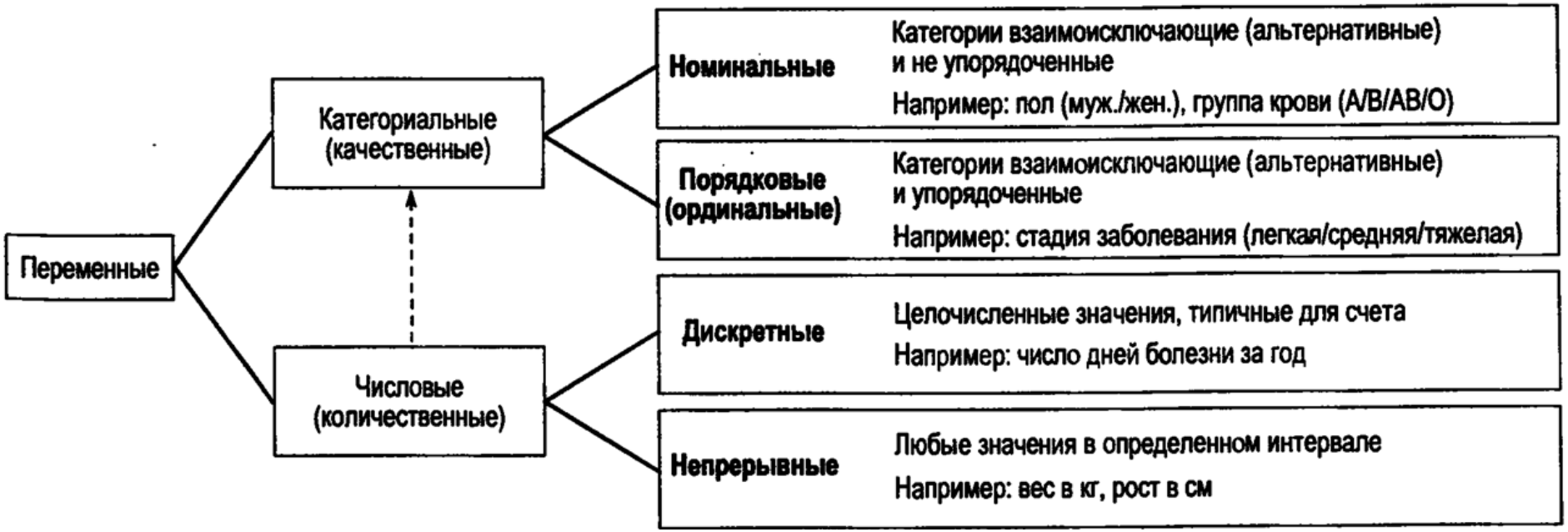
интервальная шкала (interval scale):

•размер интервалов на протяжении всей шкалы одинаковый;

•положение нулевой точки выбрано произвольно.

Примеры: температура по Цельсию, время дня, дата

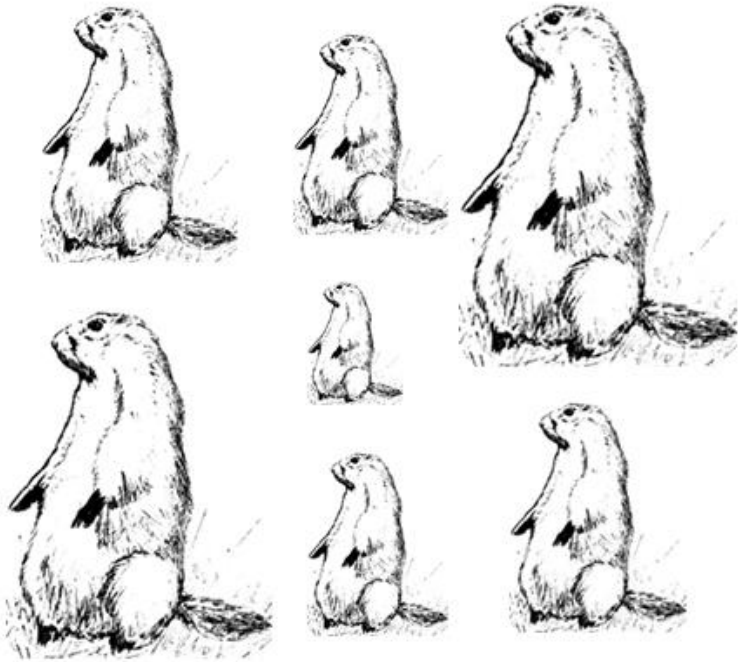
Тип данных¶


Переменные
Категориальные (качественные)
Числовые (количественные)
Номинальные
Категории взаимоисключающие (альтернативные) и не упорядоченные
Например: пол (муж./жен.), группа крови (A/B/AB/O)
Порядковые (ординальные)
Категории взаимоисключающие (альтернативные) и упорядоченные
Например: стадия заболевания (легкая/средняя/тяжелая)
Дискретные
Целочисленные значения, типичные для счета
Например: число дней болезни за год
Непрерывные
Любые значения в определенном интервале
Например: вес в кг, рост в см

Например, мы изучаем популяцию сурков

наблюдение

популяция – совокупность всех интересующих нас объектов



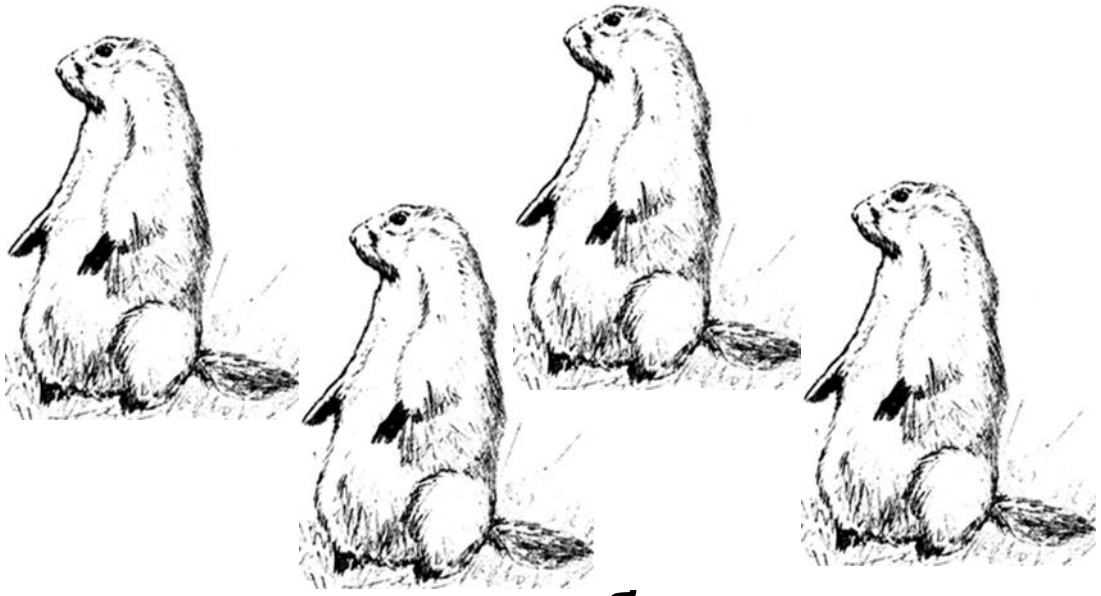
Описательная статистика: **ОПИСЫВАЕМ ВЫБОРКУ**

Индуктивная статистика: на основе свойств выборки (параметров выборки) делаем заключения о **СВОЙСТВАХ ПОПУЛЯЦИИ.**

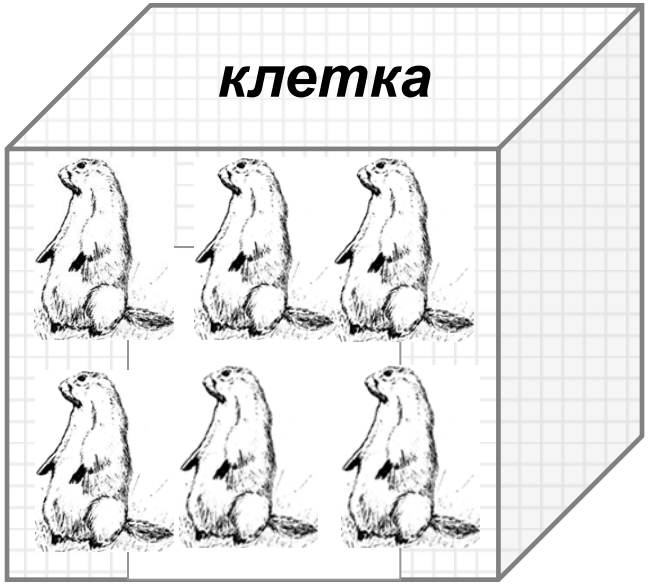
Выборка представляет собой случайным образом отобранную **часть популяции (генеральной совокупности)**

Выборка должна быть ***репрезентативной*** - её свойства должны отражать свойства популяции.

Для этого она должна быть ***случайной*** (random) – все особи в популяции должны иметь одинаковые шансы попасть в неё, и попадание в выборку одного элемента не должно влиять на попадание другого элемента.



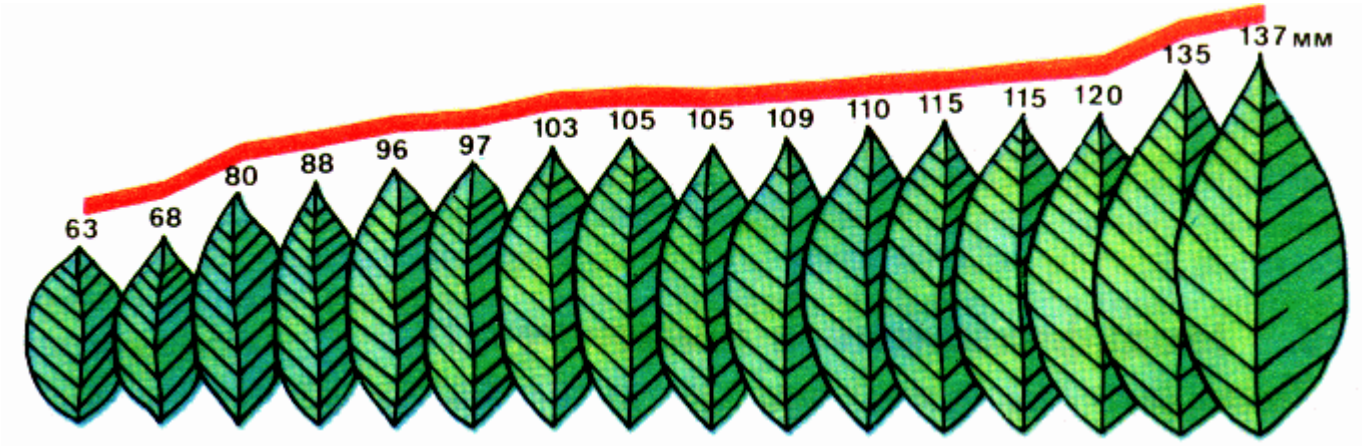
выборка



Пример: если в одну группу поместить зверьков, которые первыми вышли из клетки, а в другую – тех, кто в ней остался, выборки будут неслучайными

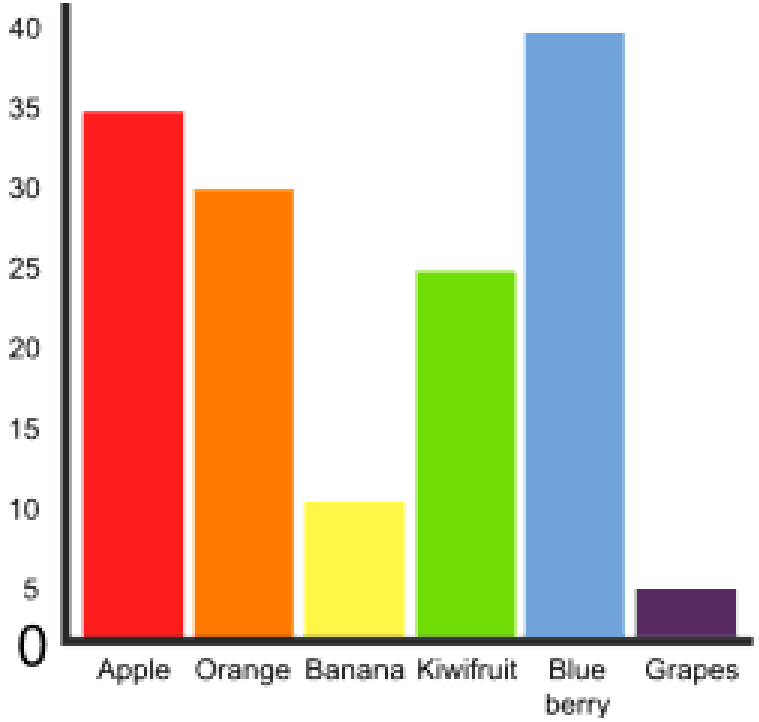
Три основные концепции в анализе данных:

