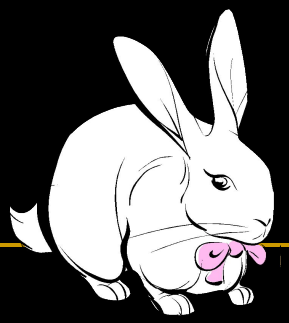
Занятие 1

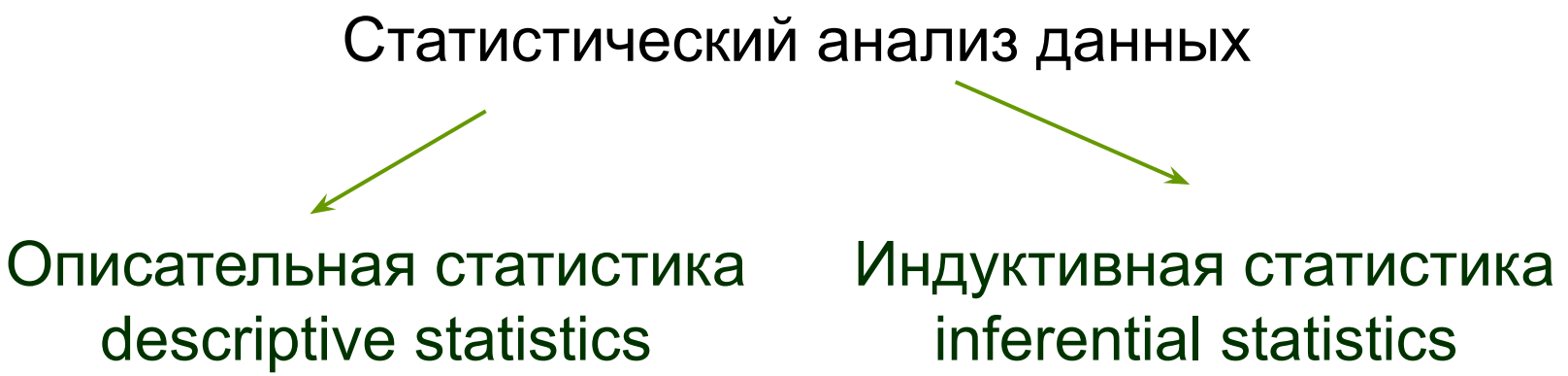
Основные понятия. Описательная статистика.

