

Занятие 1

Основные понятия.
Описательная
статистика.

Статистические методы анализа в биологии и медицине

I семестр 2020 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**

Зачтено - с 56 баллов в сумме

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** - это результаты **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 биологические явления. Значения признака определяются при помощи специальных шкал наблюдения.

В 1946 году С. 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.**

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

ДАННЫЕ

Категориальные
(качественные)

Количественные
(числовые)

Номинальные
nominal

Порядковые
ordinal

шкала
отношений
ratio scale

интервальная
шкала
interval scale

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

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

Дискретные
discrete

Непрерывные
continuous

Целочисленные
значения,
типичные для
счета

Любые значения
в определенном
интервале

← Потеря информации и точности

шкала отношений (ratio scale):

- размер интервалов на протяжении всей шкалы одинаковый;
- существует реальное нулевое значение.

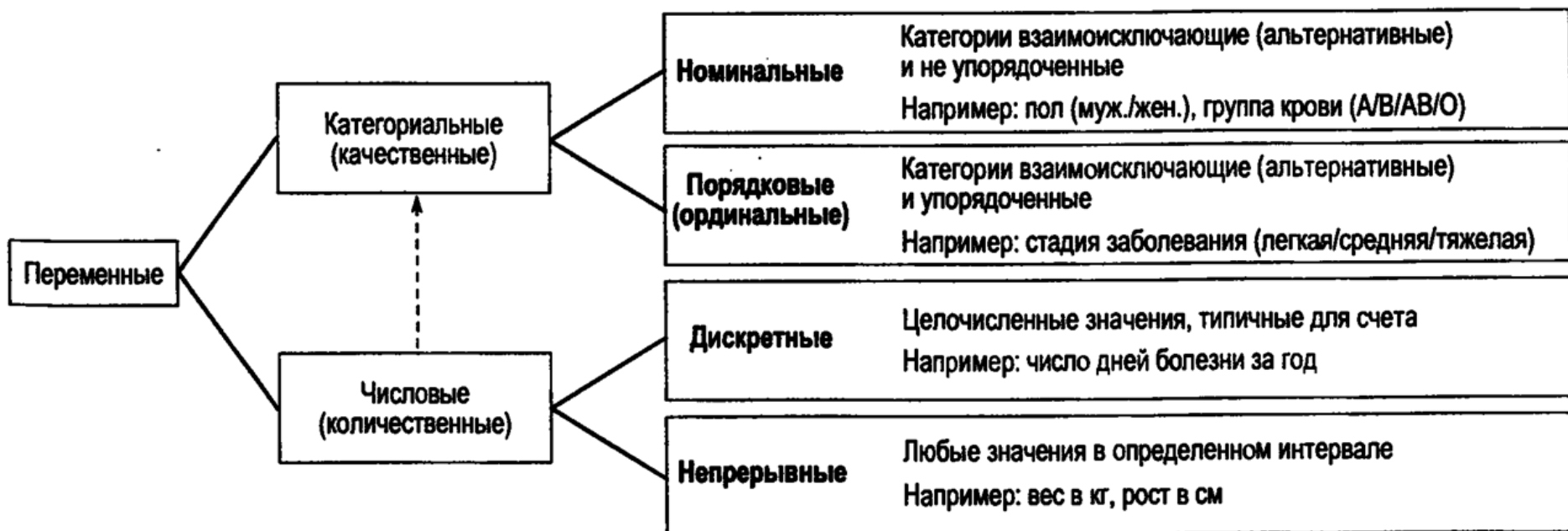
Примеры: масса тела, размер выводка, объём

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



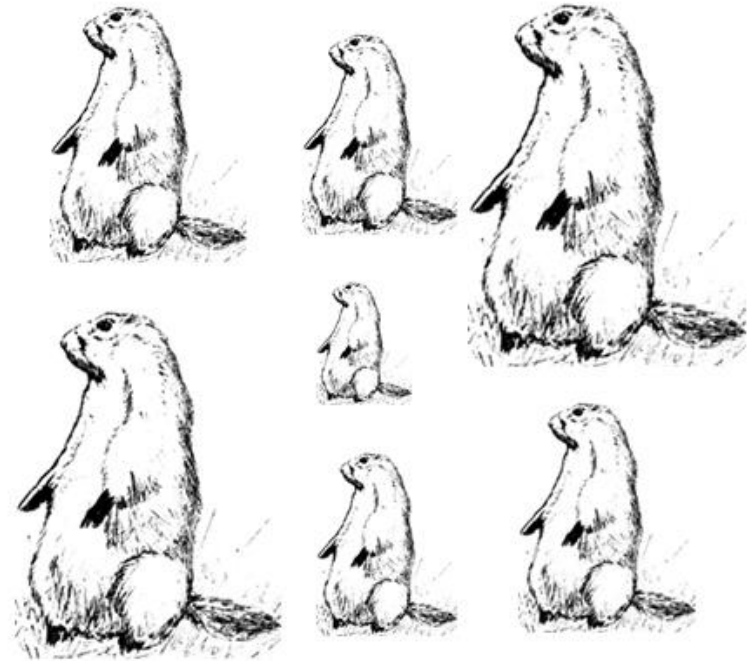
Например, мы изучаем популяцию сурков

наблюдение



популяция – совокупность всех интересующих нас объектов

ВЫБОРКА

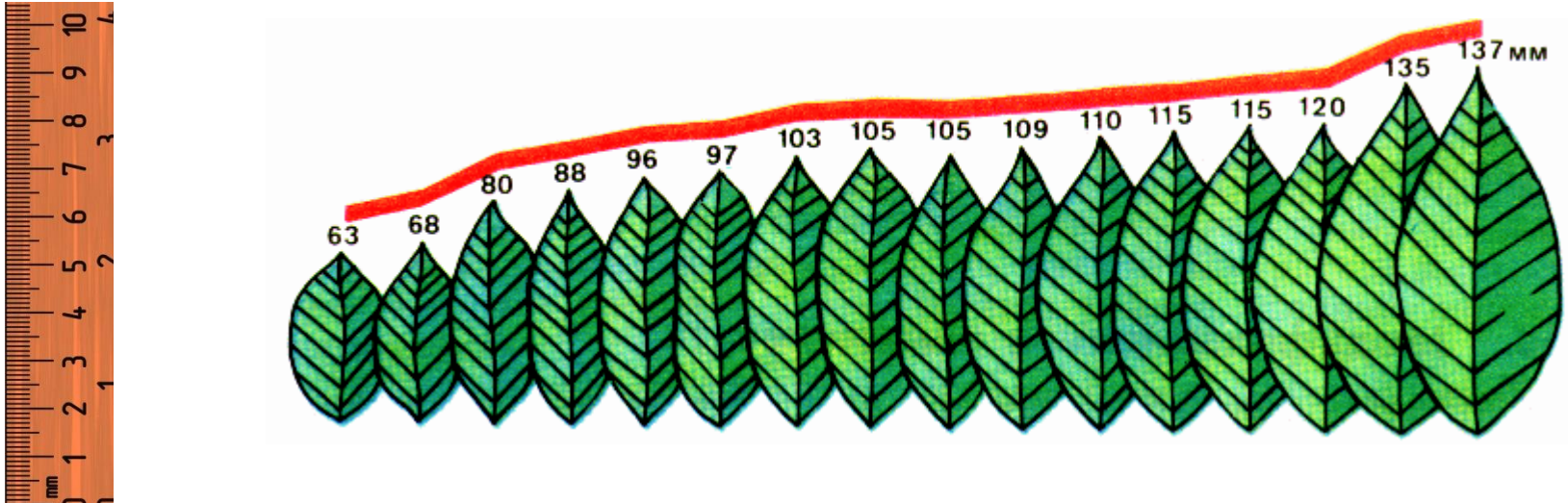


Описательная статистика: ОПИСЫВАЕМ ВЫБОРКУ

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

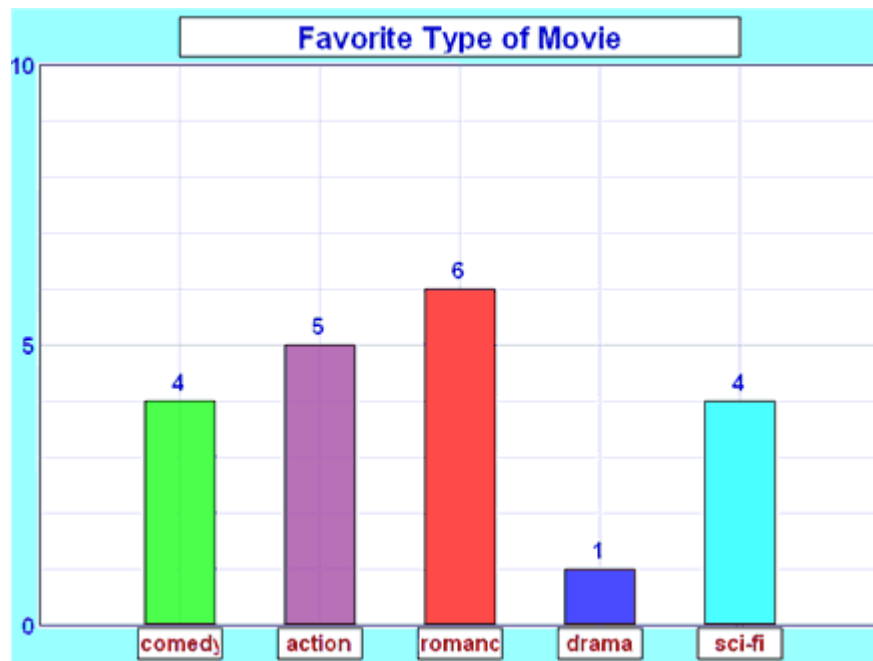
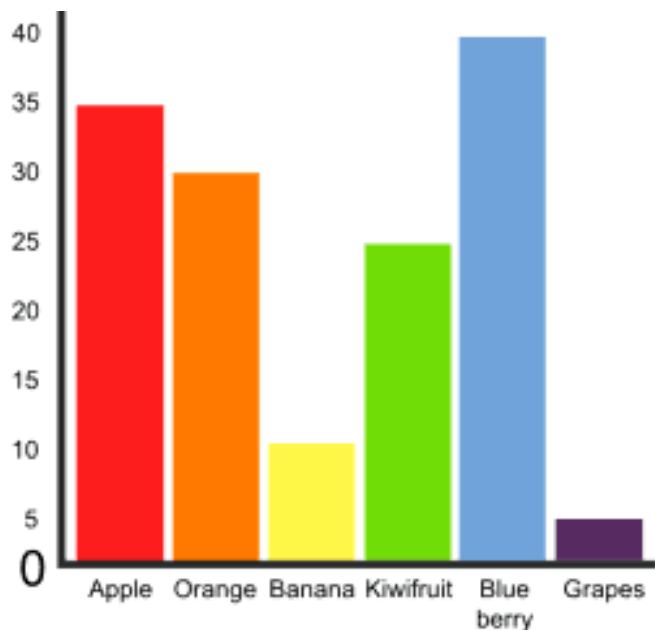
Три основные концепции в анализе данных:

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**

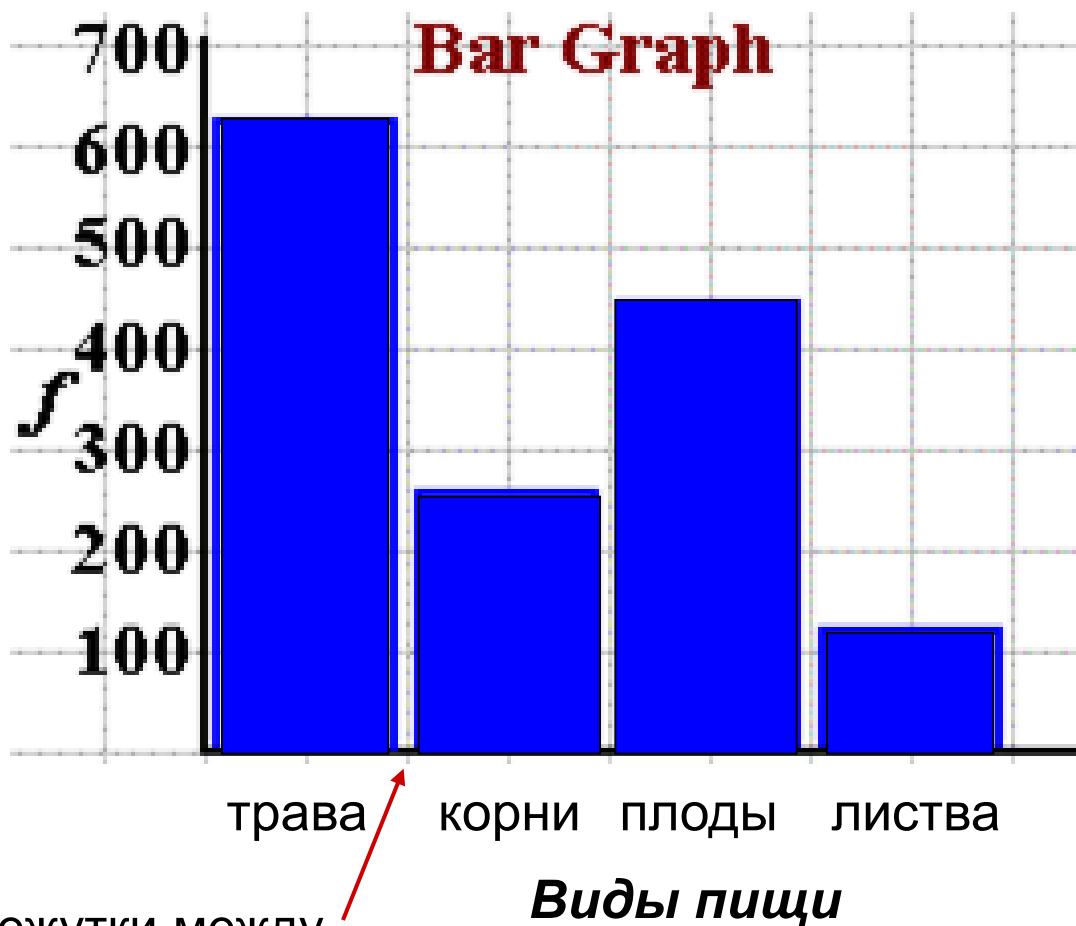


Частотное распределение переменной (frequency distribution) – это соответствие между значениями переменной и их вероятностями (на практике – количеством таких значений в выборке)

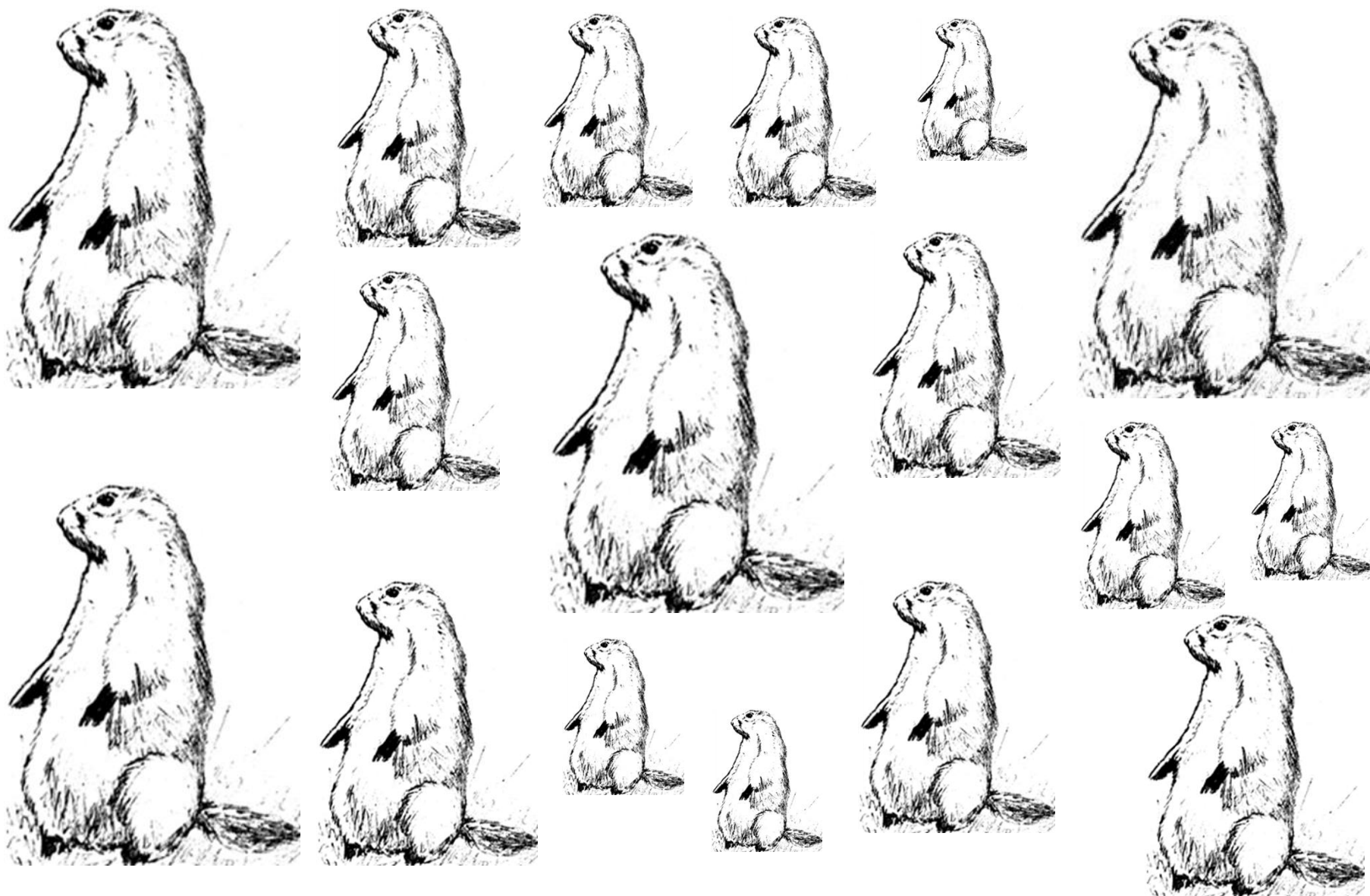
Можно представить в виде таблички или картинки.



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



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.

Частотью 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:

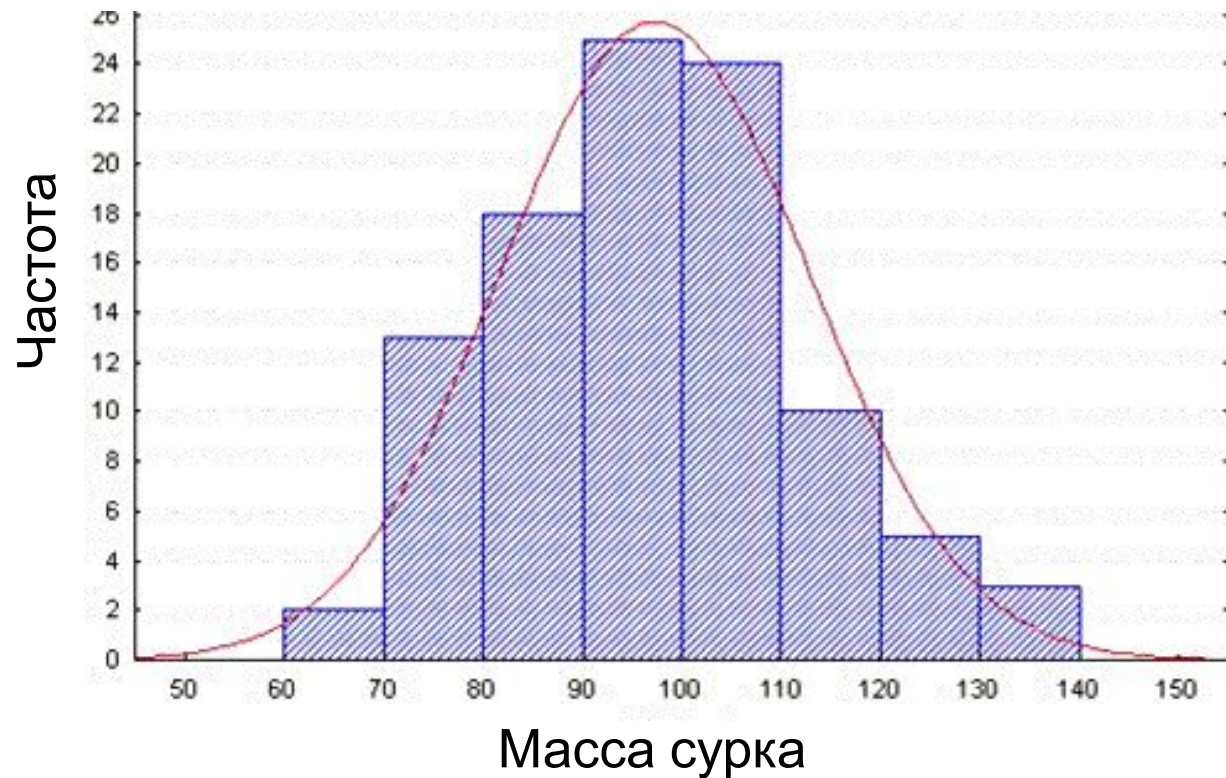
- **гистограмма**
- **полигон**
- **кумулята**