1. Что такое РАСПРЕДЕЛЕНИЕ переменной и как его описывать
2. Что такое распределение ВЫБОРОЧНЫХ СРЕДНИХ и как оно связано с распределением переменной
3. Что такое СТАТИСТИКА КРИТЕРИЯ



Частотное распределение переменной (frequency distribution) – это соответствие между значениями переменной и их вероятностями (на практике – количеством таких значений в выборке)

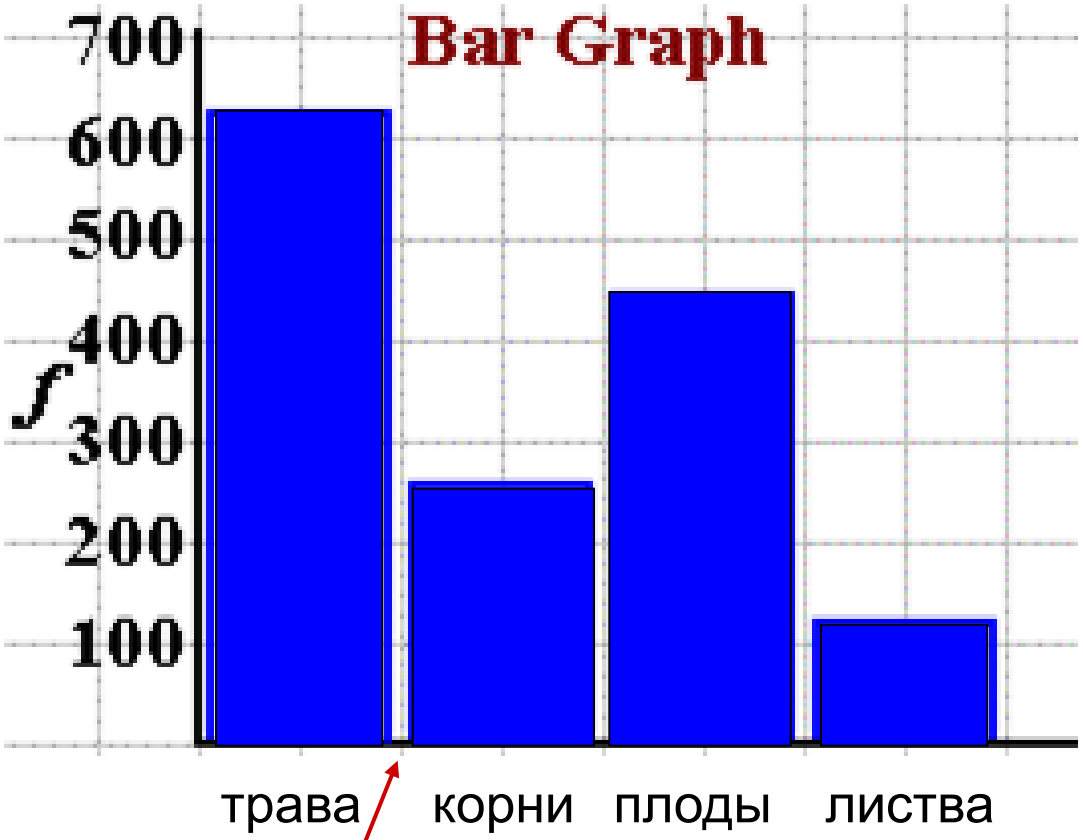
Можно представить в виде таблицы или графика



Сгруппировав данные по классам 145-149, 150-154, …,180-184,представить частотное распределение учащихся по этим группам с помощью :1) таблицы;

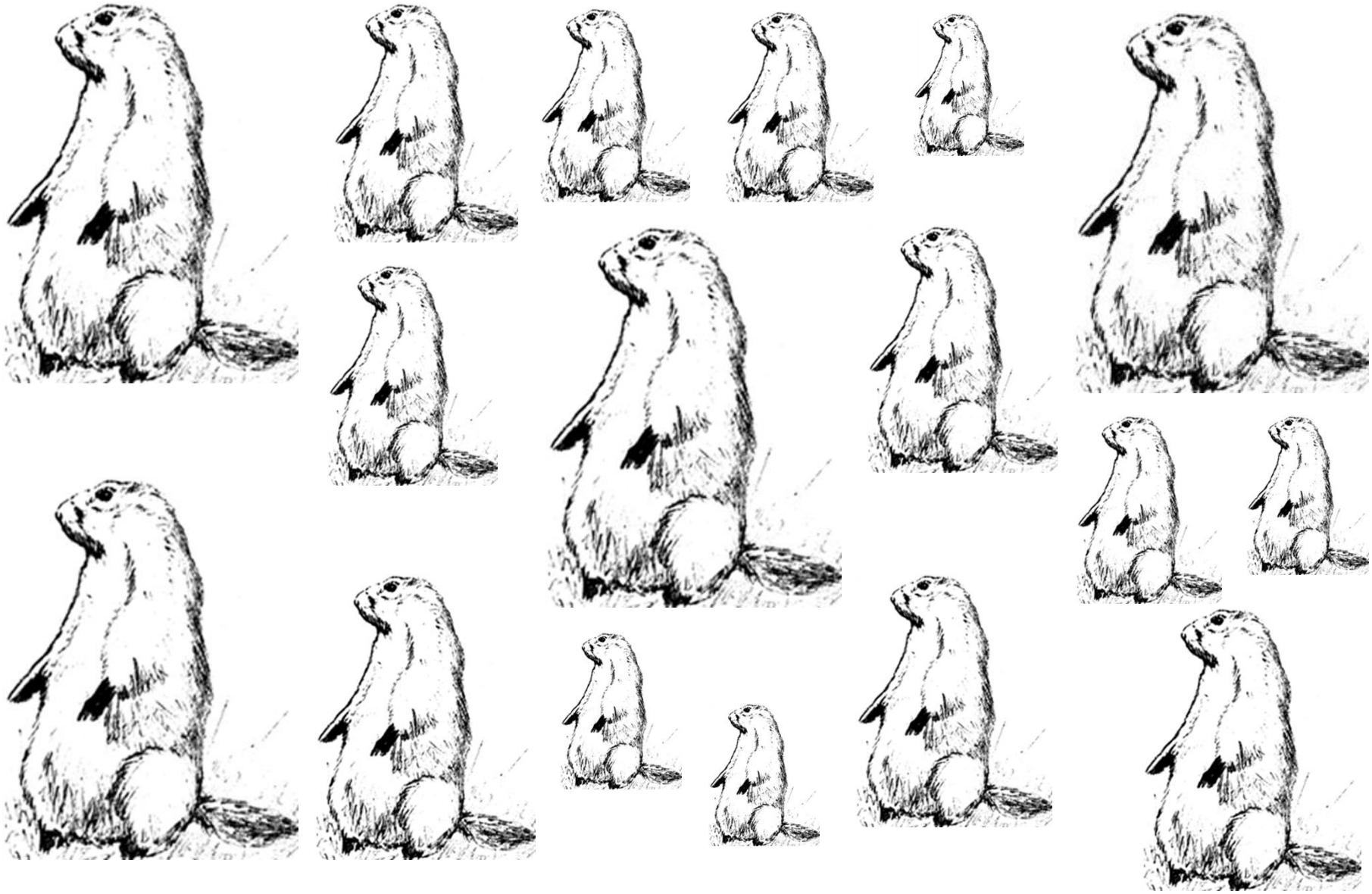
№ группы	1	2	3	4	5	6	7	8
Рост (см)	145-149	150-154	155-159	160-164	165-169	170-174	175-179	180-184
Кол-во человек	1	9	14	8	7	6	3	2

Картинка распределения **качественных** переменных (bar graph). В русском языке обозначается словом «гистограмма» (не совсем верно).



промежутки между столбиками

Оставим на некоторое время качественные переменные и обратимся только к **КОЛИЧЕСТВЕННЫМ**



Взвешиваем ***N*** сурков

1. Упорядочим по возрастанию значения переменной (выстроим сурков от меньшего к большему);

2. разобьём их на группы по равным интервалам;

3. Получаем вариационный ряд

Вариационный ряд – ряд, в котором сопоставлены (по степени возрастания или убывания) варианты и соответствующие им частоты.

Вариантами *(x_i)* считаются отдельные значения признака, которые он принимает в вариационном ряду.

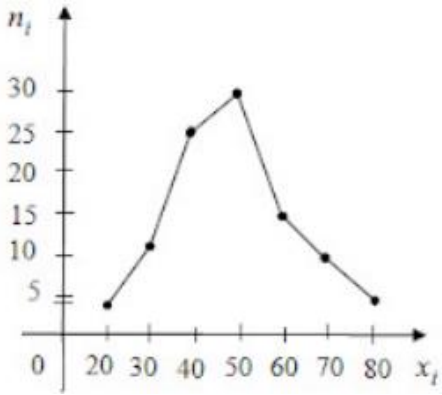
Частота *(f_i)* число, показывающее, сколько раз повторяется варианта.

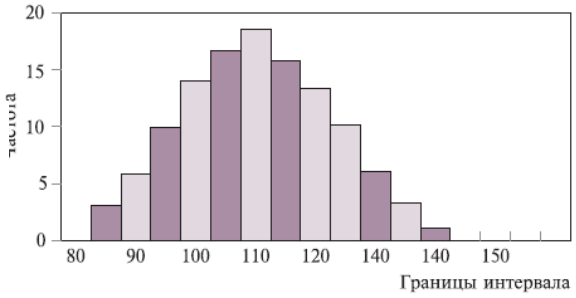
Частостью или относительной частотой *(w_i)* – называется отношение частоты к объему выборки (n).

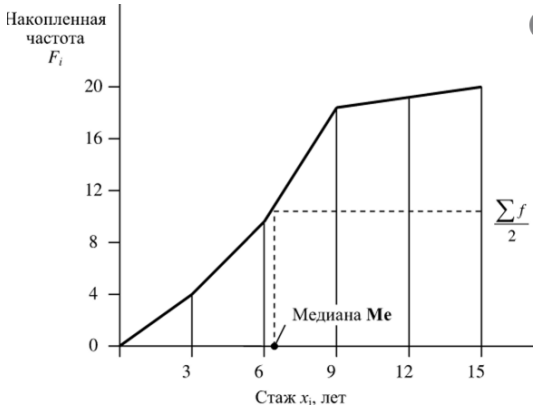
$$w_i = \frac{f_i}{n}$$

Наиболее употребительными графиками для изображения вариационных рядов, т. е. соотношений между значениями признака и соответствующими частотами или относительными частотами, являются:

- ***гистограмма***
- ***полигон***
- ***кумулята***







Гистограмма – графическое представление частотного распределения, разбитого по интервалам, где высота столбика отражает **ЧАСТОТУ**

Частота – то, сколько раз встретилось данное значение переменной

Термин «гистограмма» придуман английским математиком и статистиком ***Карлом Пирсоном***, одним из основателей современной статистики

Впервые использовал в одной из своих лекций в ***1891*** году.

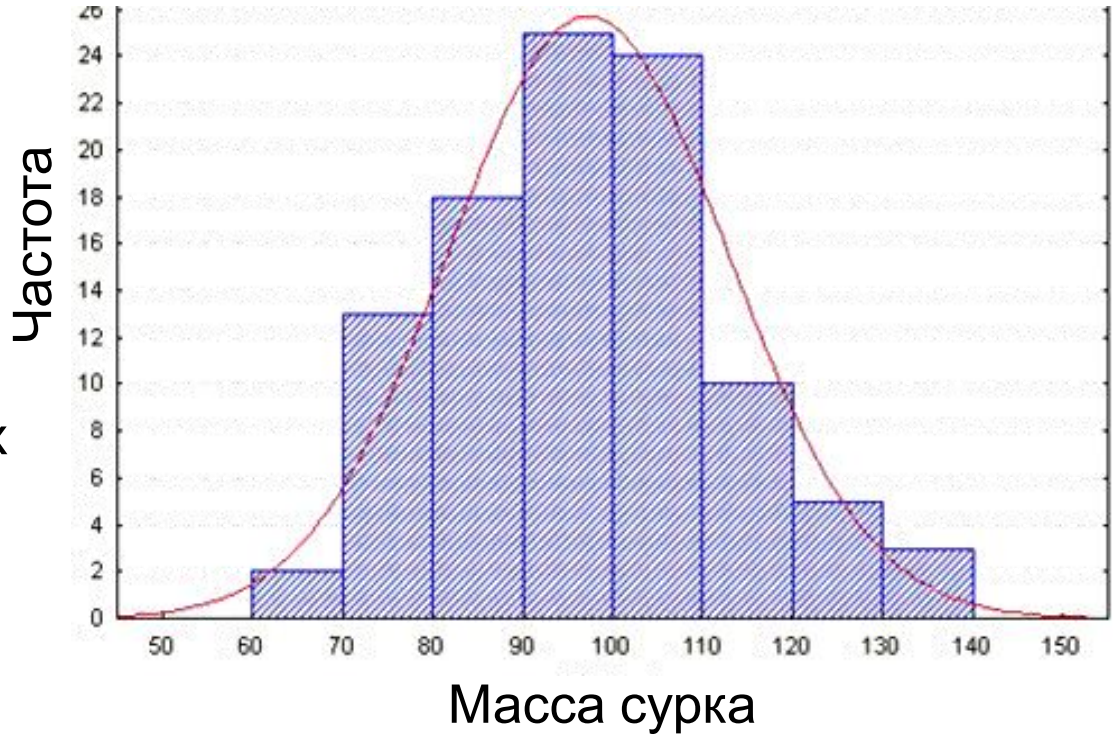
Карл Пирсон
(*Karl (Carl) Pearson 1857-1936*)