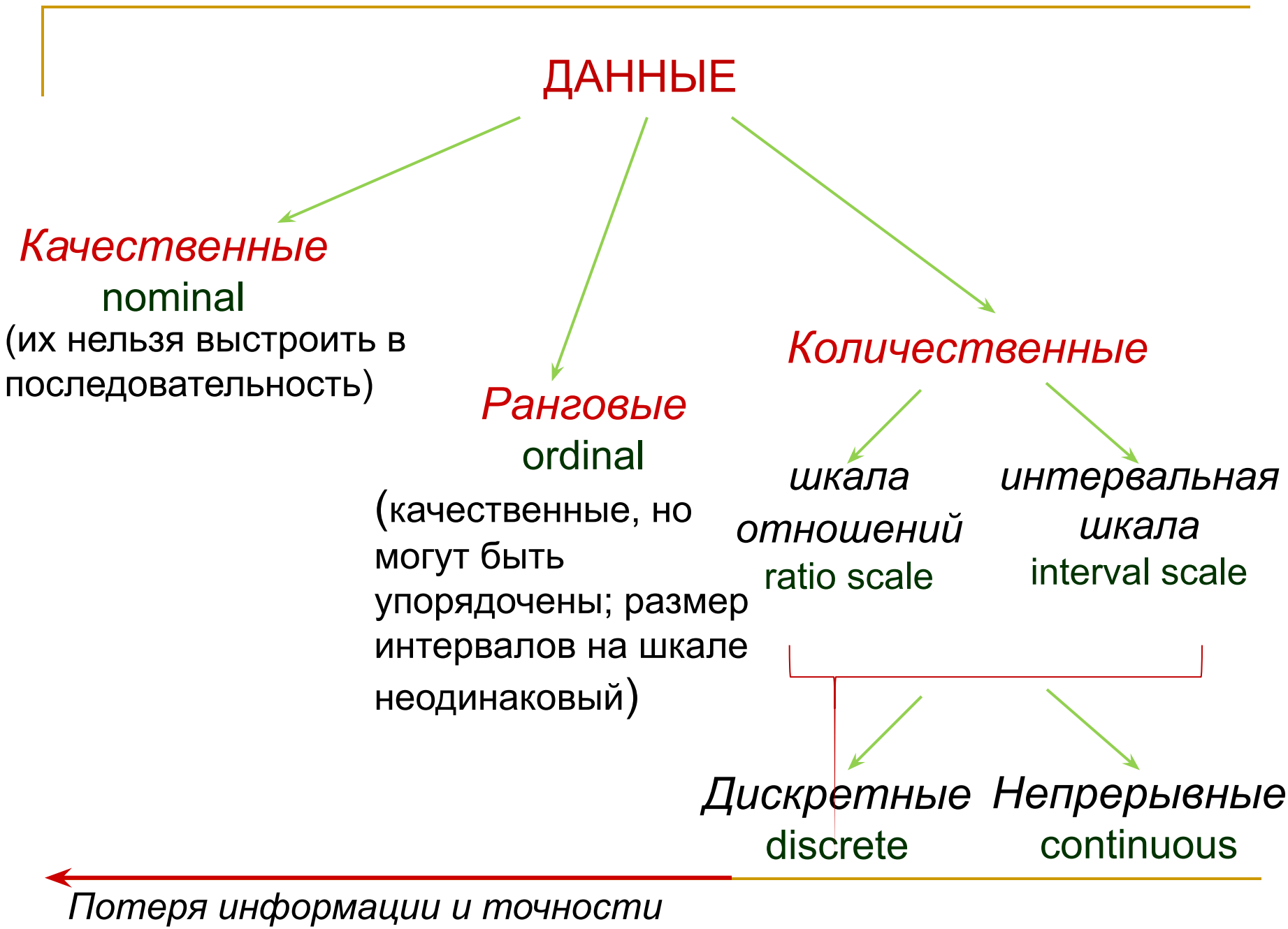


Данные – результаты некоторого количества измерений какой-либо ПЕРЕМЕННОЙ (переменных) – variable. Например:

- вес, длина тела, пол, окрас, температура .....

Статистика – инструмент для количественного анализа и интерпретации данных




ДАННЫЕ
Качественные
nominal
(их нельзя выстроить в последовательность)
Ранговые
ordinal
(качественные, но могут быть упорядочены; размер интервалов на шкале неодинаковый)
Количественные
шкала отношений
ratio scale
интервальная шкала
interval scale
Дискретные
discrete
Непрерывные
continuous
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, температура по Кельвину

интервальная шкала (interval scale):

- размер интервалов на протяжении всей шкалы одинаковый;
- положение нулевой точки выбрано произвольно.