Частота – то, сколько раз встретилось данное значение переменной

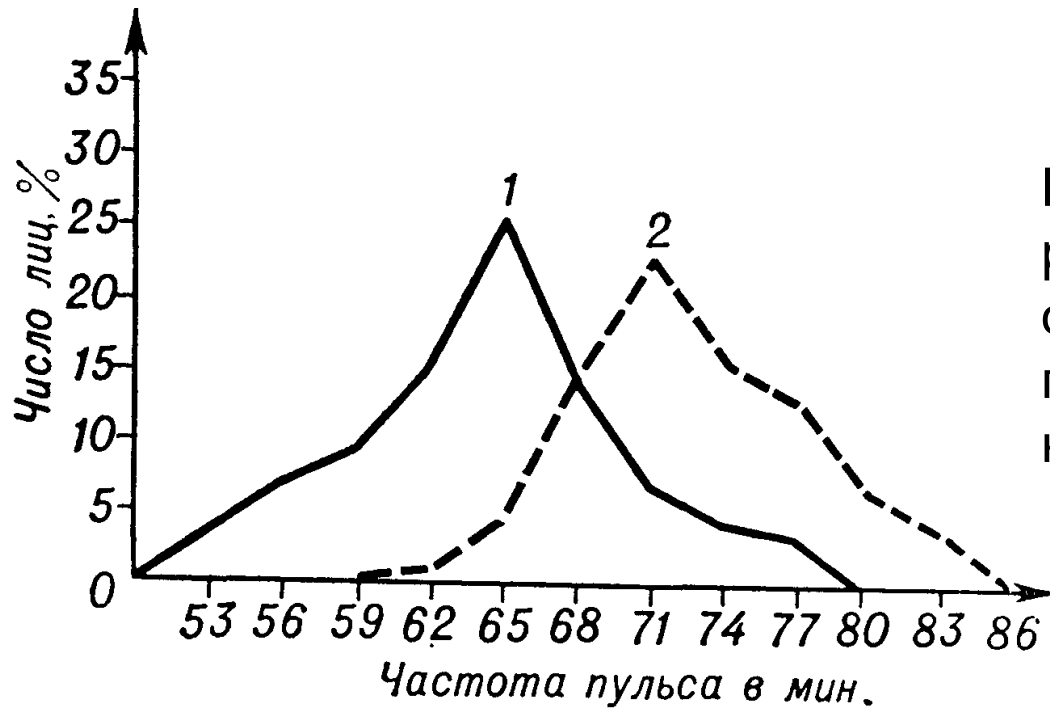
Гистограмма – графическое представление частотного распределения, разбитого по интервалам, где высота столбика отражает **ЧАСТОТУ**

Интервалы должны быть:

- одного размера,
- не должны иметь общих точек,
- для биологических данных – **10-20** интервалов



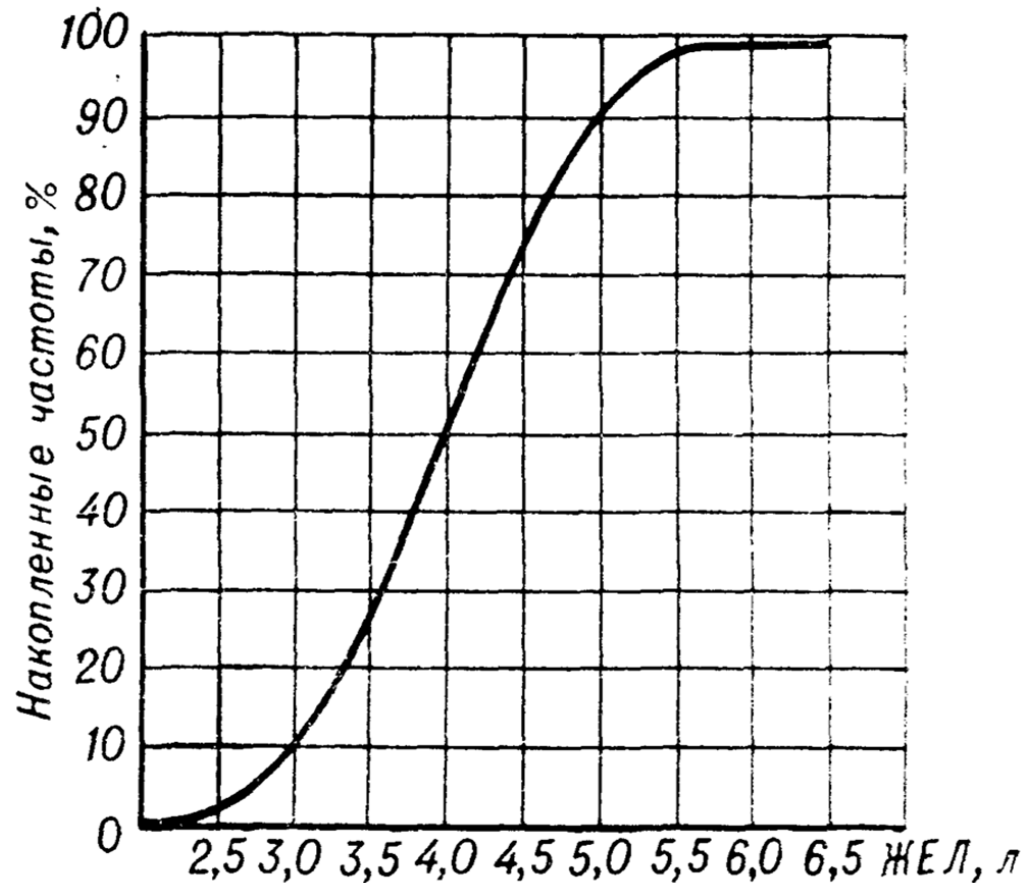
Полигон используют для дискретных рядов. На оси абсцисс в произвольном масштабе откладывают значения аргумента, а на оси ординат - значения частот или относительных частот.



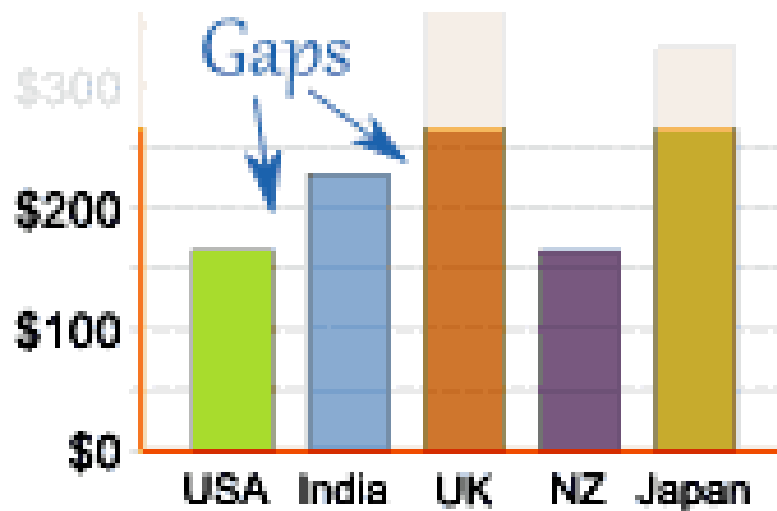
Полигон (кривая распределения) обследованных по частоте пульса: 1 — до физической нагрузки, 2 — после нагрузки.

- 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.

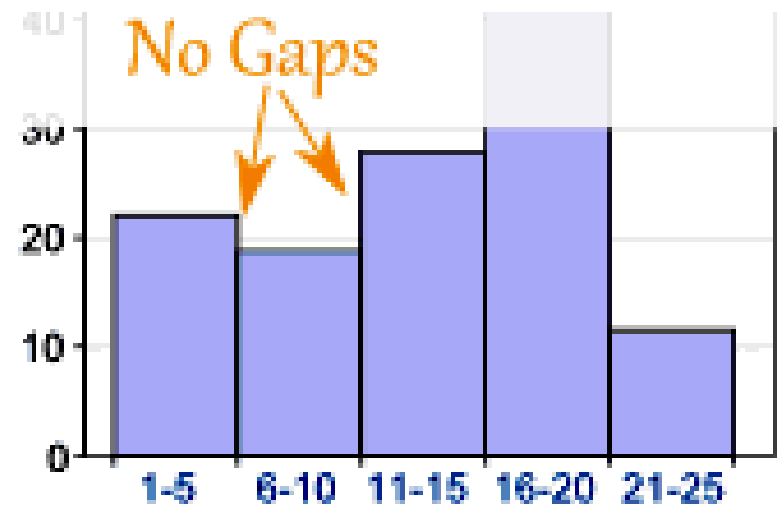


- На оси абсцисс откладывают значения аргумента, а на оси ординат - накопленные частоты.
- Строят точки, абсциссы которых равны вариантам (в случае дискретных рядов) или верхним границам интервалов (в случае интервальных рядов), а ординаты - соответствующим частотам (накопленным частотам). Эти точки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