Интервалы должны быть:

•одного размера,

•не должны иметь общих точек,

•для биологических данных – 10-20 интервалов



Определение числа групп

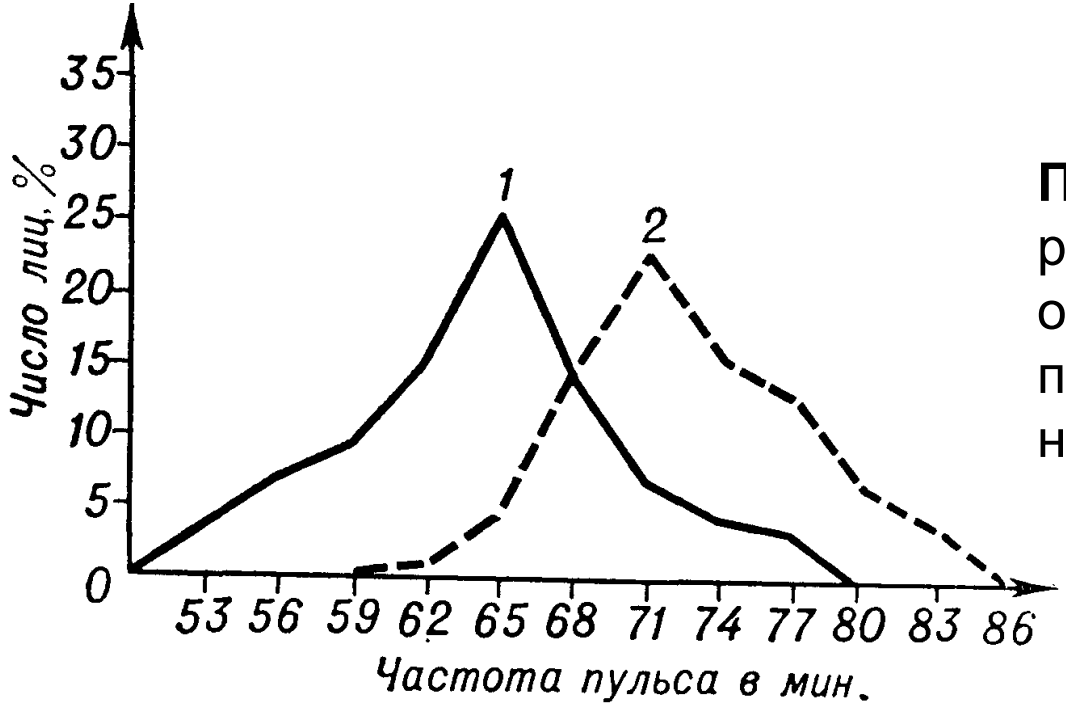
Формула Стерджесса

$$n = 1 + 3{,}322 \log N$$

где n – число групп

N – число единиц совокупности

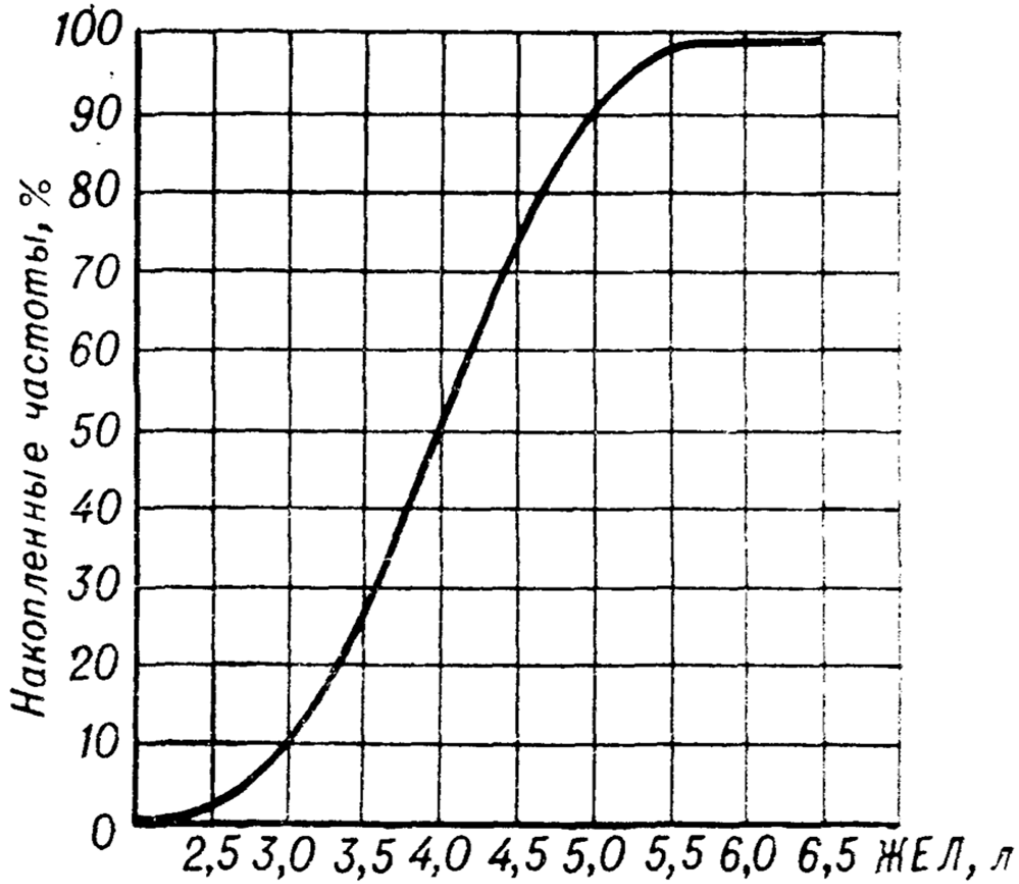
Полигон используют для дискретных рядов. На оси абсцисс в произвольном масштабе откладывают значения аргумента, а на оси ординат - значения частот или относительных частот.



Полигон (кривая распределения) обследованных по частоте пульса: 1 — до физической нагрузки, 2 — после нагрузки.

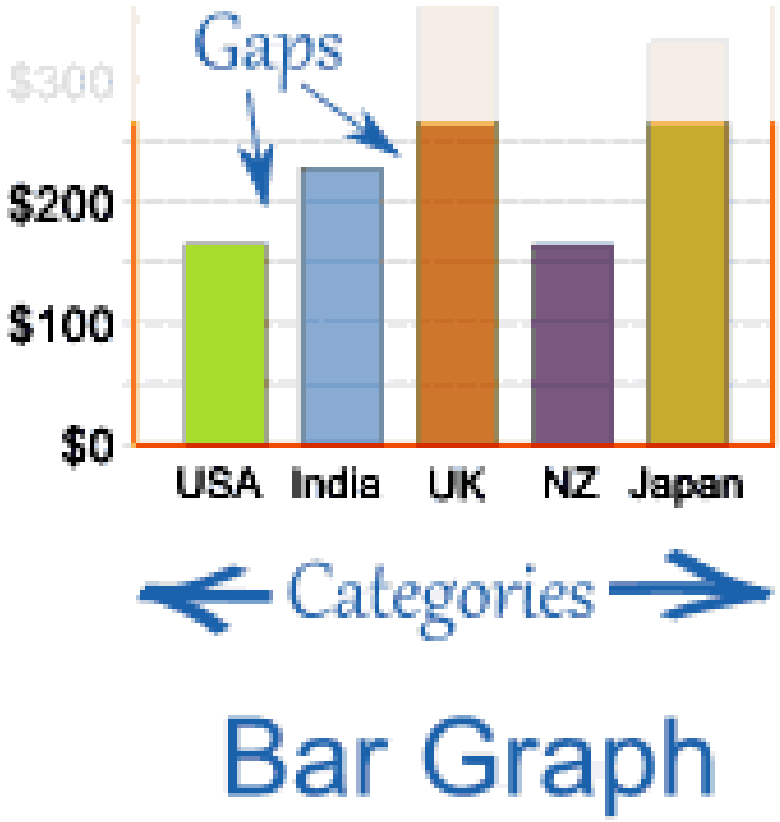
- Строят точки, координатами которых являются пары соответствующих чисел из вариационного ряда. Полученные точки последовательно соединяют отрезками прямой.
- Крайнюю левую точку соединяют с точкой оси абсцисс, абсцисса которой находится слева от рассматриваемой точки на таком же расстоянии, как абсцисса ближайшей справа точки.

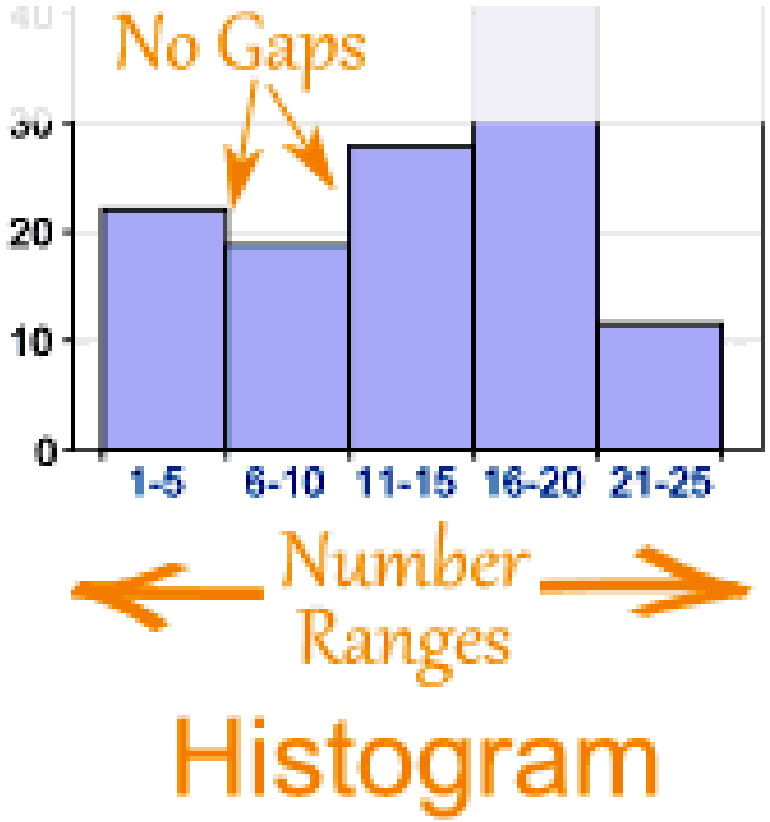
Кумулята служит для графического изображения кумулятивного (с накоплением) вариационного ряда.



- На оси абсцисс откладывают значения аргумента, а на оси ординат - накопленные частоты.
- Строят точки, абсциссы которых равны вариантам (в случае дискретных рядов) или верхним границам интервалов (в случае интервальных рядов), а ординаты - соответствующим частотам (накопленным частотам). Эти точки соединяют отрезками прямой.

Отличие графического представления распределения качественных и количественных признаков





Как описать частотное распределение переменной?

Три **ОСНОВНЫЕ ХАРАКТЕРИСТИКИ**, которыми можно почти полностью описать большинство **распределений**

1. **Меры положения или центральной тенденции**
2. **Меры вариации или изменчивости**
3. **Меры формы**

Наша цель – оценить **параметры популяции** *(population parameters)* с помощью **показателей из выборки** *(sample statistics)*

Речь идёт не только о количественных данных, но и о качественных

Меры положения или центральной тенденции распределения

Количественные данные*

«Середина»

Среднее значение
(mean)

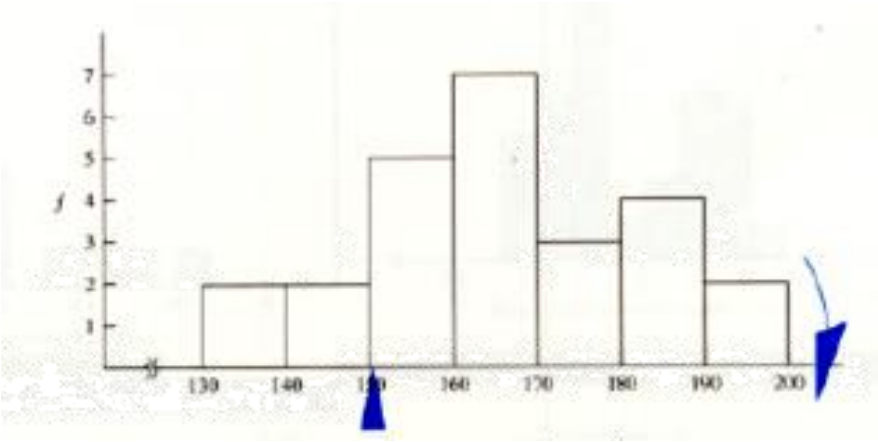
Медиана
(median)

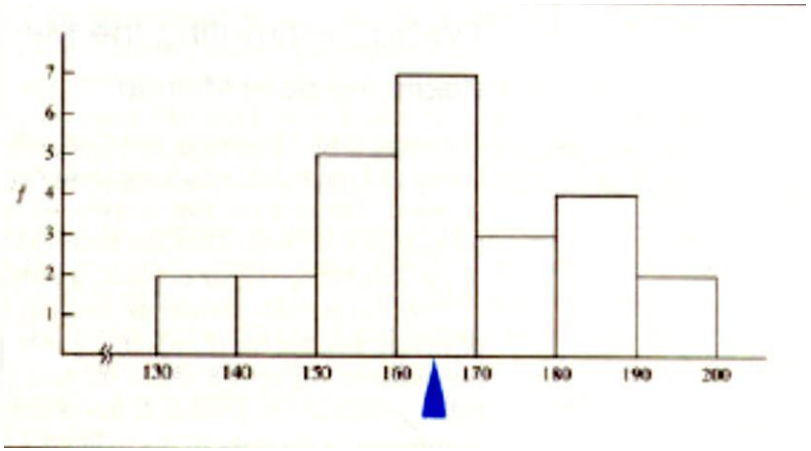
Мода
(mode)

Все они могут служить оценками популяционного среднего. Среднее в выборке – наиболее эффективная и несмещённая оценка.