Примеры: температура по Цельсию, время дня, дата

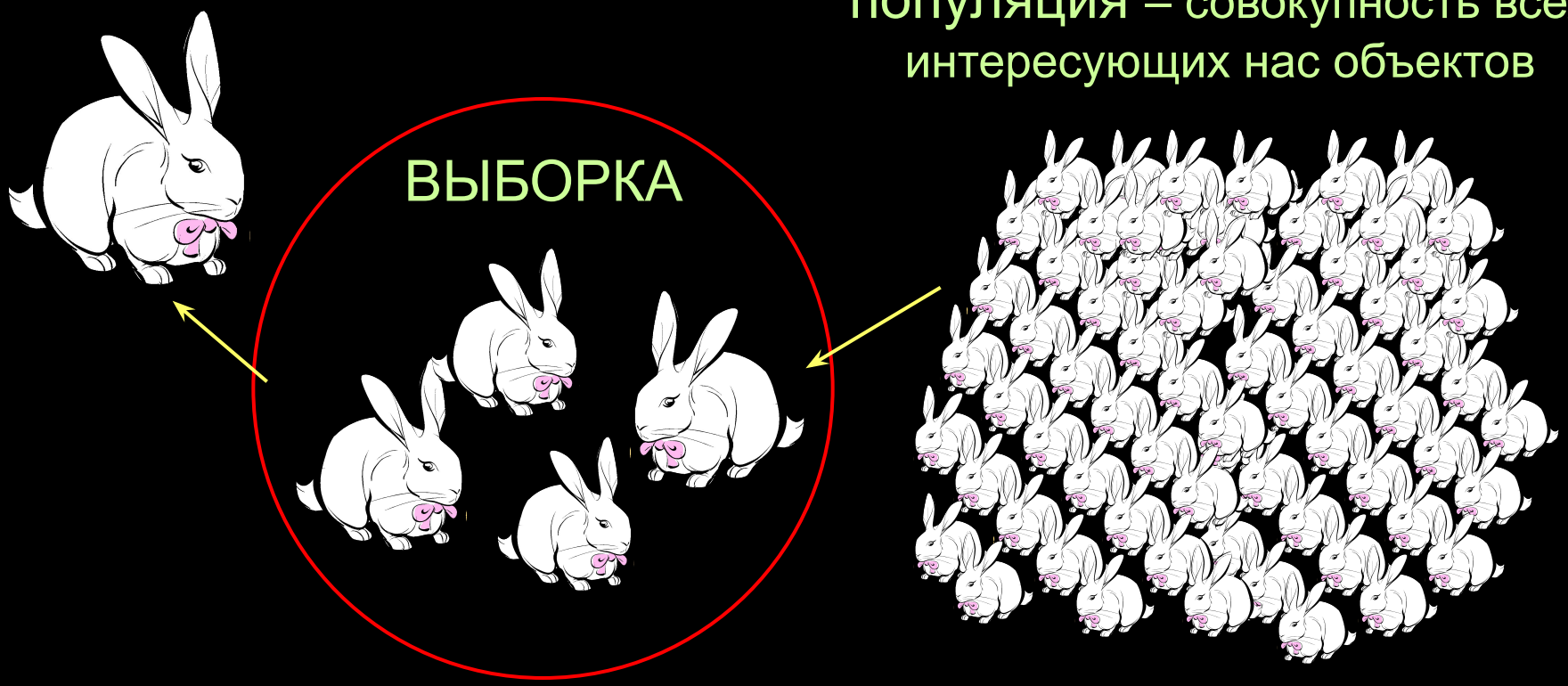
Непрерывные переменные:

1. Не нужно писать много знаков после запятой – количество знаков показывает точность измерения (= ошибку измерения)
2. Если почему-то необходимо округлить числа, чётные округляют в меньшую сторону, нечётные – в большую (2.5 в 2, 3.5 в 4);

рост, вес Ани, Тани и Мани

наблюдение

популяция – совокупность всех интересующих нас объектов



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

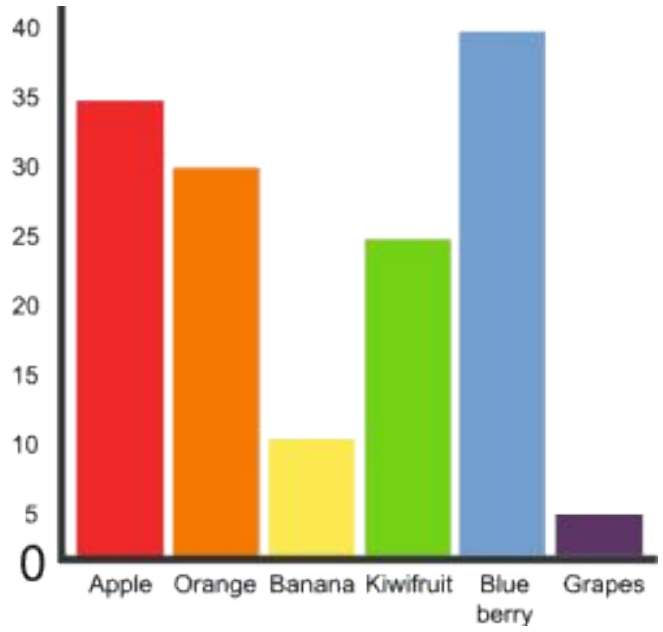
Три основные концепции в анализе данных:

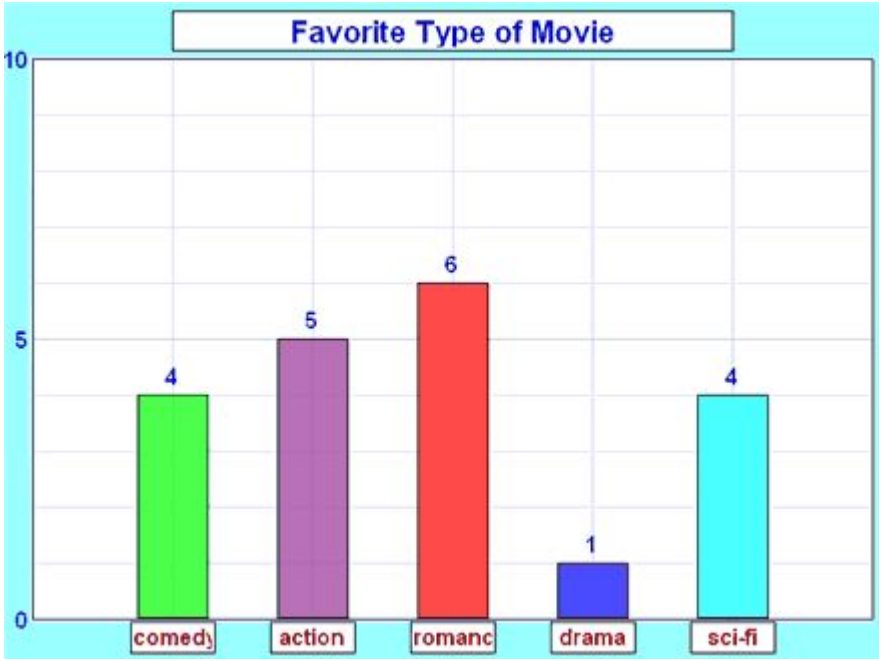
1. Что такое РАСПРЕДЕЛЕНИЕ переменной и как его описывать
2. Что такое распределение ВЫБОРОЧНЫХ СРЕДНИХ и как оно связано с распределением переменной
3. Что такое СТАТИСТИКА КРИТЕРИЯ

Необходимо для обдумывания и обсуждения данных

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.