← Categories →

Bar Graph

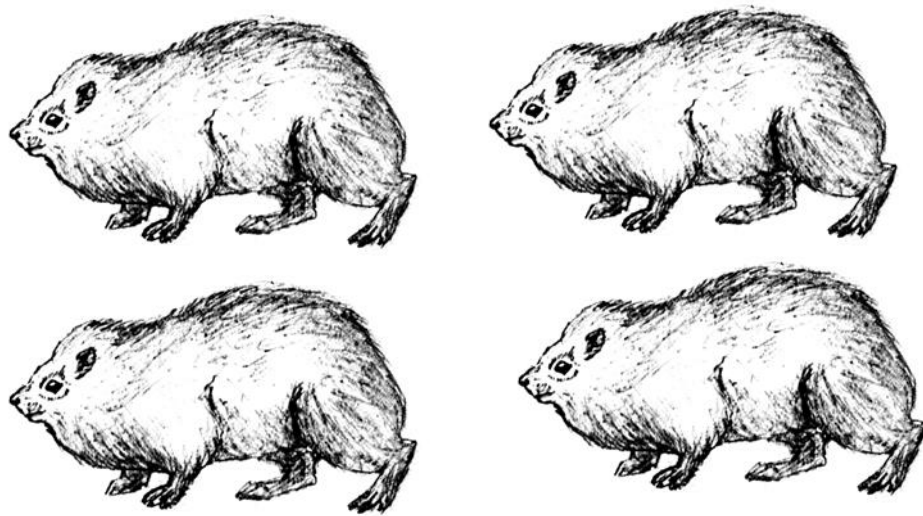


← Number Ranges →

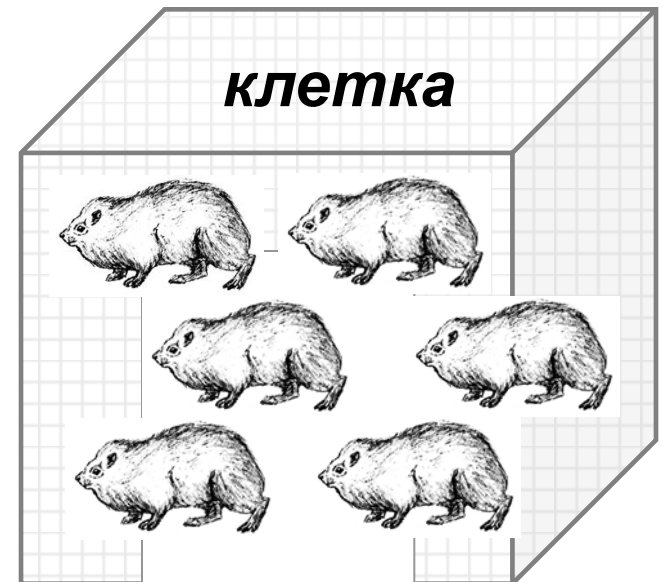
Histogram

Выборка должна быть **РЕПРЕЗЕНТАТИВНОЙ**, т.е. её свойства должны отражать свойства популяции.

Для этого она должна быть **СЛУЧАЙНОЙ** (random) – т.е., 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

Как описать частотное распределение переменной?

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

1. «**Середина**» распределения;
2. «**Ширина**» распределения;
3. **Форма** распределения

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

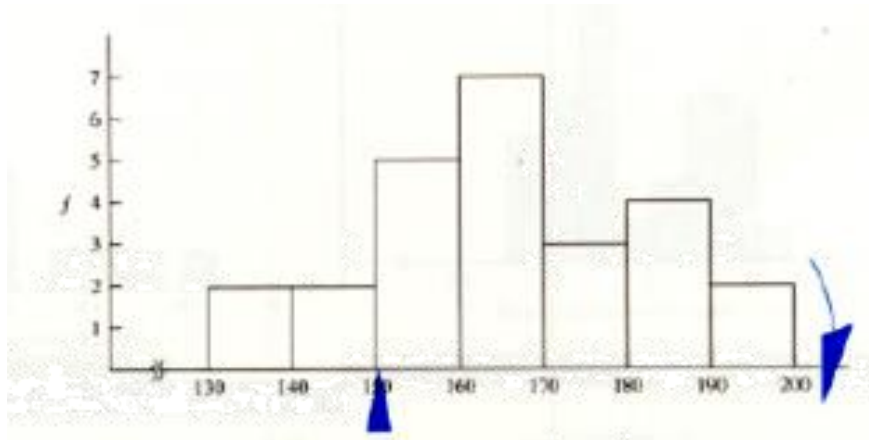
«Середина» распределения



Все они могут служить оценками популяционного среднего.

Среднее в выборке – наиболее эффективная и **несмещённая** оценка.

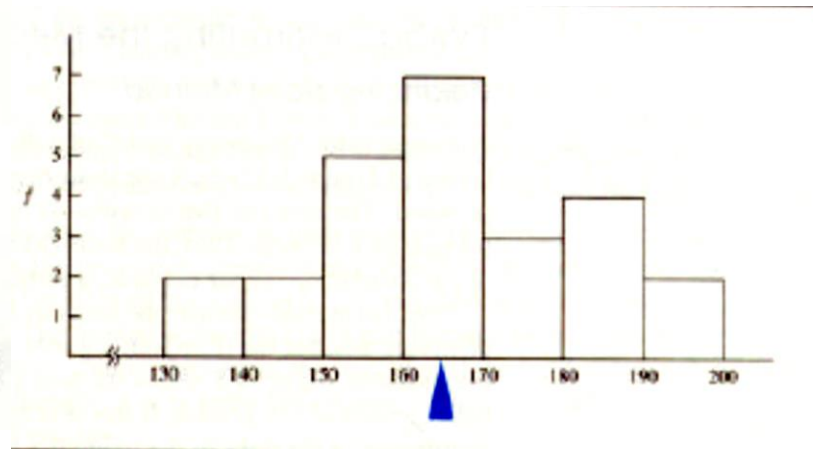
Среднее значение – сумма всех значений переменной, делённая на количество значений



*«balancing point» method

Среднее для **выборки**

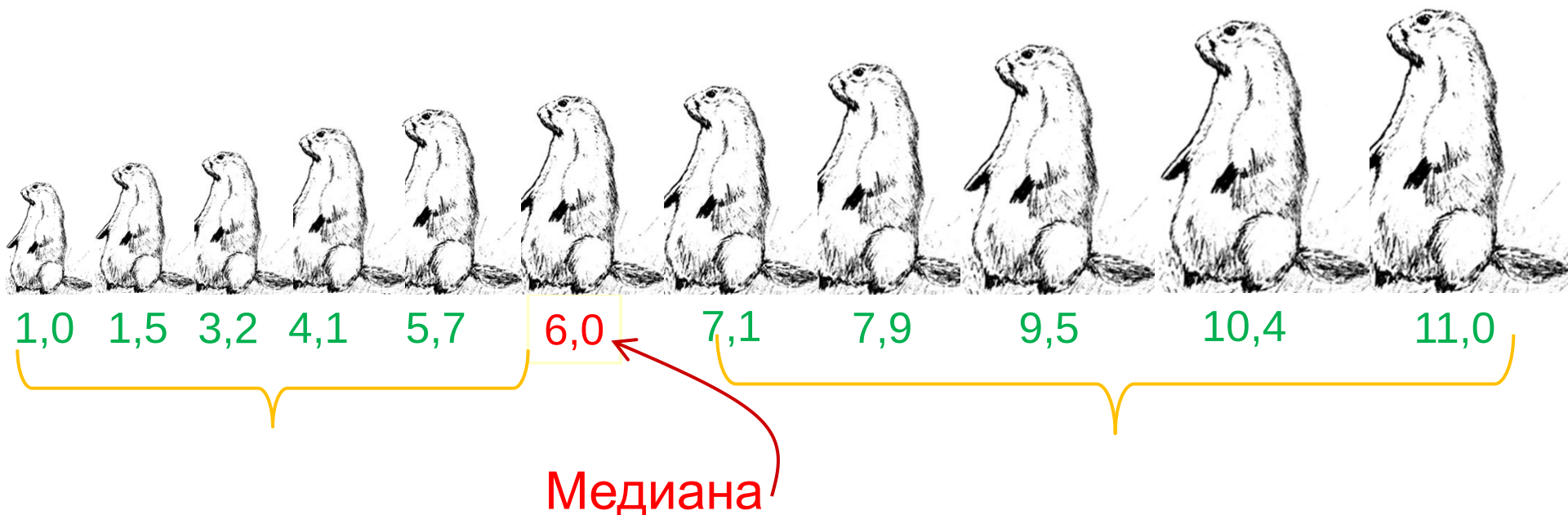
$$\bar{X} = \frac{\sum_i X_i}{n}$$



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** (... процентиля);
- ✓ **N** (... квантили).