* - описательная статистика для качественных данных рассматривается в лекции, посвященной анализу качественных данных

Среднее значение – сумма всех значений переменной, делённая на количество значений





*«balancing point» method

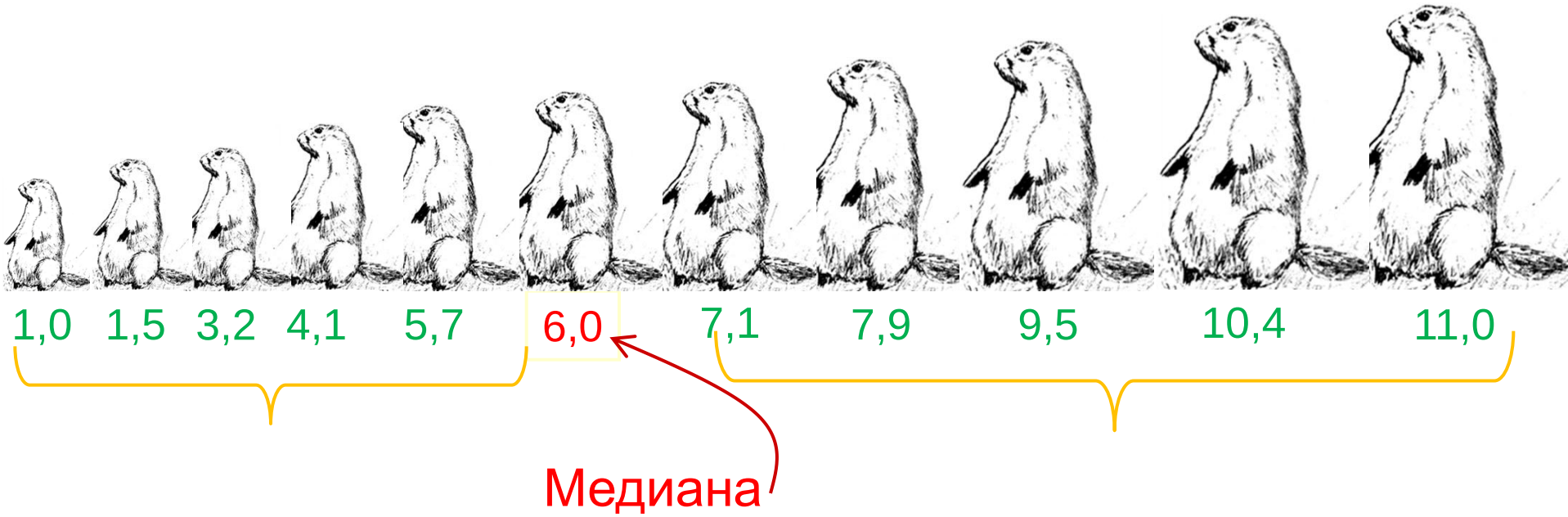
Среднее для **выборки**

$$\overline{X} = \frac{\sum_i X_i}{n}$$

Среднее для **популяции**

$$\mu = \frac{\Sigma X}{N}$$

Медиана (median) – значение, которое делит распределение пополам (его площадь в т.ч.): половина значений больше медианы, половина – не больше.



Имеет смысл не только для количественных переменных, но и для порядковых! (не для номинальных).

- ✓Если распределение не симметричное, медиана лучше характеризует центр распределения.
- ✓она содержит меньше информации, чем среднее (определяется только рангом измерений, а не их значениями)
- ✓но зато она не чувствительна к «аутлаерам» и может применяться даже в случае, если не для всех особей измерения точные.

Распределение можно поделить не только на ДВЕ равные части, но и на:

- ✓четыре (значения, стоящие на границах - квартили);
- ✓восемь (... октили);
- ✓сто (... процентили);
- ✓N (... квантили).

Квартили (quartiles) делят распределение на четыре части так, что в каждой из них оказывается поровну значений (2-я квартиль = медиана).

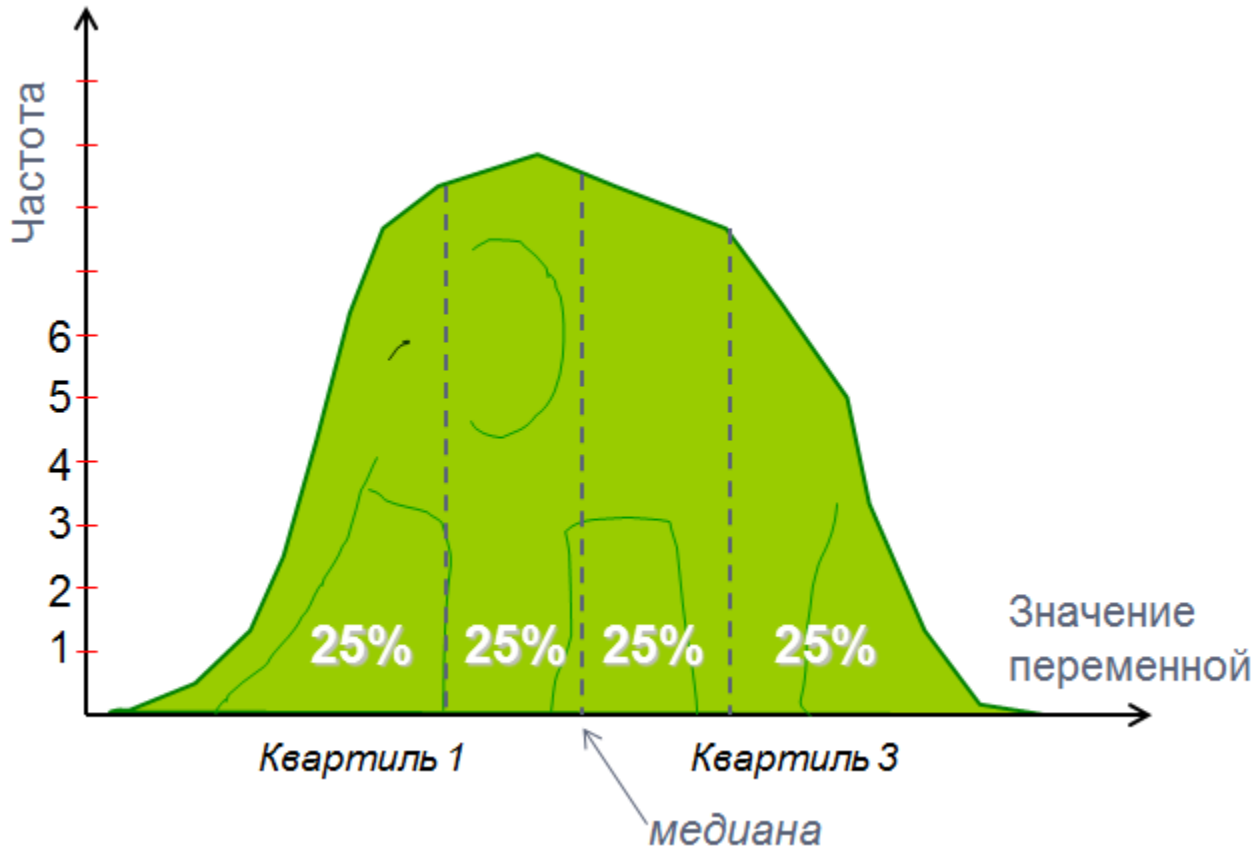
1-я квартиль = 25% процентиль

3-я квартиль = 75% процентиль

Межквартильный размах

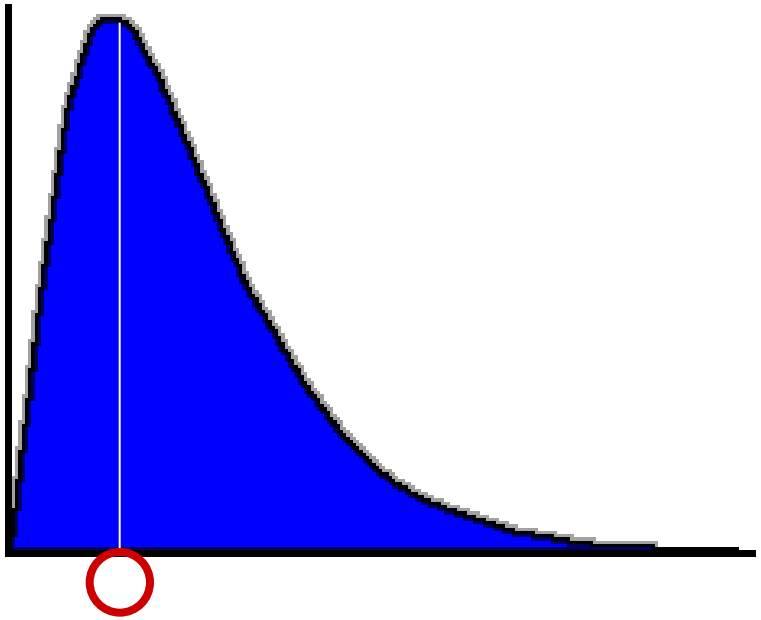
– разница между третьей

и первой квартилями.



Мода (mode) – наиболее часто встречающееся значение

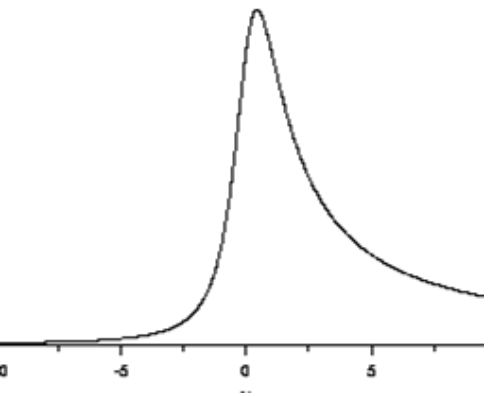
Существует не только для количественных, но и для ранговых, и для качественных переменных



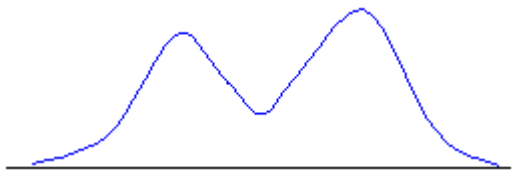
В первую очередь биолога интересует количество мод в распределении, а не мода как таковая

По количеству «максимумов» (мод) распределение:

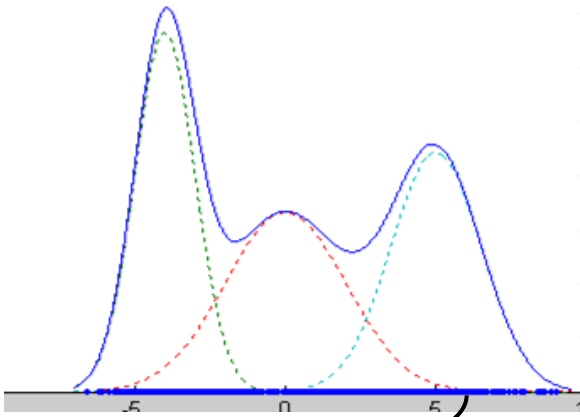
унимодальное



бимодальное

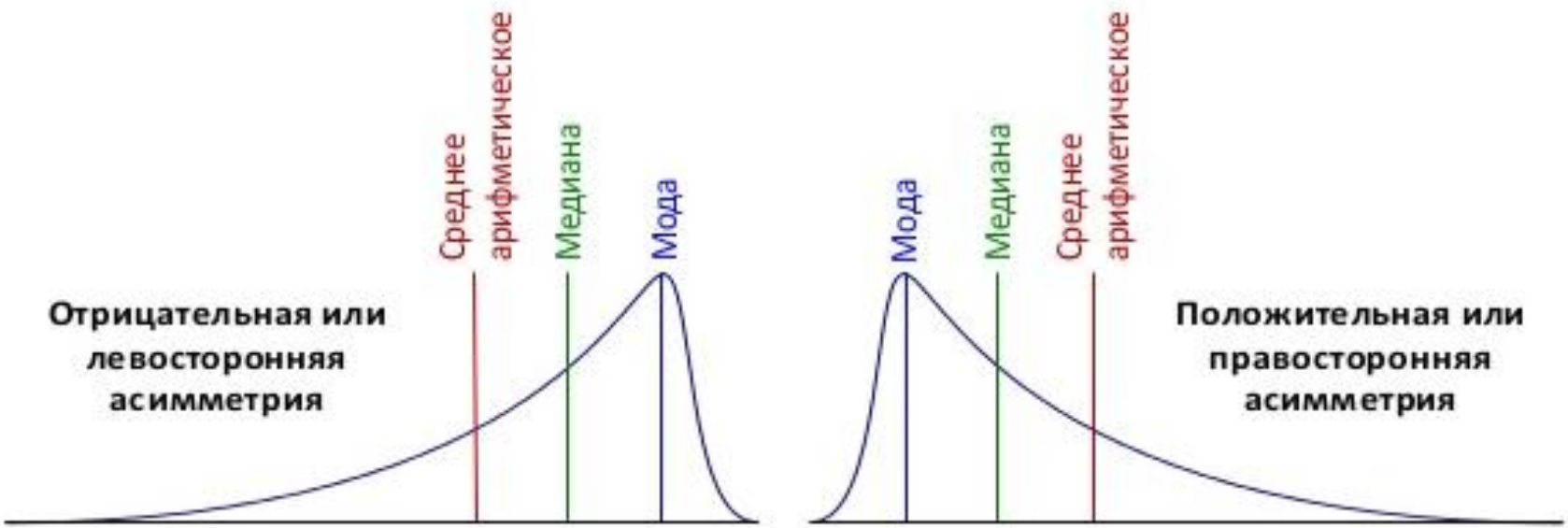


мультимодальное



обычно возникают, если популяция имеет естественные обособленные подгруппы

Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения



К появлению перекоса чувствительнее всего среднее значение

*«**Середина**» распределения*

Тип средней	Преимущества	Недостатки
Среднее	▪ Используются все значения набора данных ▪ Определяется математически выполнимым алгебраическим выражением	▪ Искажено выбросами* ▪ Искажается асимметричными данными
Медиана	▪ Не искажается выбросами ▪ Не искажается асимметричными данными	▪ Игнорирует большую часть информации ▪ Не определяется алгебраически
Мода	▪ Легко определяется для категориальных данных	• Игнорирует большую часть информации • Не определяется алгебраически

*** - выброс (аномальное значение) - наблюдение, которое отличается от основной части данных и несовместимо с остальными данными**

Частотное распределение переменной (frequency distribution)

«**Ширина**» распределения = **Разброс***

Размах (range)

Стандартное отклонение (standard deviation)

Дисперсия (variance)

Размах (range) – разность между максимальным и минимальным значениями $= X_n - X_1$

Хорош тем, что легко считается и имеет «биологический смысл».

Плох тем, что зависит лишь от 2-х точек из распределения. Недооценивает истинный разброс в популяции. Если в статье приводится размах, следует привести ещё какую-нибудь характеристику разброса.

* Это лишь основные параметры разброса

Стандартное отклонение или средне квадратичное отклонение (standard deviation)

Для **выборки**:

$$s = \sqrt{\frac{\boxed{\sum_i (X_i - \overline{X})^2}}{n-1}}$$

Поправка на то, что в выборке разброс всегда будет меньше, чем во всей популяции

Сумма квадратов (sum of squares = SS)

Для популяции:

$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$$

Стандартное отклонение зависит ото всех значений переменной.

Измеряется в тех же единицах, что и переменная!

Дисперсия (variance)

Для **выборки**:

$$s^2 = \frac{\sum_i (X_i - \overline{X})^2}{n-1}$$

Для популяции:

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$$

Равна стандартному отклонению в квадрате и содержит почти ту же информацию;
измеряется в единицах переменной, <u>возведённых в квадрат</u> (что не всегда удобно).

Дисперсия используется скорее в различных статистических тестах, а не в описательной статистике

Коэффициент вариации *
(Coefficient of variation)

$$CV = \frac{s \cdot 100}{\overline{X}}$$

Это отношение стандартного отклонения к средней арифметической для выборки, которое выражено в процентах

Даёт понять, насколько на самом деле велик разброс в данных, независимо от масштаба измерений.

Не годится для данных, измеренных по интервальной шкале (температура, время и пр.)

Чем больше значение коэффициента вариации, тем выше изменчивость (вариабельность) признака в выборке

Обычно используют 3 пороговых значения:
CV – 5% - низкая изменчивость,
CV – 10% - средняя изменчивость,
CV – 15% - высокая изменчивость

* Мера относительного разброса случайной величины

Разброс распределения

Мера Рассеяния	Преимущества	Недостатки
Размах	▪ Легко определить	▪ Используют только два наблюдения
Дисперсия	▪ Использует каждое наблюдение ▪ Имеет алгебраическое выражение для вычисления	▪ Единица измерения квадрат исходных данных ▪ Восприимчива к выбросам ▪ Не подходит для асимметричных распределений данных
Стандартное отклонение	▪ Использует каждое наблюдение ▪ Имеет алгебраическое выражение для вычисления ▪ Единицы измерения те же, что и у исходных данных ▪ Легко объяснима	▪ Восприимчива к выбросам ▪ Не подходит для асимметричных распределений данных

Представление мер описательной статистики:

I. Количественные признаки с нормальным распределением

Стандарт 1. Среднее ± стандартная ошибка среднего*

В символьной форме: $\bar{x} \pm m$; Англ. mean ± standard error of mean.
В публикациях встречаются обозначения: M ± m, Mean ± SE, Mean ± SEM.
Например масса студентов: **50 ± 1,5**

Стандарт 2. Среднее; стандартное отклонение

В символьной форме: $\bar{x}$; s. Англ. mean; standard deviation.
В публикациях встречаются обозначения : Mean; SD, M; s.d.
Например масса студентов : **50; 5**

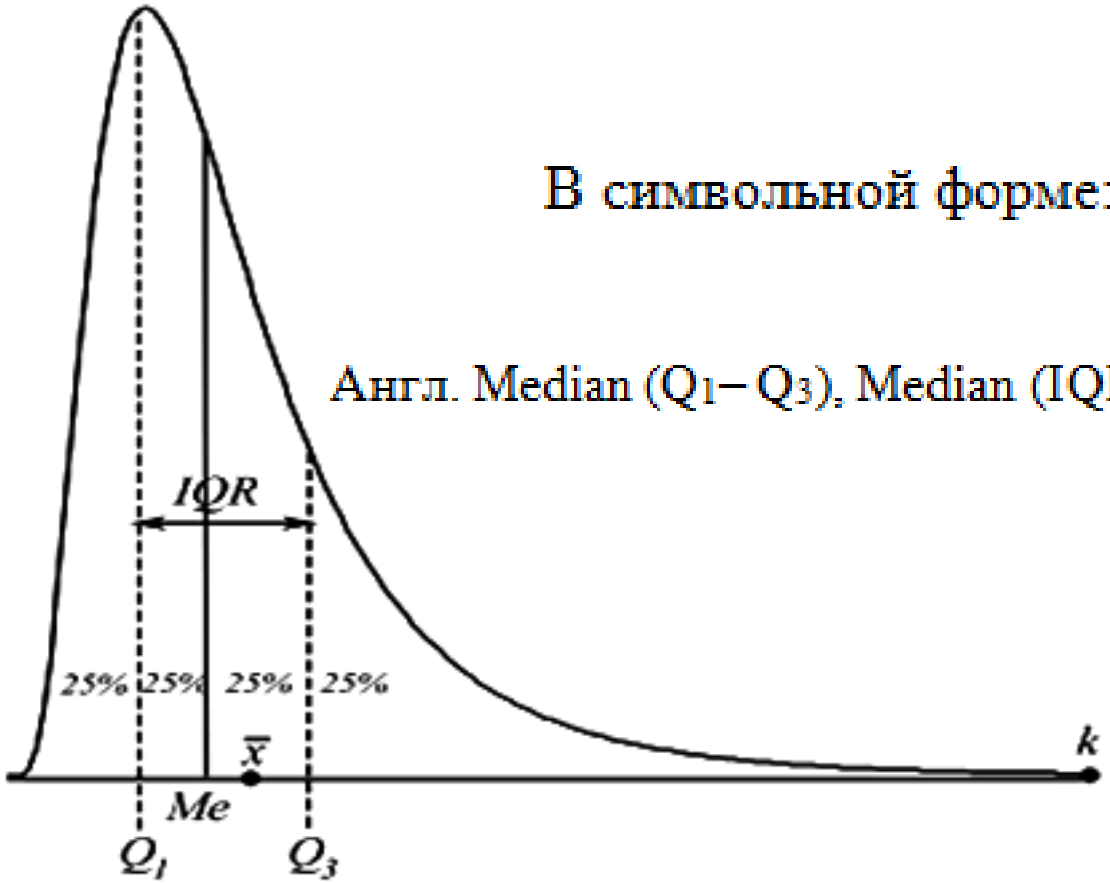
Стандарт 3. Среднее [Доверительный интервал*]

В символьной форме: $\bar{x}$ [ДИ; $\overline{\text{ДИ}}$], $\bar{x}$ (95% ДИ; от ДИ до $\overline{\text{ДИ}}$)
где ДИ – нижняя граница доверительного интервала; $\overline{\text{ДИ}}$ - верхняя граница интервала
Например масса студентов : **50 [30; 70]**

* - о стандартной ошибке среднего и доверительном интервале см. в лекции далее

II. Количественные признаки с ненормальным распределением и порядковые признаки

СТАНДАРТ. Медиана (межквартильный размах*)



В символьной форме: Me (Q_1– Q_3)

Англ. Median (Q_1– Q_3), Median (IQR = Q_1– Q_3), Median ($q_1$to q_3)

* ***Межквартильный размах*** - интервал между нижней квартилью (25-й процентиль) — верхней квартилью (75-й процентиль)

III. Качественные номинальные признаки *

Такие признаки нельзя упорядочить естественным образом.

Примеры: вид организма, сорт сельскохозяйственной культуры, диагноз пациента. Эти данные представляют собой частоты

Частоты: 1) абсолютные (в штуках) или 2) относительные (в долях единицы, в процентах, в промилле и др.).

СТАНДАРТ Частота [Доверительный интервал]

* - подробнее об описательной статистике качественных признаков смотрите в лекции анализ качественных данных

Форма *распределения*

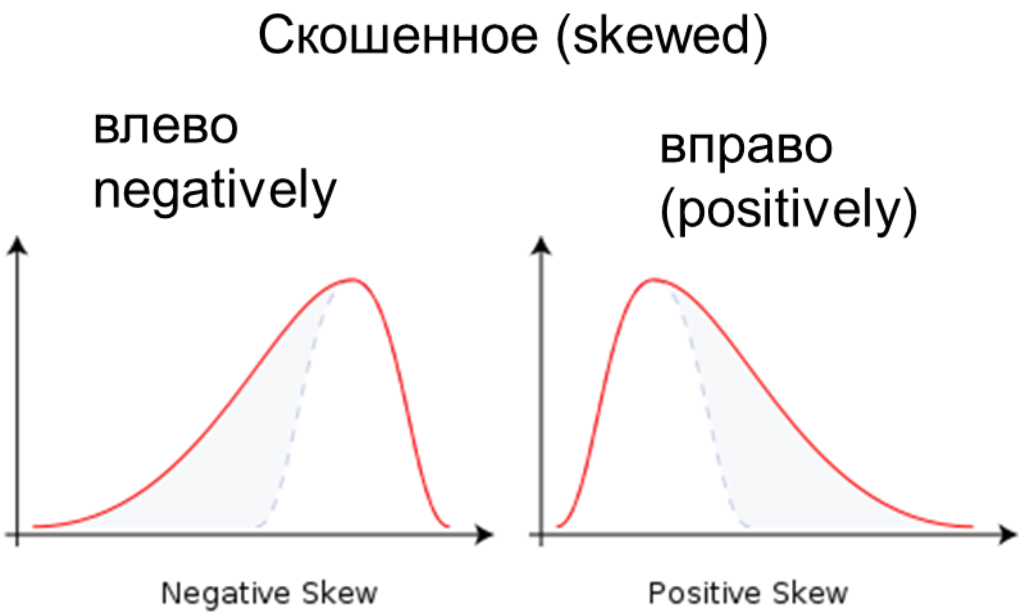
Асимметрия, скошенность (skewness)

Характеристики:

As=0 симметричное

As<0 левосторонее

As>0 правостороннее



As - коэффицие́нт асимметри́и

При ***положительной скошенности*** распределения *правый хвост*, как правило, *толще левого*, а вершина смещена влево. Такое распределение часто называют *скошенным вправо* (исходя *из соотношения толщины хвостов*, а не из положения вершины)