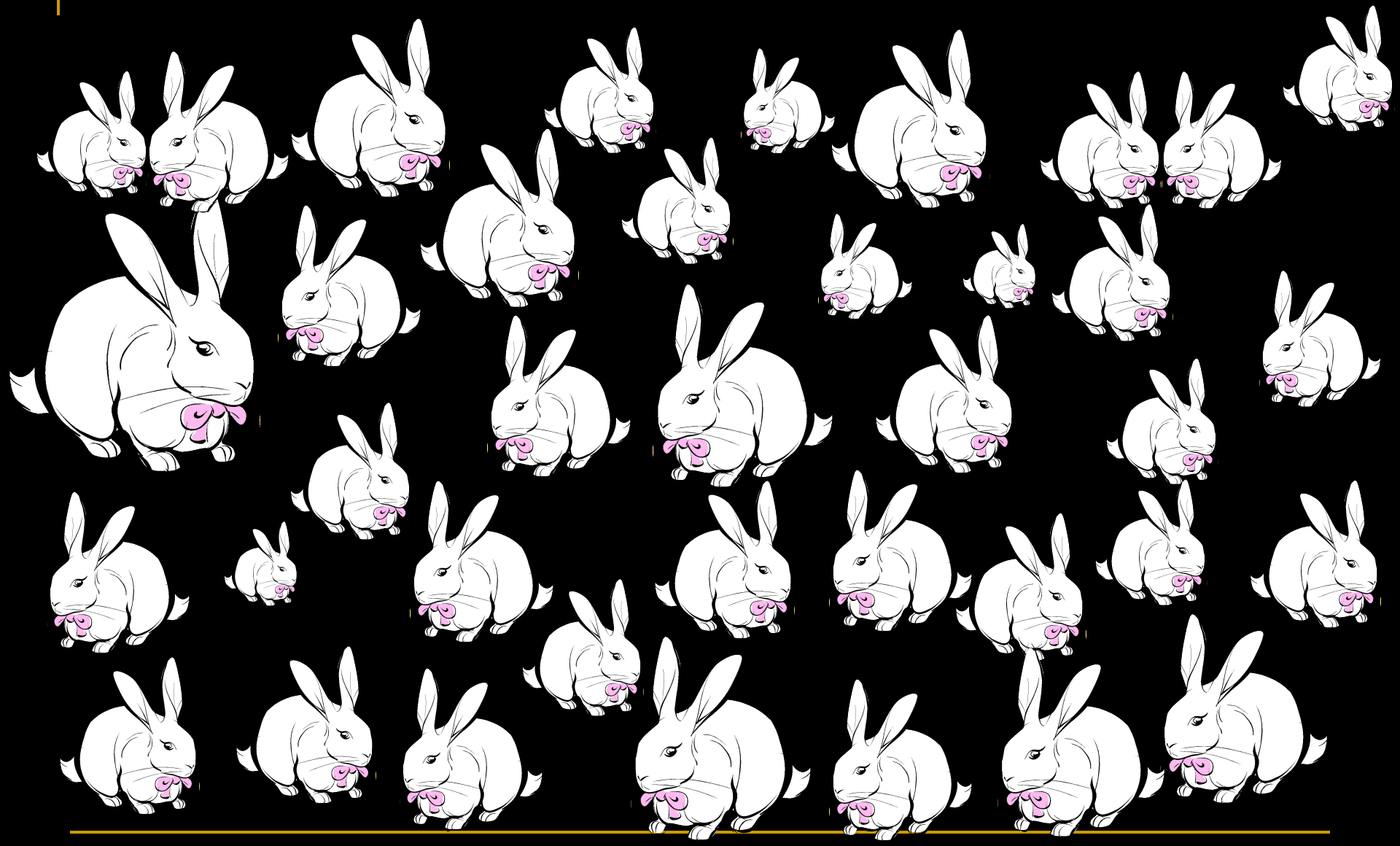


Частотное распределение переменной (frequency distribution)

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

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

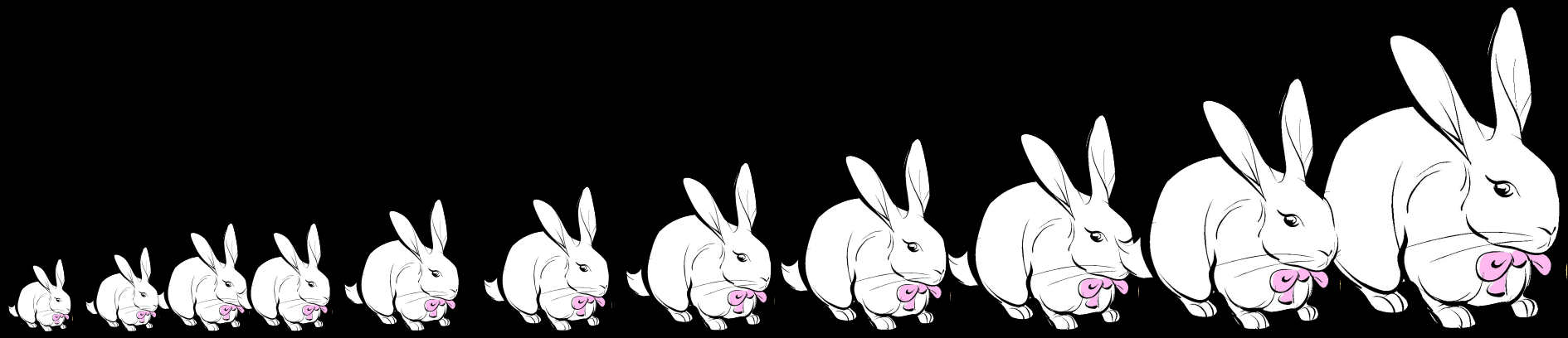
Частотное распределение переменной (frequency distribution)



Взвешиваем ***N*** кроликов

Частотное распределение переменной (frequency distribution)

1. Упорядочим по возрастанию значения переменной (выстроим кроликов от меньшего к большему);
2. разобьём их на группы по равным интервалам.