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

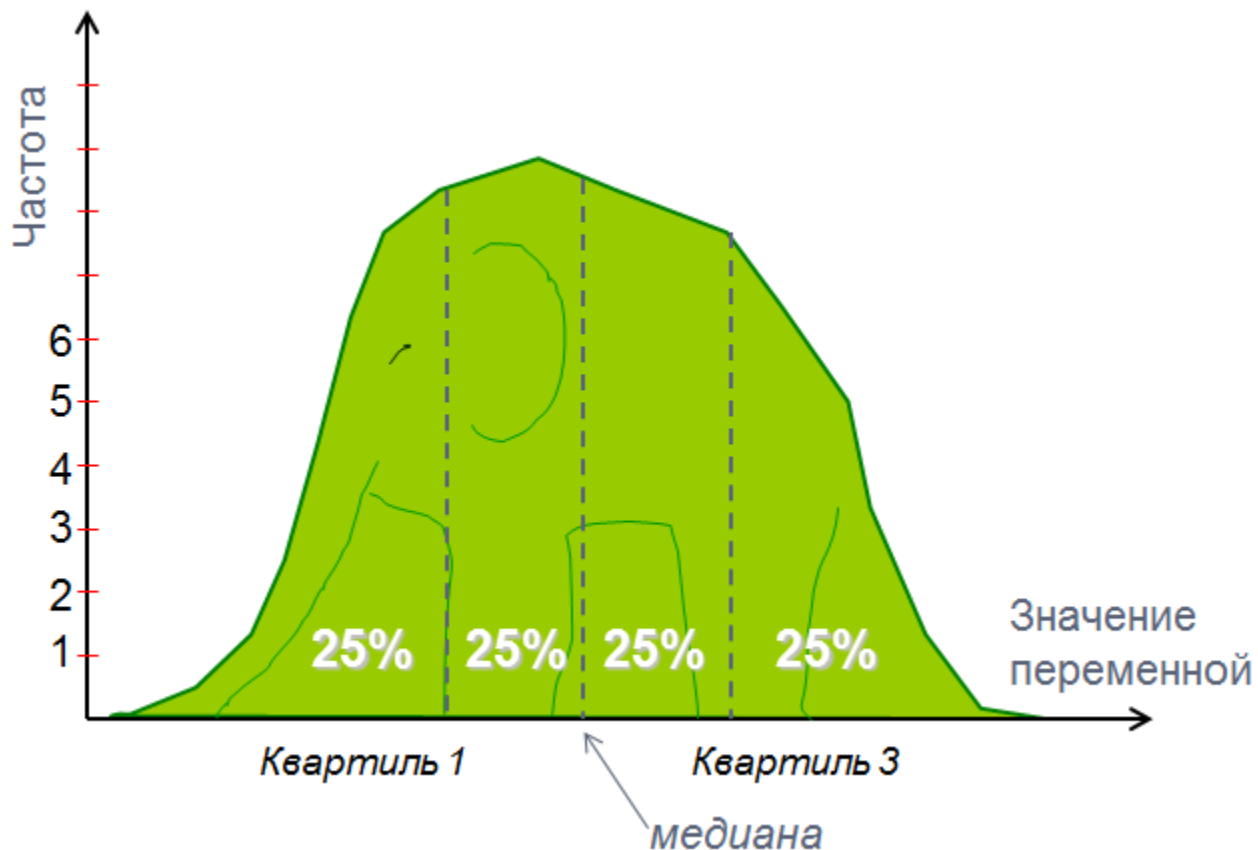
1-я квартиль = 25%
процентиль

3-я квартиль = 75%
процентиль

**Интерквартильный
размах**

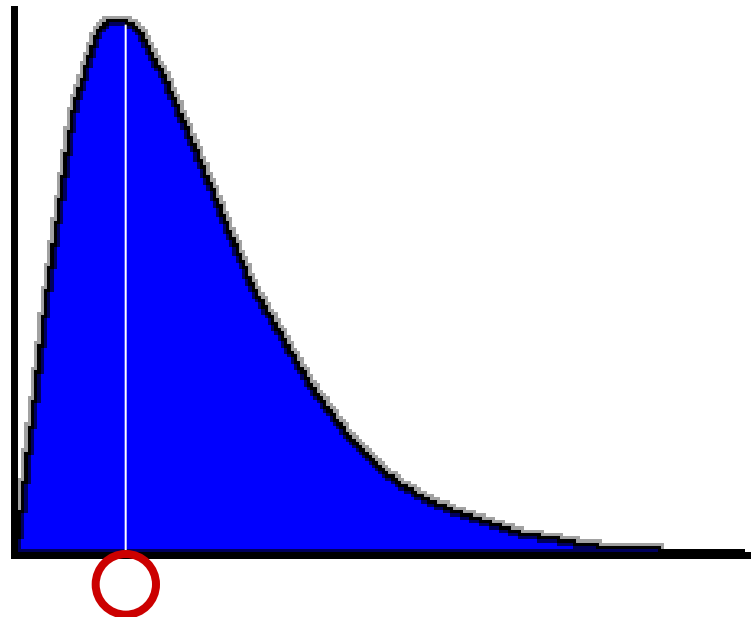
– разница между
третьей

и первой
квартилями.



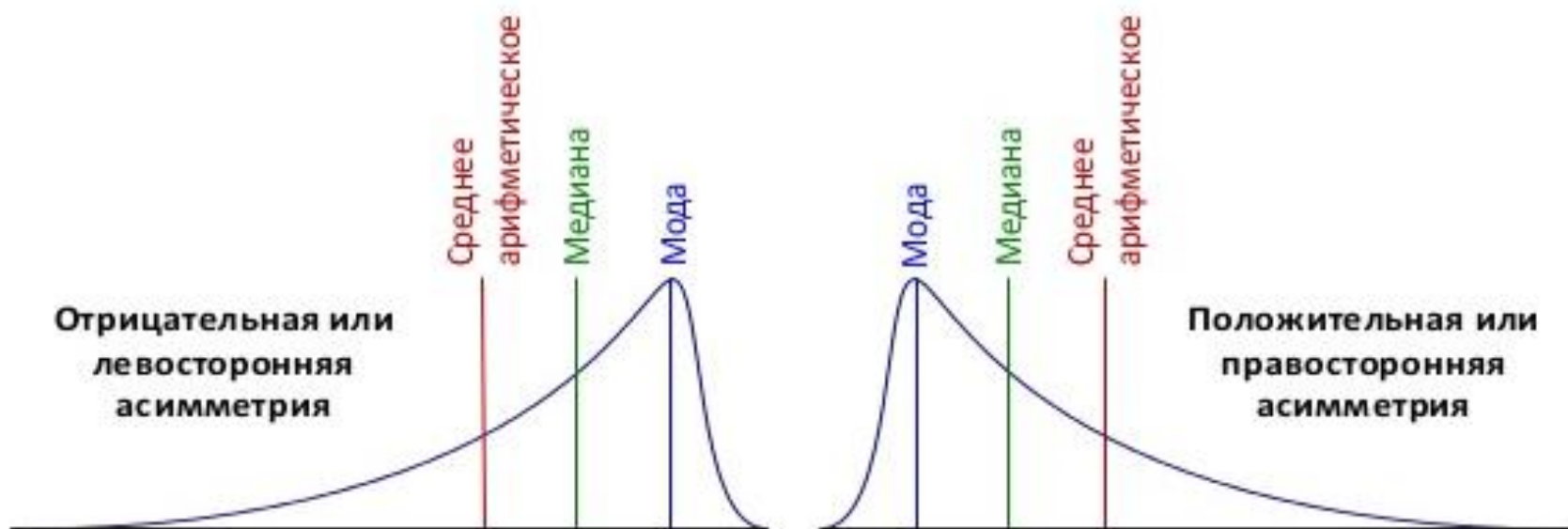
Мода (mode) – наиболее часто встречающееся значение

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

Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения



К появлению перекоса чувствительнее всего среднее значение

Частотное распределение переменной (frequency distribution)

«Середина» распределения

Тип средней	Преимущества	Недостатки
Среднее	• Используются все значения набора данных • Определяется математически выполнимым алгебраическим выражением • Известно выборочное распределение (раздел 6)	• Искажено выбросами • Искажается асимметричными данными
Медиана	• Не искажается выбросами • Не искажается асимметричными данными	• Игнорирует большую часть информации • Не определяется алгебраически • Усложняется в выборочном распределении
Мода	• Легко определяется для категориальных данных	• Игнорирует большую часть информации • Не определяется алгебраически • Нензвестно выборочное распределение
Среднее геометрическое	• До обратного преобразования имеет те же самые преимущества, что и среднее	• Подходит, если логарифмическое преобразование образует симметричное распределение
Взвешенное среднее	• Те же самые преимущества, что и у среднего • Приписывает соответствующий вес каждому наблюдению • Алгебраически определяется	• Вес должен быть известен или оценен

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 размах в популяции. Если в статье приводится размах, следует привести ещё какую-нибудь характеристику разброса.

* Это лишь основные параметры разброса

Стандартное отклонение (standard deviation)

Для **выборки**:

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

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

Для популяции:

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

Сумма квадратов
(*sum of squares = SS*)

Стандартное отклонение зависит от всех значений переменной.

Измеряется в тех же единицах, что и переменная!

Дисперсия (variance)

Для **выборки**:

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

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, возведённых в квадрат (что не всегда удобно).

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

Коэффициент вариации
(Coefficient of variation)

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

- 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

Частотное распределение переменной (frequency distribution)

Разброс распределения

Мера рассеяния	Преимущества	Недостатки
Размах	• Легко определить	• Использует даже 2 наблюдения • Искажается выбросами • Имеет тенденцию к увеличению при росте объема выборки
Размахи, основанные на процентилях	• Не подвержены влиянию выбросов • Не зависят от размера выборки • Подходят для асимметричных распределений данных	• Грубый расчет • Невозможно рассчитать для маленьких выборок • Может использоваться даже при двух наблюдениях • Не имеет алгебраического выражения для вычисления
Дисперсия	• Использует каждое наблюдение • Имеет алгебраическое выражения для вычисления	• Единица измерения — квадрат исходных данных • Восприимчива к выбросам • Не подходит для асимметричных распределений данных
Стандартное отклонение	• Те же самые преимущества, что и у вариации • Единицы измерения те же, что и у исходных данных • Легко объяснима	• Восприимчива к выбросам • Не подходит для асимметричных распределений данных

Параметры разброса для качественных данных: Индексы разнообразия (*indices of diversity*)

Показывают, насколько равномерно данные распределены по категориям. Разнообразие считается высоким, когда распределение более-менее равномерное, и низким, когда превалирует 1-2 категории

Индекс Шеннона-Винера (или Шеннона-Уивера)

$$H = - \sum_{i=1}^k p_i \log p_i$$

p = доля объектов в той или иной категории;
 k – число категорий.

Этих индексов много для разных целей; это показатели
ОПИСАТЕЛЬНОЙ статистики!

Частотное распределение переменной (frequency distribution)

Форма распределения

Асимметрия, скошенность (skewness)

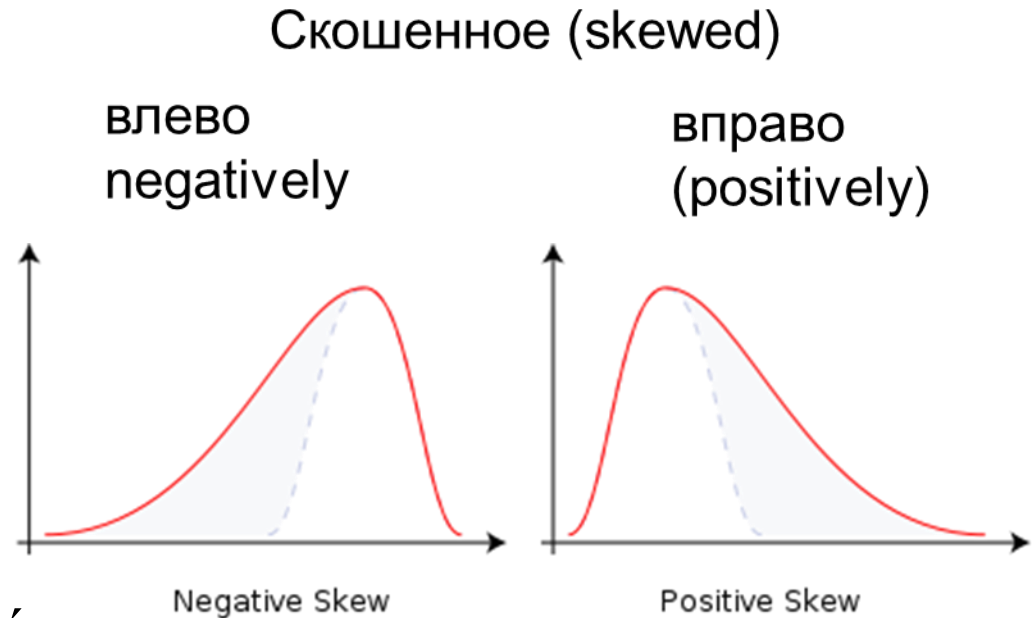
Характеристики:

$As=0$ симметричное

$As<0$ левостороннее

$As>0$ правостороннее

As - коэффициент асимметрии



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.