При ***отрицательной скошенности***, соответственно, *левый хвост*, как правило, *толще правого*, а вершина смещена вправо. Такое распределение называют *скошенным влево*

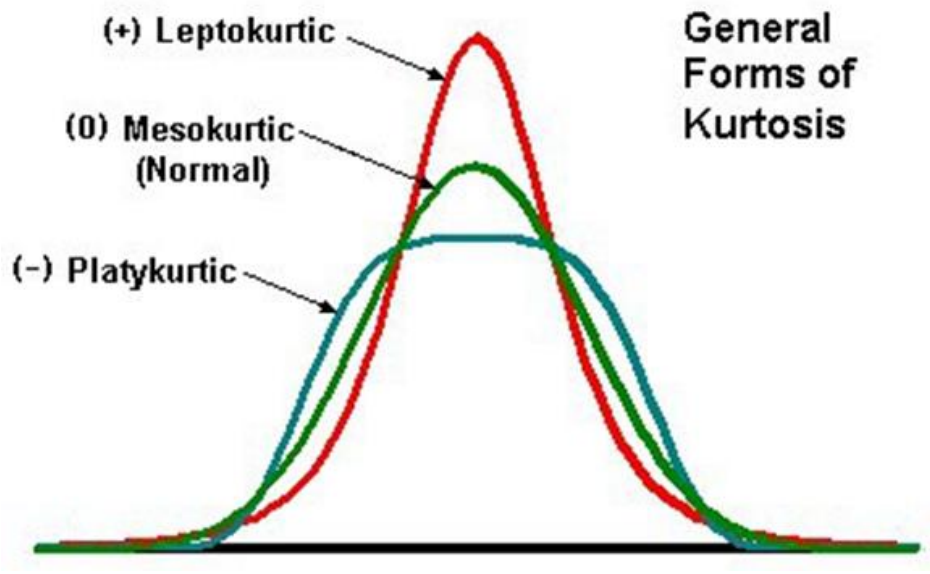
Эксцесс, куртозис (excess, kurtosis)

Характеристики:

Ex=0 нормальное

Ex<0 плосковершинное

Ex>0 островершинное



Куртозис (kurtosis) является показателем, отражающим остроту вершины и толщину хвостов одномерного распределения.

Leptokurtic distribution — это распределение с положительным эксцессом, имеющее острую вершину и толстые хвосты, а ***platykurtic distribution*** — это распределение с отрицательным эксцессом, имеющее плоскую вершину и тонкие хвосты.

Соответственно, ***mesokurtic distribution*** — это распределение, похожее по куртозису на нормальное распределение.

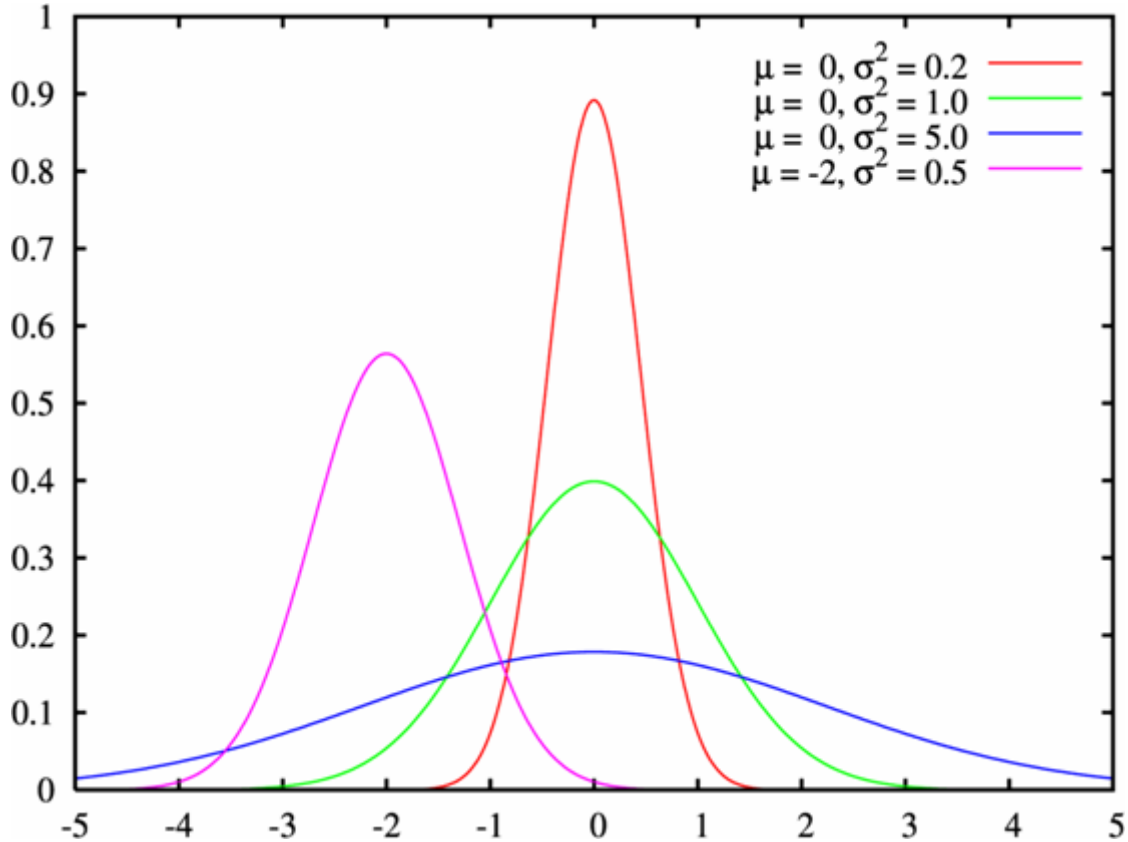
Нормальное распределение (Гауссово):

первое знакомство

✓Унимодальное
✓Симметричное
✓Асимптотическое

Это <u>непрерывное</u> распределение

$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

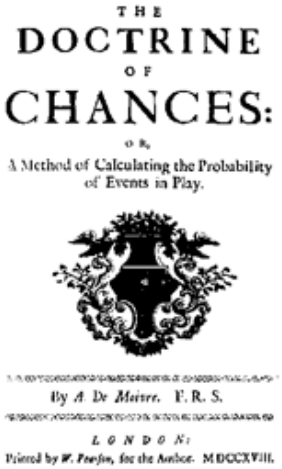


Собственно говоря – это целое семейство нормальных кривых. Форма зависит от значений: μ — математическое ожидание, σ^2 — дисперсия

Муавр Абрахам (Moivre Abraham)

"Учение о случаях" 1718 г.

(1667-1754)



THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability
of Events in Play.

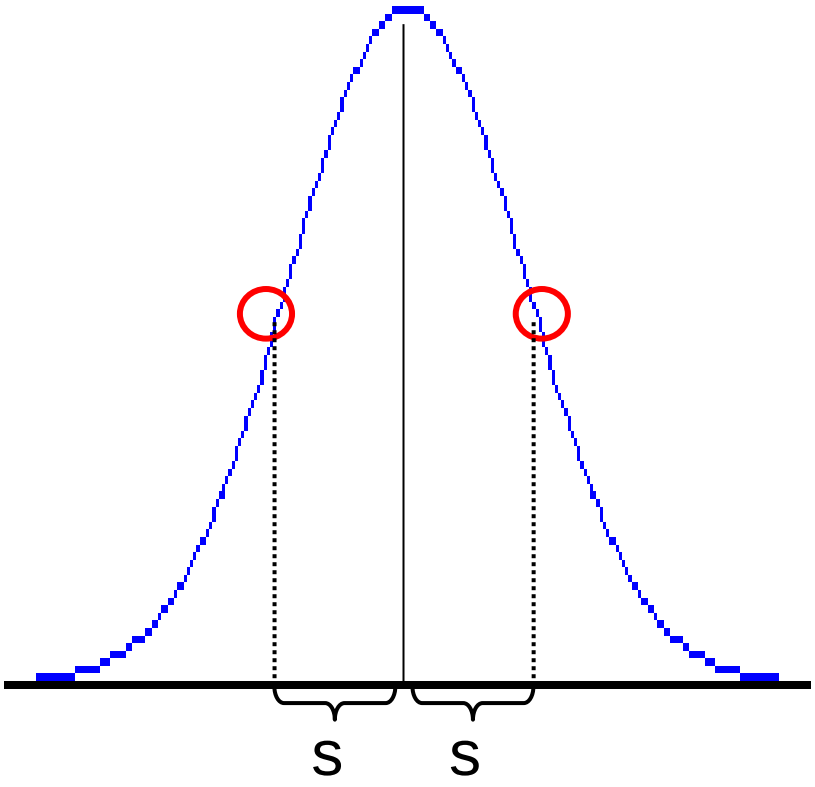
By A. De Moivre, F. R. S.

LONDON:
Printed by W. Pearson, for the Author. MDCCXVIII.

Название распределения в честь Гаусса, возможно, не совсем справедливо – первым его описал вовсе не он. Однако оно, в первую очередь, стало известно благодаря его работам

Впервые это распределение как предел биномиального распределения появилось в ***1738*** году во втором издании работы ***Муавра*** «Доктрина случайностей»

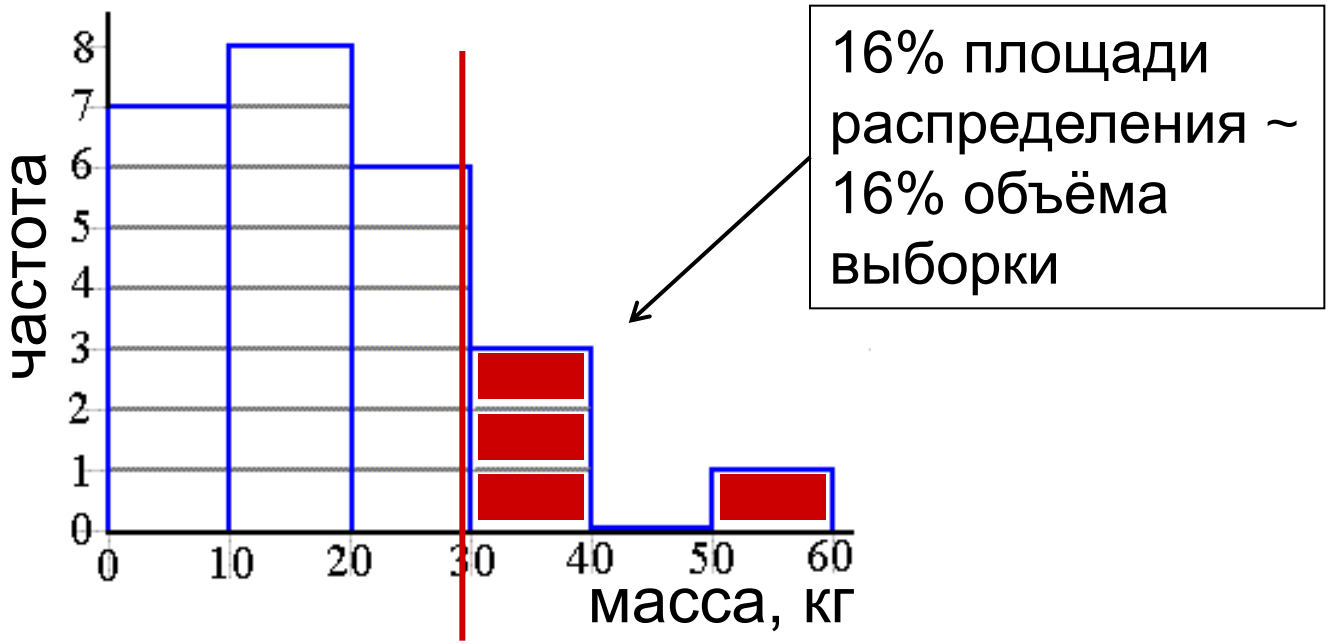
Стандартное отклонение (standard deviation):
для нормального распределения = дистанции от среднего значения до каждой из точек перегиба



«Площадь распределения»

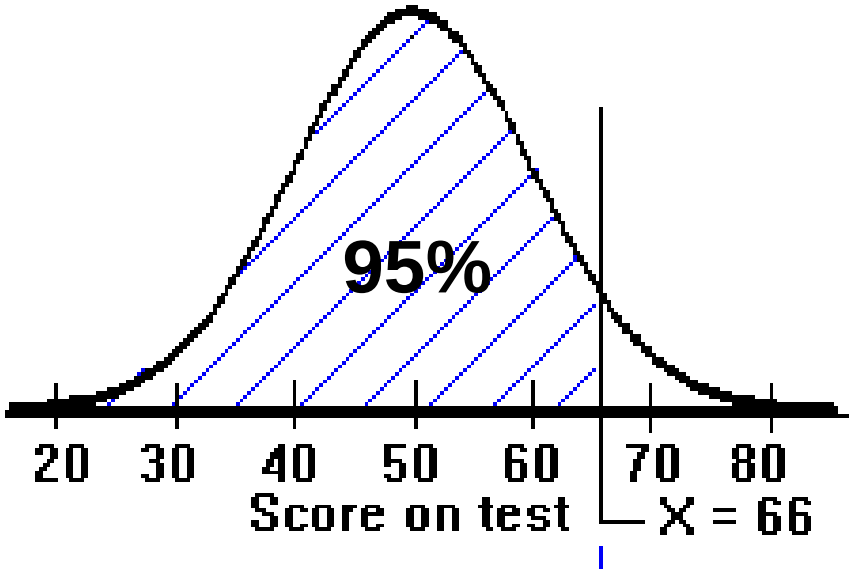
Площадь, которую занимает график распределения, соответствует <u>количеству измерений</u> в выборке.