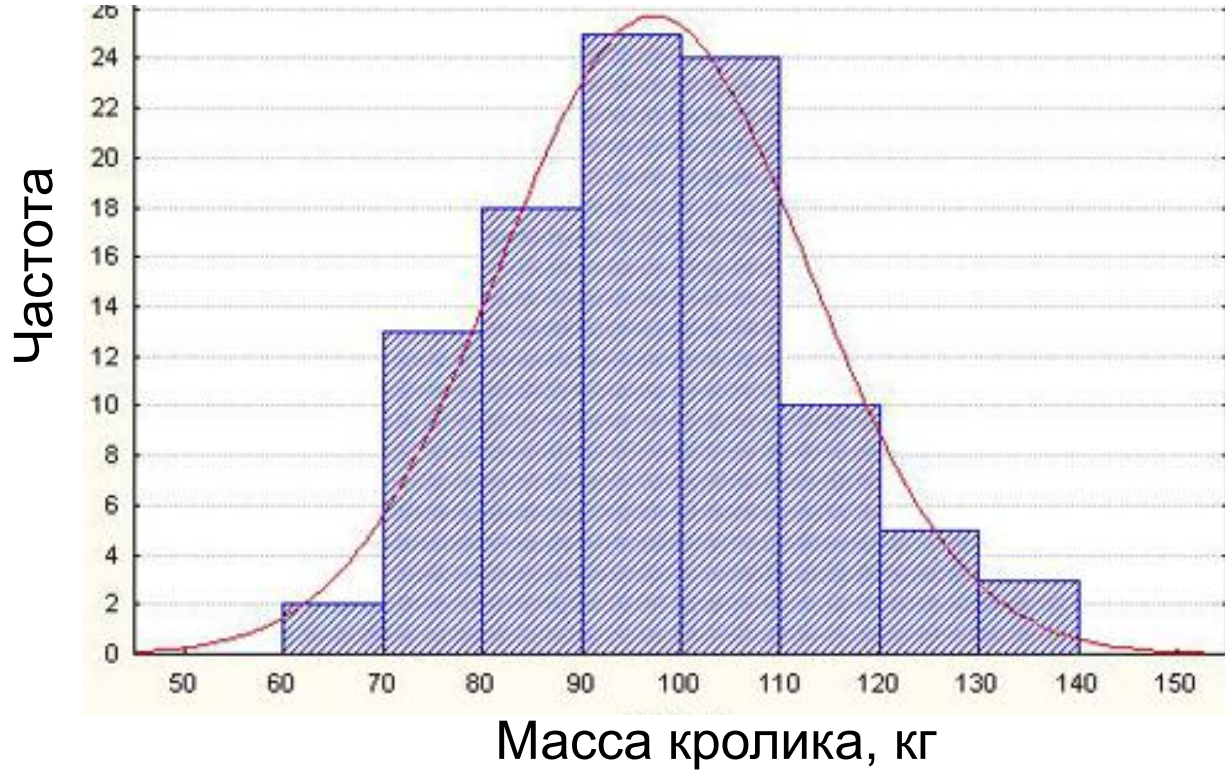
Частотное распределение переменной (frequency distribution)