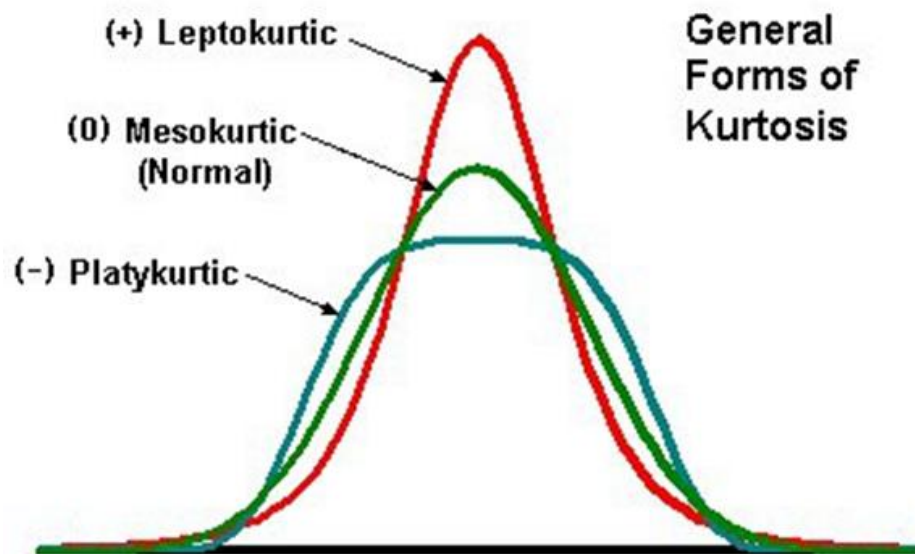
Эксцесс, куртозис (*excess, kurtosis*)

Характеристики:

$E_x=0$ нормальное

$E_x<0$ плосковершинное

$E_x>0$ островершинное

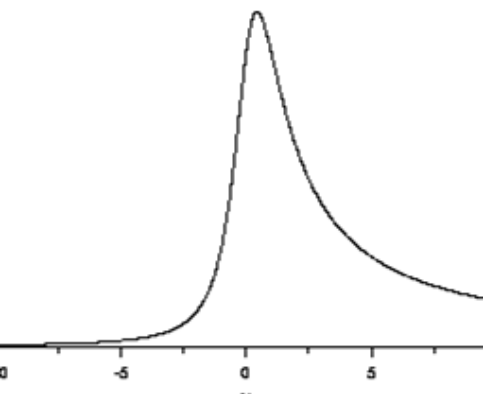


Куртозис (kurtosis) является показателем, отражающим остроту вершины и толщину хвостов одномерного распределения.

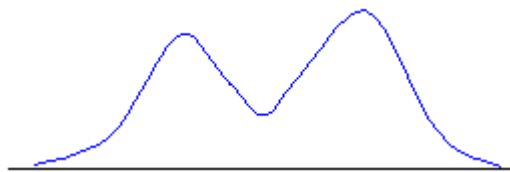
Leptokurtic distribution — это распределение с положительным эксцессом, имеющее острую вершину и толстые хвосты, а **platykurtic distribution** — это распределение с отрицательным эксцессом, имеющее плоскую вершину и тонкие хвосты. Соответственно, **mesokurtic distribution** — это распределение, похожее по куртозису на нормальное распределение.

По количеству «максимумов» (мод) распределение:

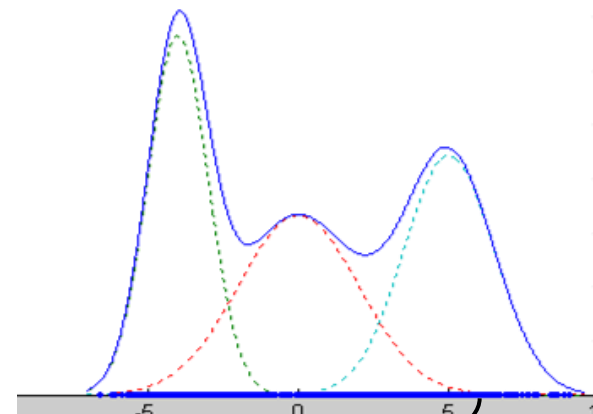
унимодальное



бимодальное



мультимодальное

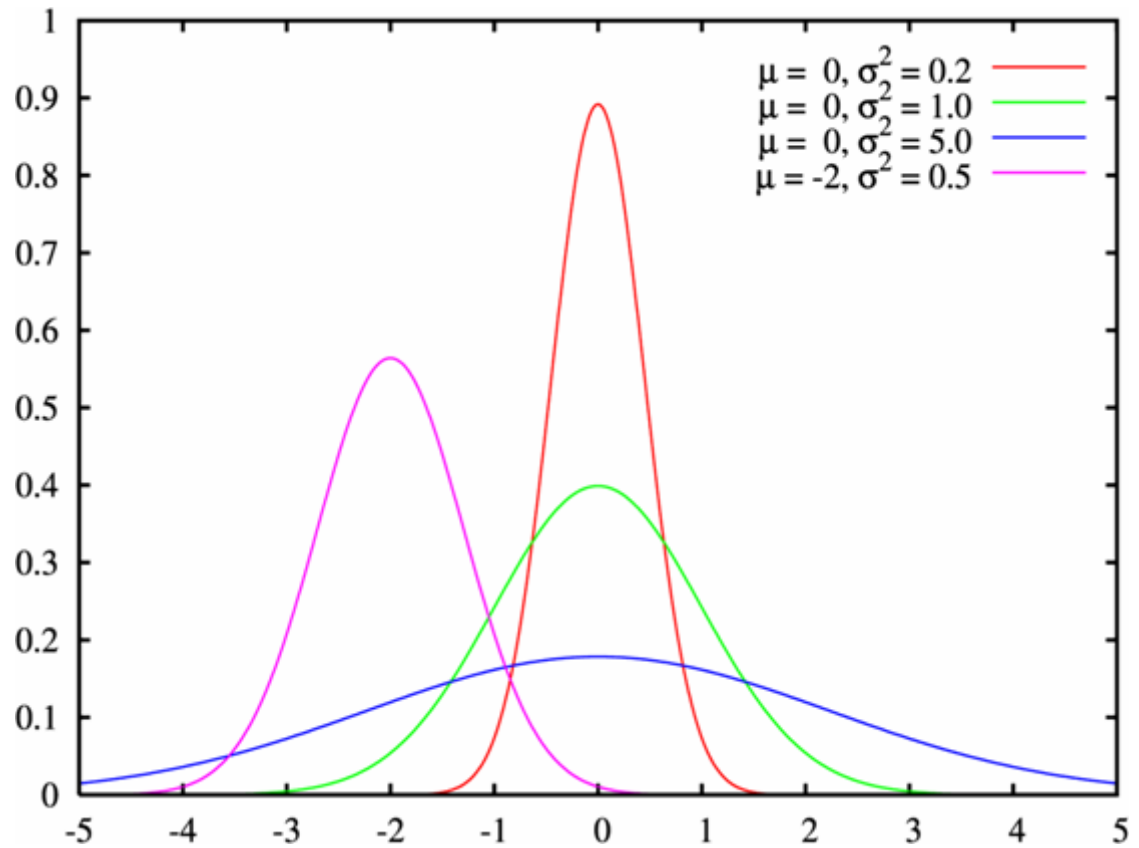


обычно возникают, если популяция имеет естественные обособленные подгруппы

Нормальное распределение (Гауссово): первое знакомство

- ✓ Унимодальное
- ✓ Симметричное
- ✓ Асимптотическое

Это
непрерывное
распределение



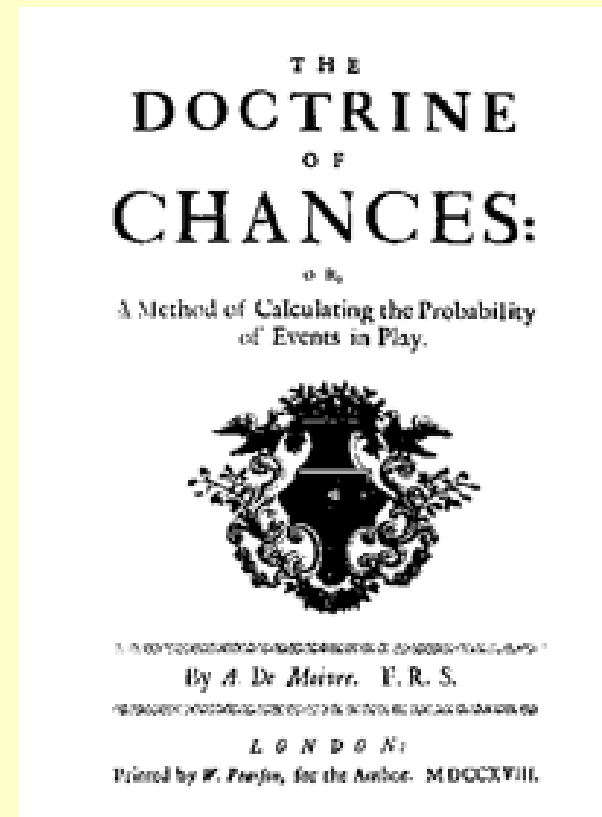
Название в честь Гаусса не совсем справедливо – первым его описал вовсе не он. Впервые нормальное распределение как предел биномиального распределения появилось в 1738 году во втором издании работы Муавра «Доктрина случайностей»