Отрезая часть распределения на графике, мы отделяем эквивалентную часть от выборки



Процентили и z-оценка

95% процентиль – значение переменной, левее которого находится 95% значений переменной



Процентили и z-оценка

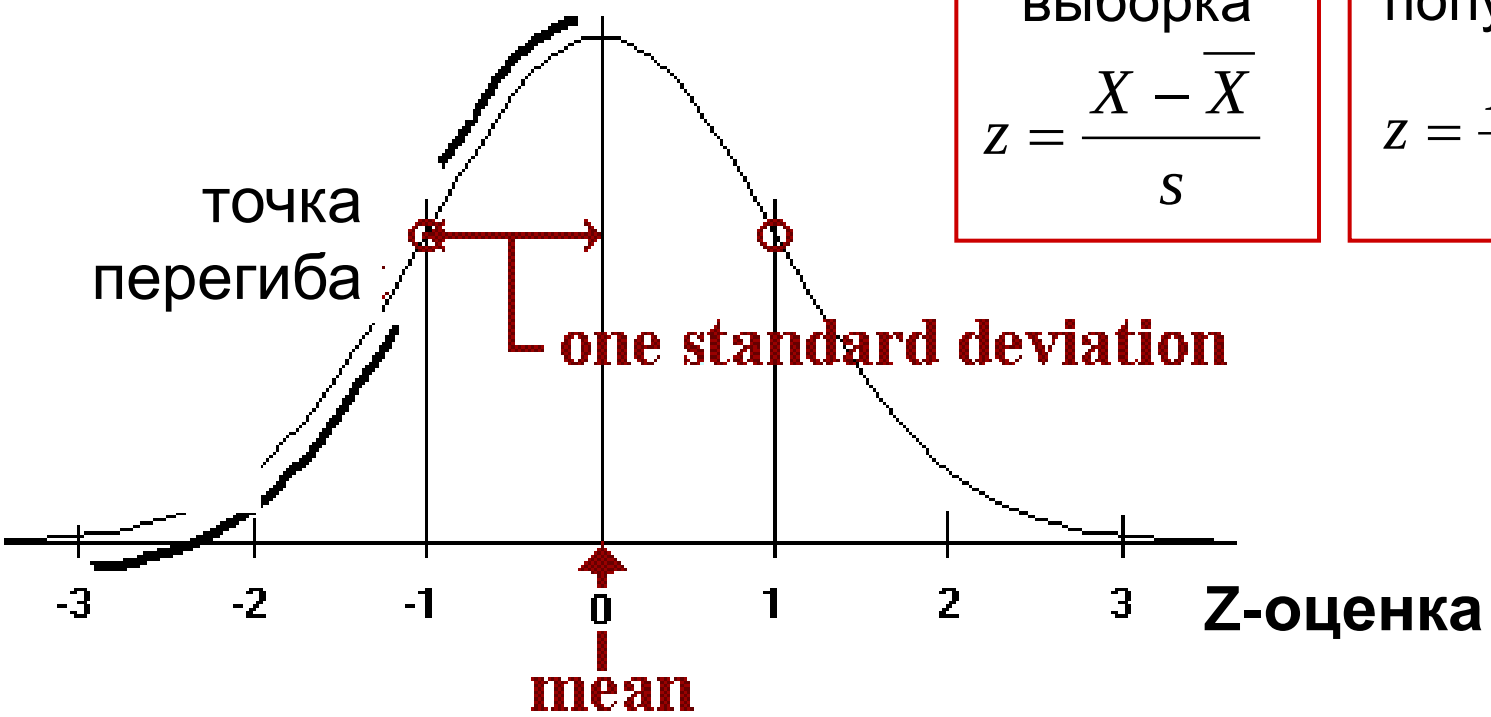
Z-оценка (z-scores) – переменная, соответствующая <u>количеству стандартных отклонений</u> относительно среднего значения

выборка

$$z = \frac{X - \overline{X}}{s}$$

популяция

$$z = \frac{X - \mu}{\sigma}$$



$z =$

Площадь нормального распределения

Нормальное распределение определяется лишь 2-мя параметрами – μ и σ.

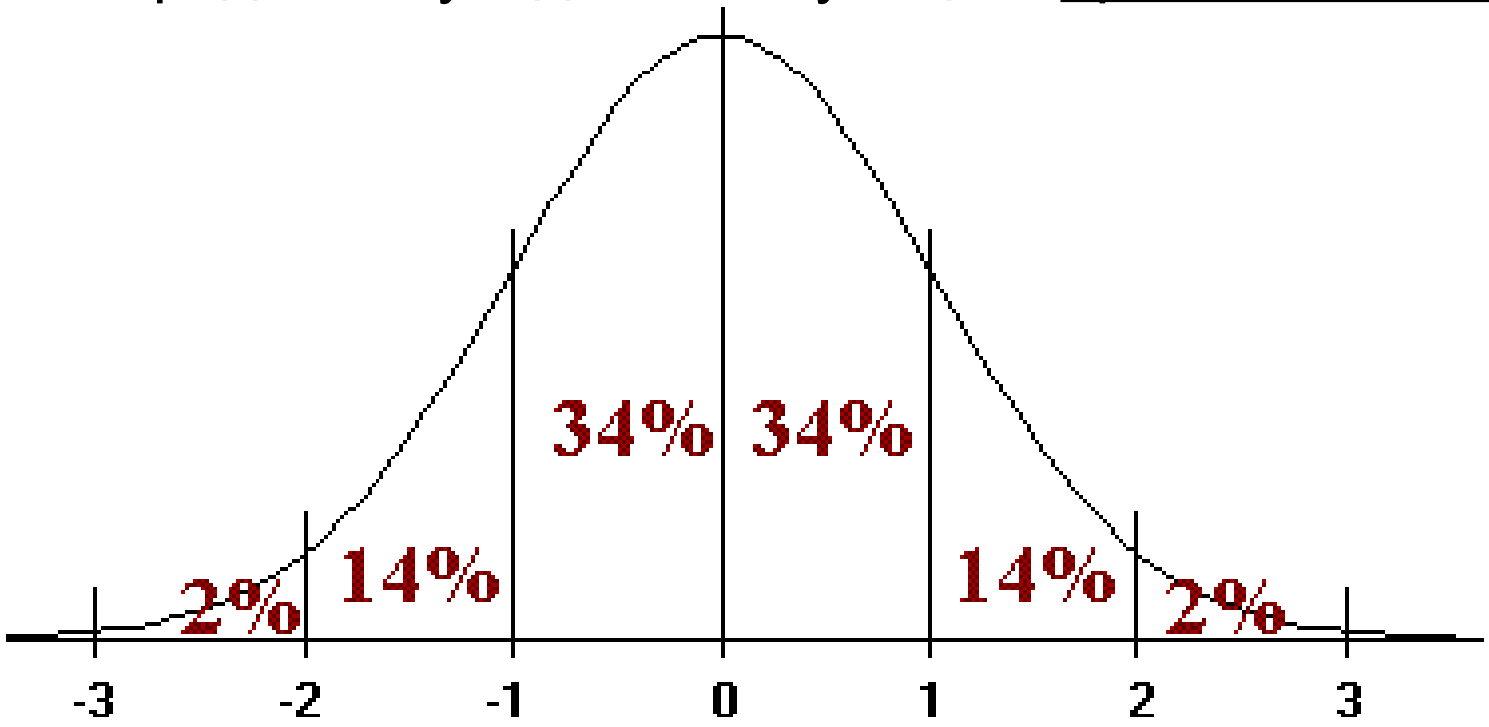
$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

<u>Необыкновенное свойство</u>:

Относительные площади под участками нормального распределения всегда одинаковы!

Площадь нормального распределения

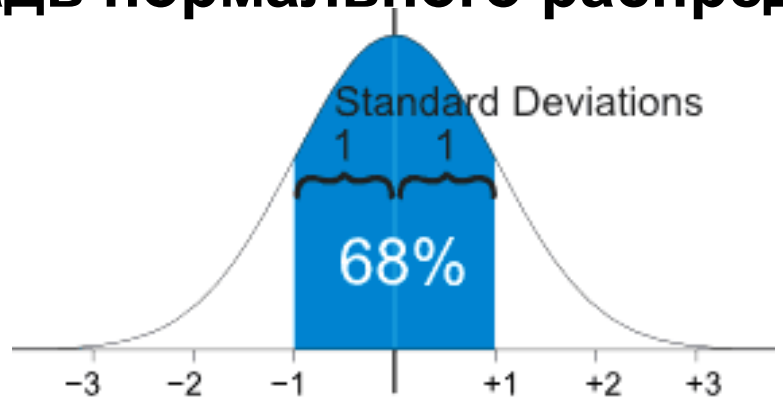
Откладывая от среднего значения стандартное отклонение (в ту или другую сторону) мы всегда отрезаем строго определённую долю популяции, приблизительно:

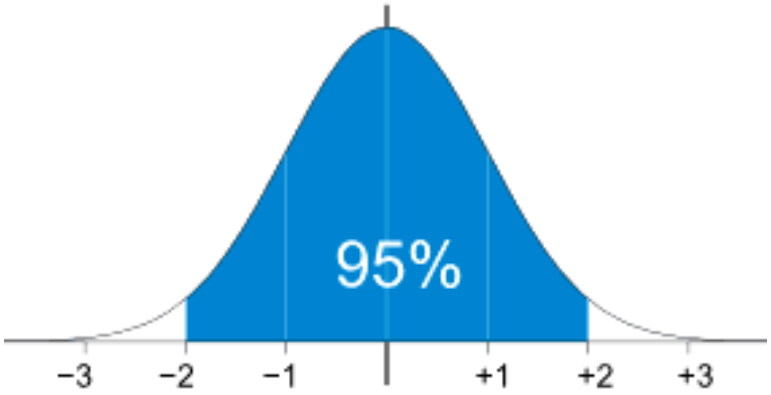


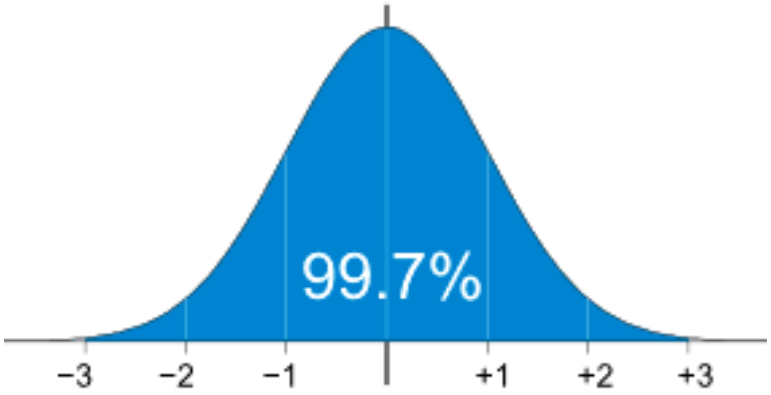
Z-оценка

(количество стандартных отклонений)

Площадь нормального распределения







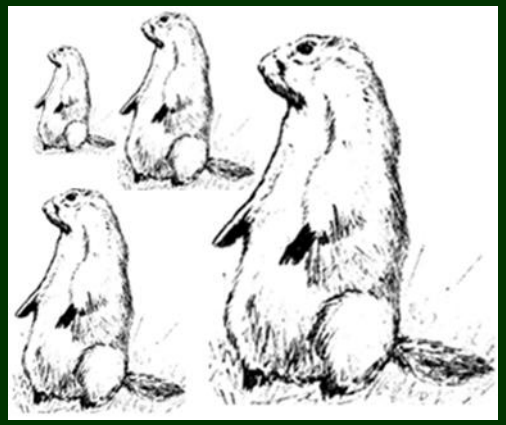
Распределение выборочных средних (sampling distribution of the means)

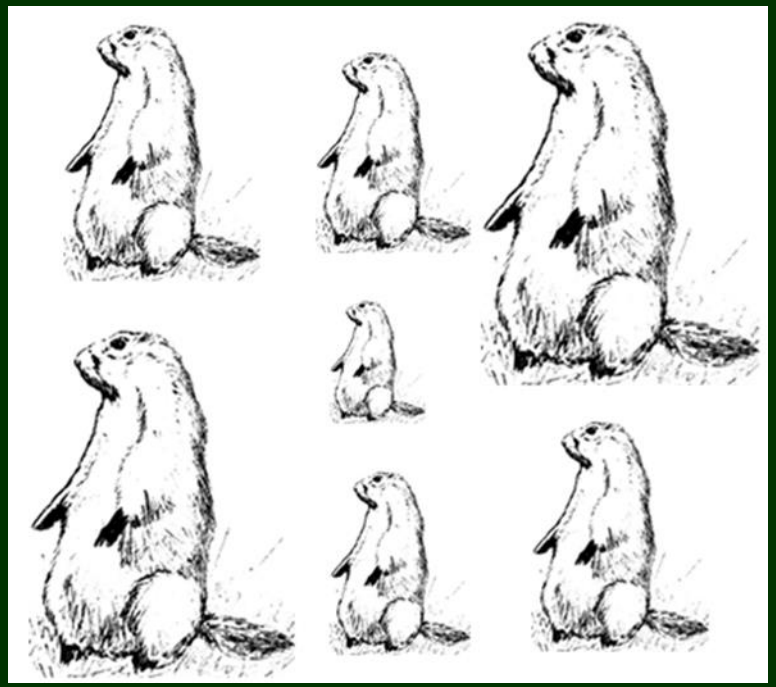
Три основные концепции в анализе данных:

1. Что такое РАСПРЕДЕЛЕНИЕ переменной и как его описывать
2. Что такое распределение ВЫБОРОЧНЫХ СРЕДНИХ и как оно связано с распределением переменной
3. Что такое СТАТИСТИКА КРИТЕРИЯ

популяция

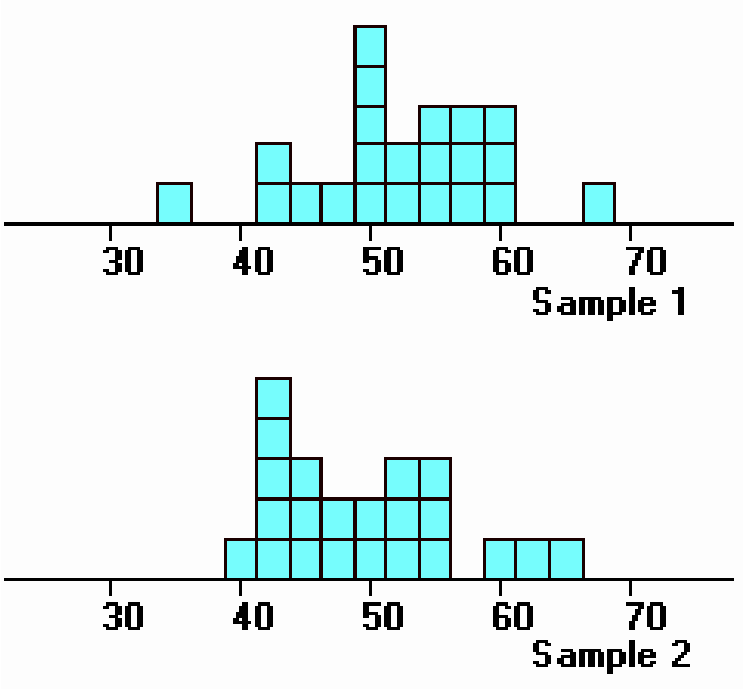
выборка





Распределение выборочных средних (sampling distribution of the means)

Ещё раз центральный статистический вопрос: что мы можем сказать обо всей ПОПУЛЯЦИИ, если всё, что у нас есть, это лишь ВЫБОРКА из неё?



На 1-м курсе института 25 групп по 22 студента.

Средняя масса студента – μ=50 кг, σ = 5 кг.

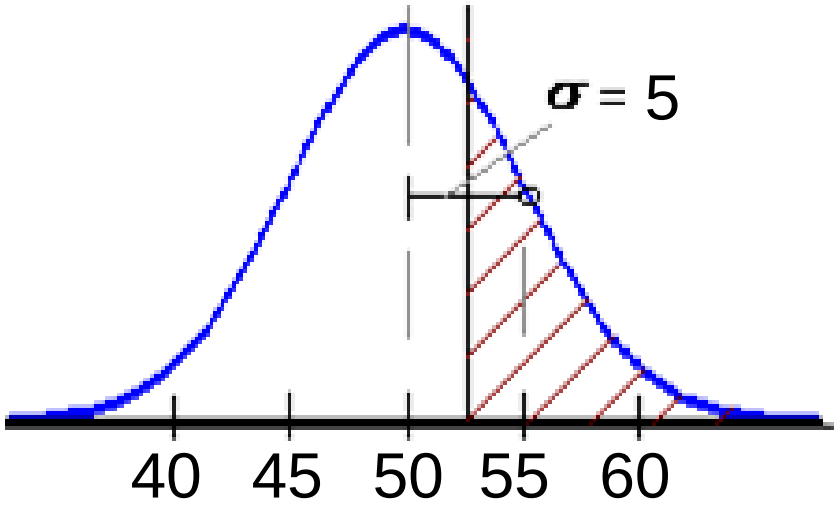
Посчитаем средние массы для каждой группы!

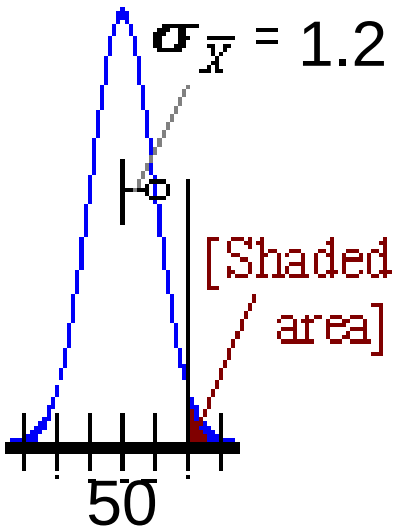
.....

Форма распределений маленьких выборок не обязательно должна удовлетворять критериям нормального распределения.

Распределение выборочных средних (sampling distribution of the means)

Мы посчитали средние массы студентов в КАЖДОЙ группе, и теперь построим **распределение** из этих СРЕДНИХ значений!





Оно будет намного УЖЕ распределения всех студентов 1-го курса, и УЖЕ, чем каждое из распределений из отдельных групп

Это и будет распределение выборочных средних (sampling distribution of the means)

	Популяция (1-й курс)		Выборка (группа)		Распределение выборочных средних
среднее	μ	$\approx$	$\overline{X}$	$\approx$	$\mu_{\overline{X}}$
стандартное отклонение	σ	$\approx$	s	>>	$\sigma_{\overline{X}}$

Стандартная ошибка среднего (Standard error = SE)

Стандартная ошибка (умноженная приблизительно на 2) формирует доверительный интервал, в рамках которого находятся значения переменной в генеральной совокупности (с вероятностью 95%)

Доверительным интервалом (ДИ)называется интервал, в который попадают измеренные в эксперименте значения, соответствующие ***доверительной вероятности***

Метод доверительных интервалов разработал американский статистик **Ежи Нейман** (термин введен им в **1934** году), исходя из идей Рональда Фишера

Ежи Нейман
(*Jerzy Neyman 1894-1981)*

ДИ, используемый в статистике при *интервальной оценке* статистических параметров, более предпочтительней при небольшом объёме выборки, чем *точечная оценка*

Точечная и интервальная оценка

Термины точечная оценка (point estimation) и интервальная оценка (interval estimation) впервые использовал в ***1943*** году американский математик ***Генри Шеффе*** в статье *«Statistical Inference in the NonParametric Case»*

Генри Шеффе
(*Henry Scheffé 1907-1977*)

Точечная оценка определяется одним числом (среднее значение, стандартное отклонение и т.д.)

Интервальная оценка определяется двумя числами – концами интервала (доверительный интервал)

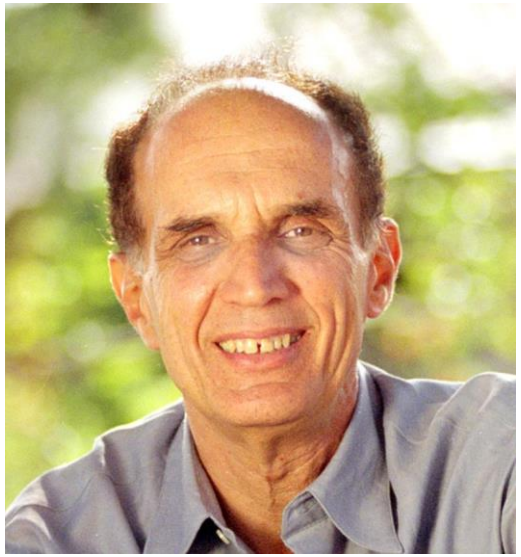
Бутстреп (bootstrap, bootstrapping)

Доверительный интервал можно построить с использованием ***ресэмплинг-техник*** (***resampling***): *методом складного ножа* или более современным *методом бутстрепа*.

Бутстреп — это современная ресэмплинг-техника, то есть техника, основанная на взятии повторных (re…) выборок (…sample).

Понятие введено в **1977** году **Брэдли Эфроном** (первая публикация относится к 1979 году).

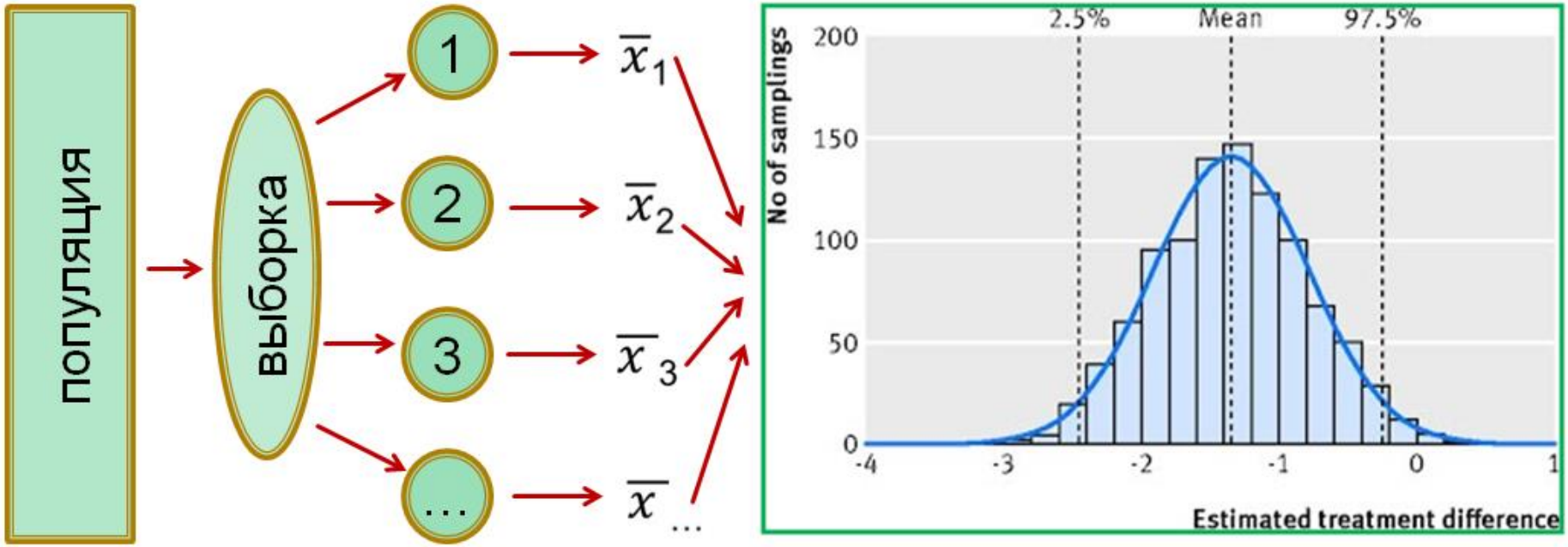
Данный метод является методом *непараметрической статистики**.



Брэдли Эфрон
(Bradley Efron p.1938)

* - о непараметрической статистике см. далее - в лекции, посвященной этим методам

Например - ***выборка*** 25 студентов, измерили рост. Выписываем значения на карточки, перемешиваем и случайным образом вытягиваем одну. Записываем результат и возвращаем карточку обратно. Повторяем процедуру ещё 24 раза. Получаем ***сгенерированную выборку***. Рассчитываем интересующую нас статистику. В нашем случае – среднее значение. Повторяем множество раз (1тыс. и более). Получаем ***множество средних значений***. Строим распределение средних и с концов отрезаем по 2,5% площади. Остаётся 95% ДИ для среднего, вычисленный с помощью процедуры бутстрепа **процентильным методом**.



ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

Определяет форму, среднее и разброс в распределении выборочных средних

- **Форма:** с увеличением размера выборок (групп) распределение выборочных средних приближается к нормальному распределению (независимо от формы распределения популяции).

- **Среднее:** среднее значение в распределении выборочных средних равно среднему значению в популяции, т.е., $\mu_{\overline{X}} = \mu$

- **Разброс:** распределение выборочных средних Уже распределения популяции на $\sqrt{n}$, где *n* – объём выборки, т.е.

$$\boxed{SE = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$$

Следствие:

если некоторая величина отклоняется от среднего под воздействием слабых, независимых друг от друга факторов, она имеет нормальное распределение.

Поэтому нормальное распределение так широко распространено в природе!

Спасибо за внимание!