Историческая справка

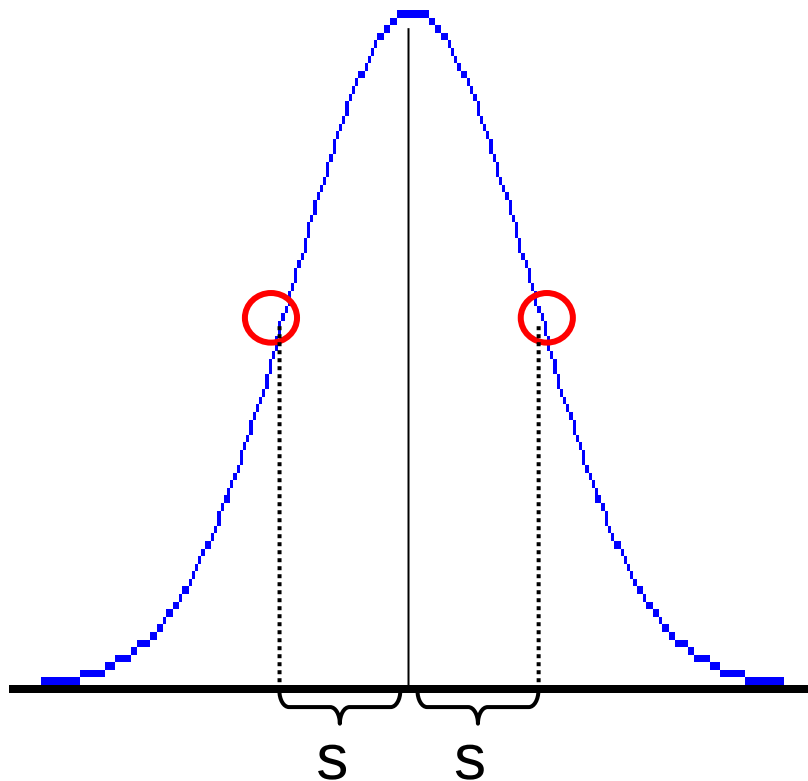
- **Муавр Абрахам**
(Moivre Abraham)
(1667-1754)



"Учение о случаях"
1718 г.



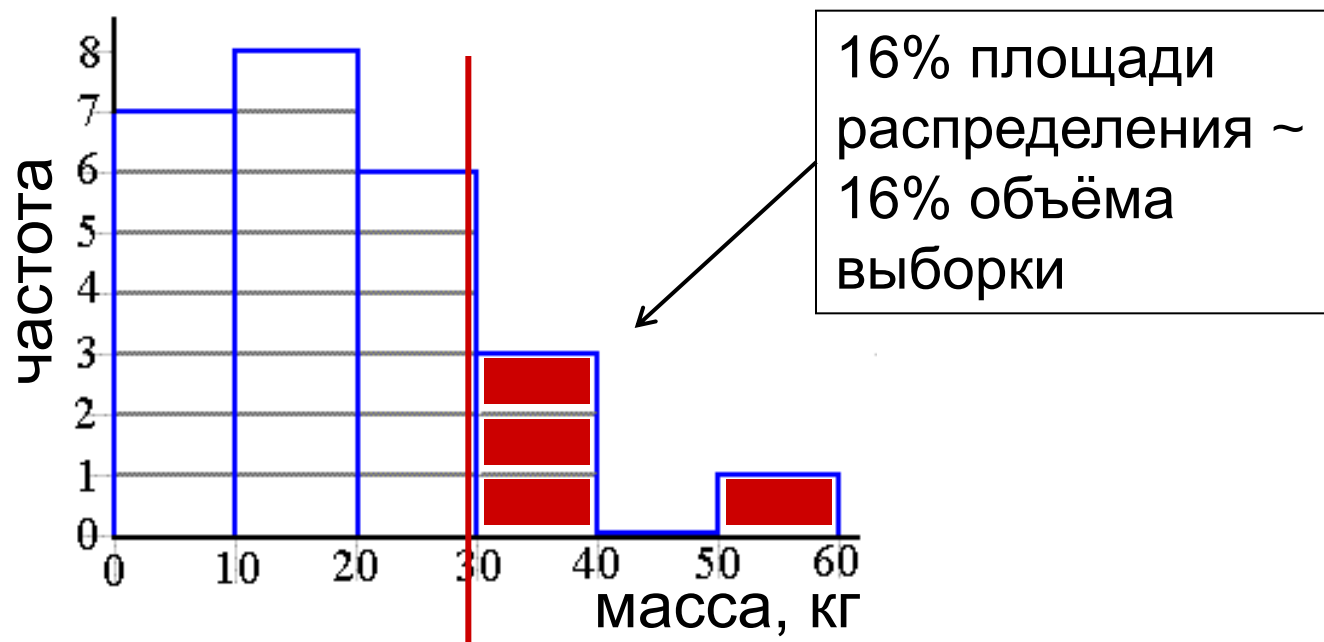
Стандартное отклонение (standard deviation):
для нормального распределения = дистанции от среднего
значения до каждой из точек перегиба



«Площадь распределения»

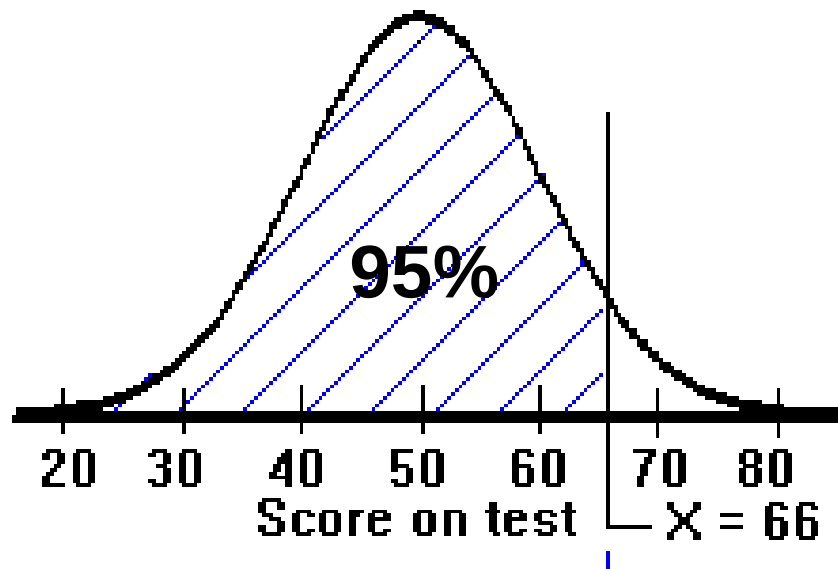
Площадь, которую занимает график распределения, соответствует количеству измерений в выборке.

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



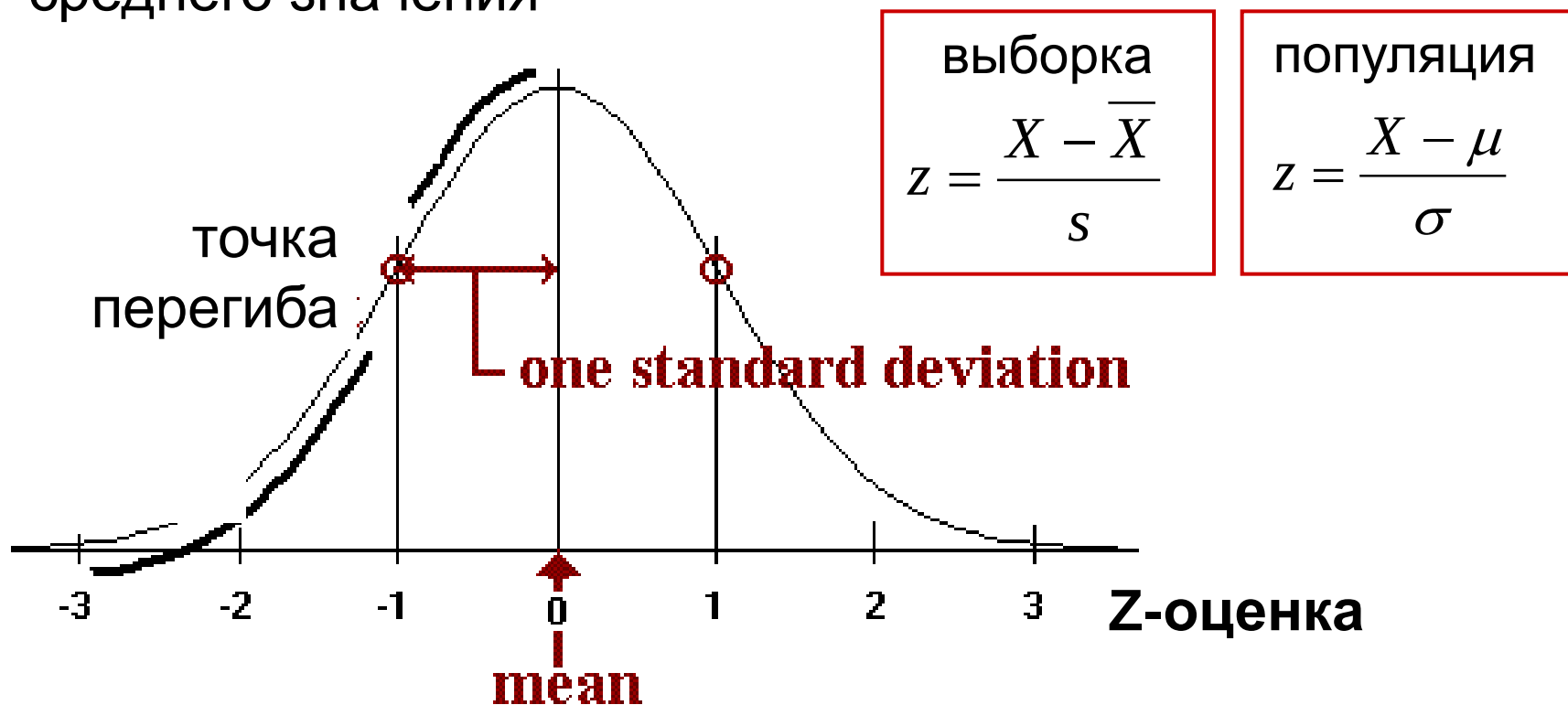
Процентили и z-оценка

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



Процентили и z-оценка

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



Площадь нормального распределения

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

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

Необыкновенное свойство:

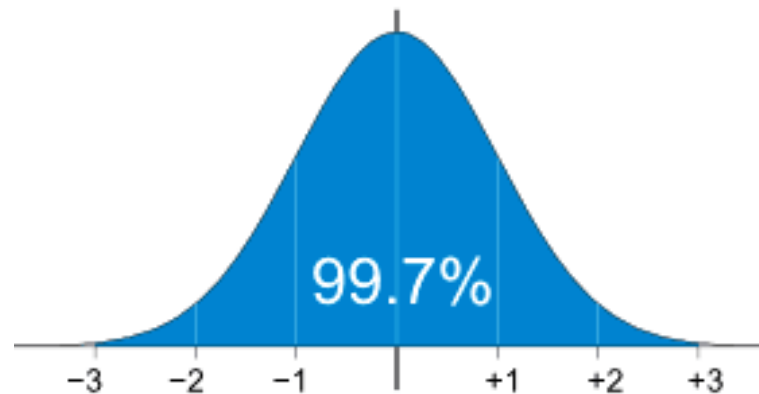
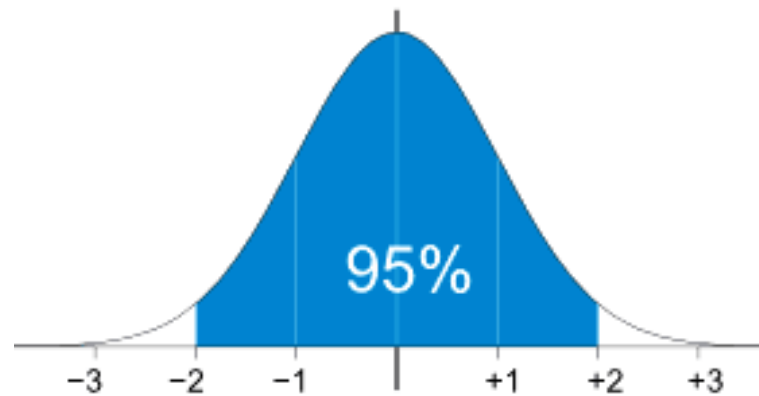
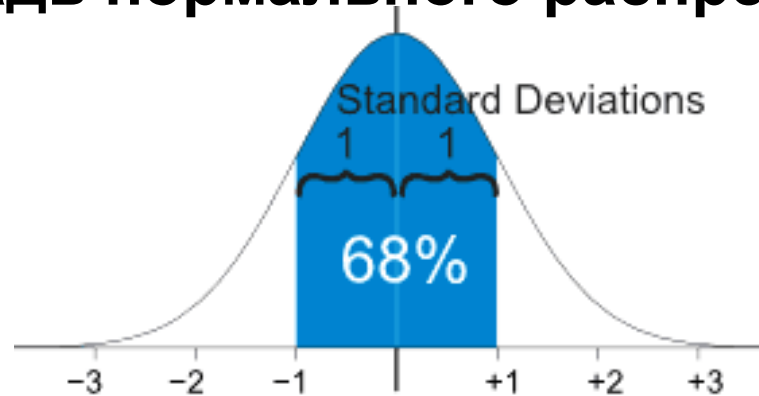
Относительные площади под участками нормального распределения всегда одинаковы!

Площадь нормального распределения

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



Площадь нормального распределения

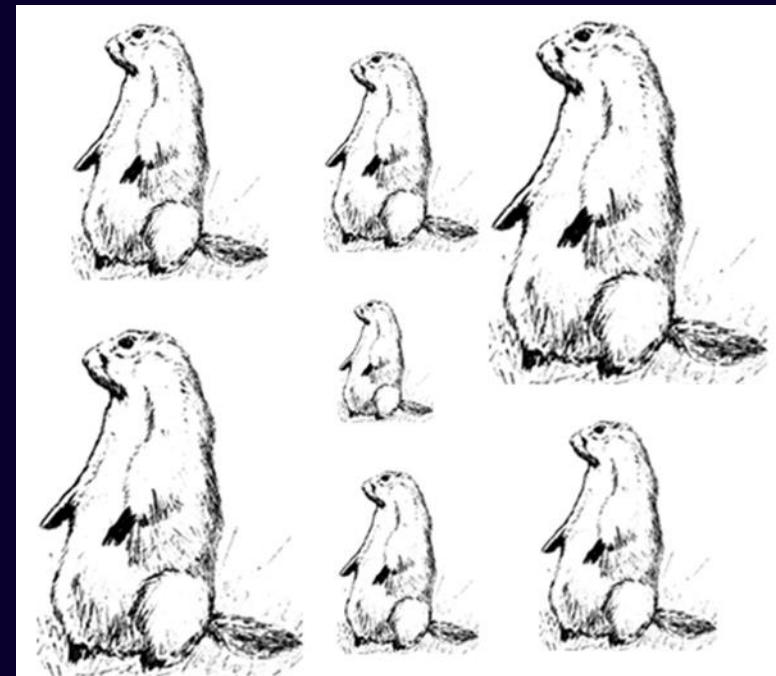


Распределение выборочных средних (sampling distribution of the means)

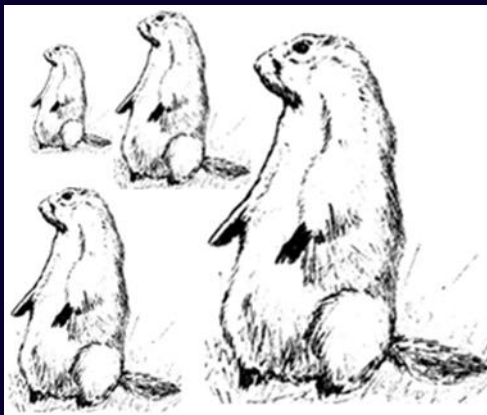
Три основные концепции в анализе данных:

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

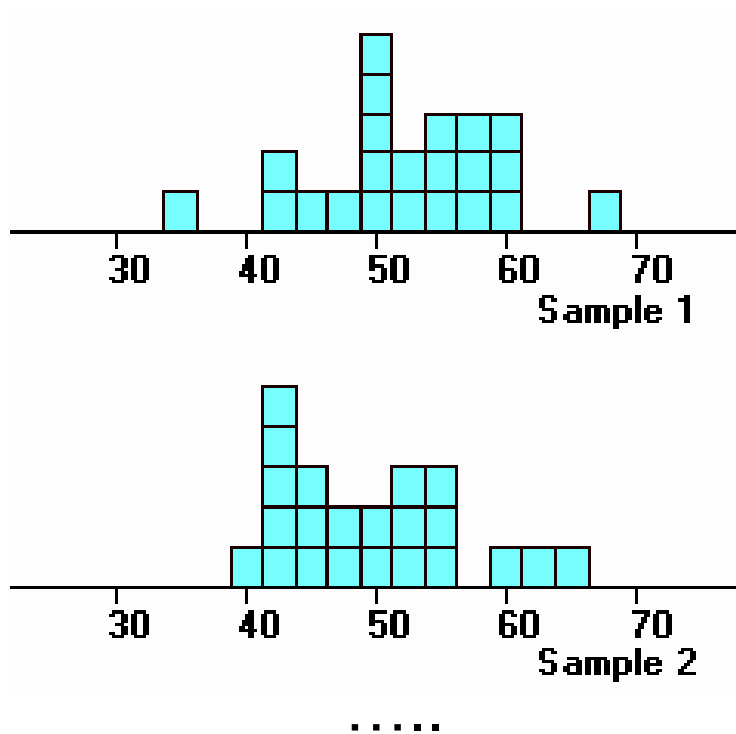


выборка



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.

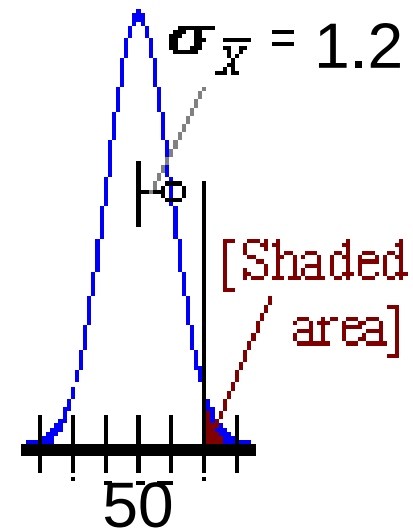
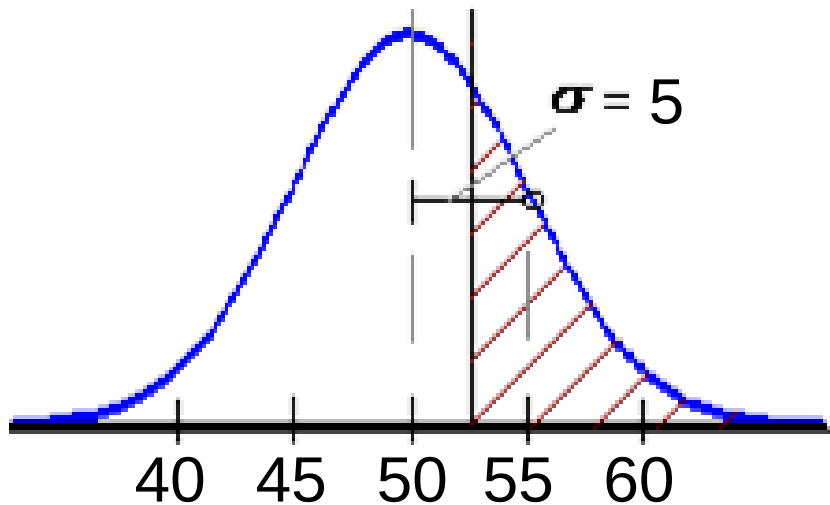
Средняя масса студента – $\mu=50$ кг, $\sigma = 5$ кг.

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)

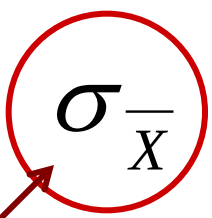
Мы посчитали средние массы студентов в КАЖДОЙ группе, и теперь построим **распределение** из этих СРЕДНИХ значений!



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)

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



 Стандартная
ошибка среднего
(Standard error = SE)

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 средних равно среднему значению в популяции, т.е., $\mu_{\bar{X}} = \mu$
- **Разброс:** распределение выборочных средних уже распределения популяции на $\sqrt{n}$, где n – объём выборки, т.е.

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

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!