Частота – то, сколько раз встретилось данное значение переменной

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

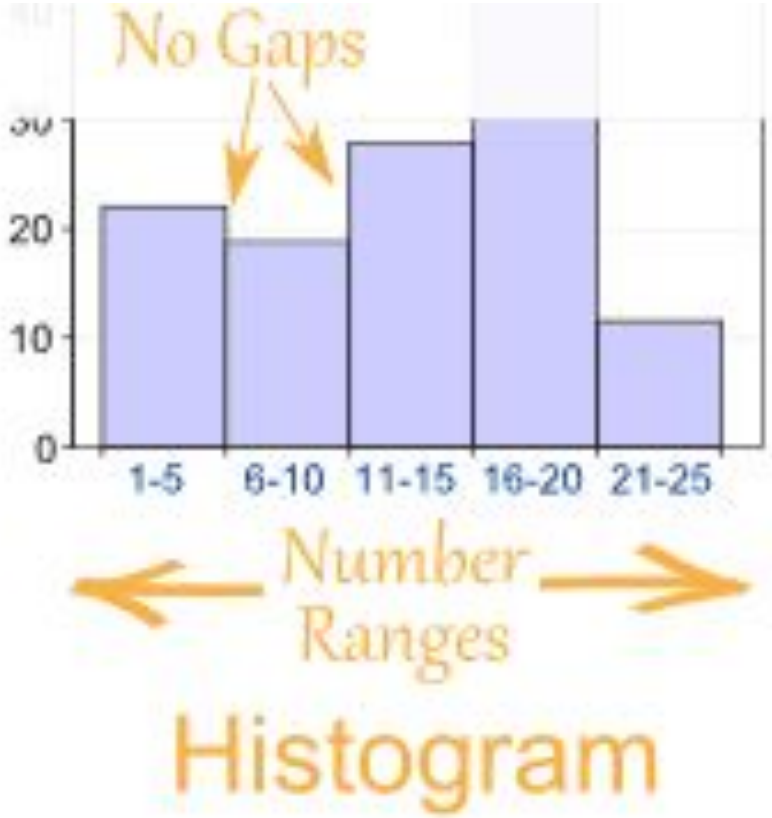
Интервалы должны быть:

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

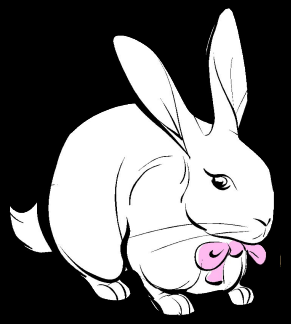


Полигон частот (frequency polygon)

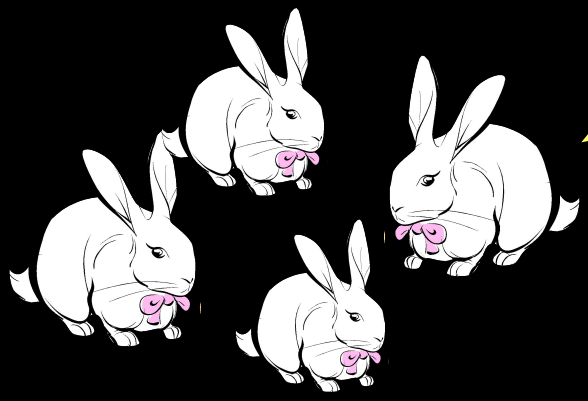
Частотное распределение переменной (frequency distribution)



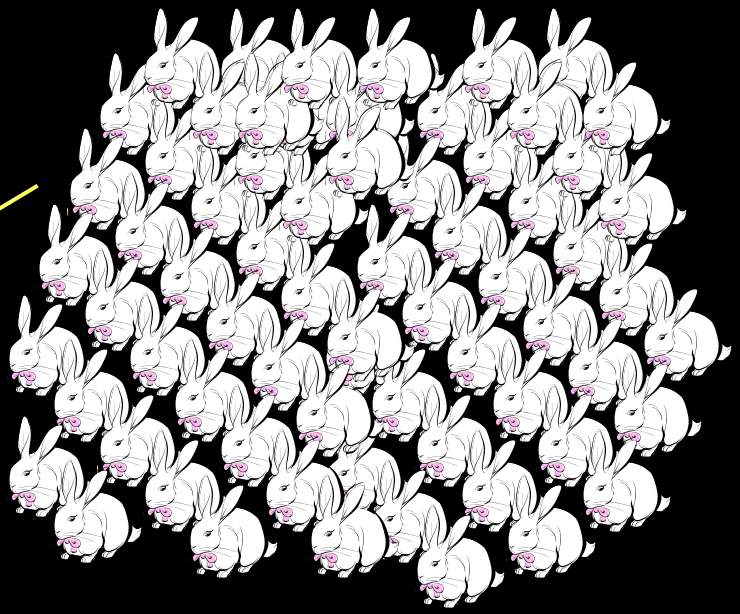
наблюдение



ВЫБОРКА



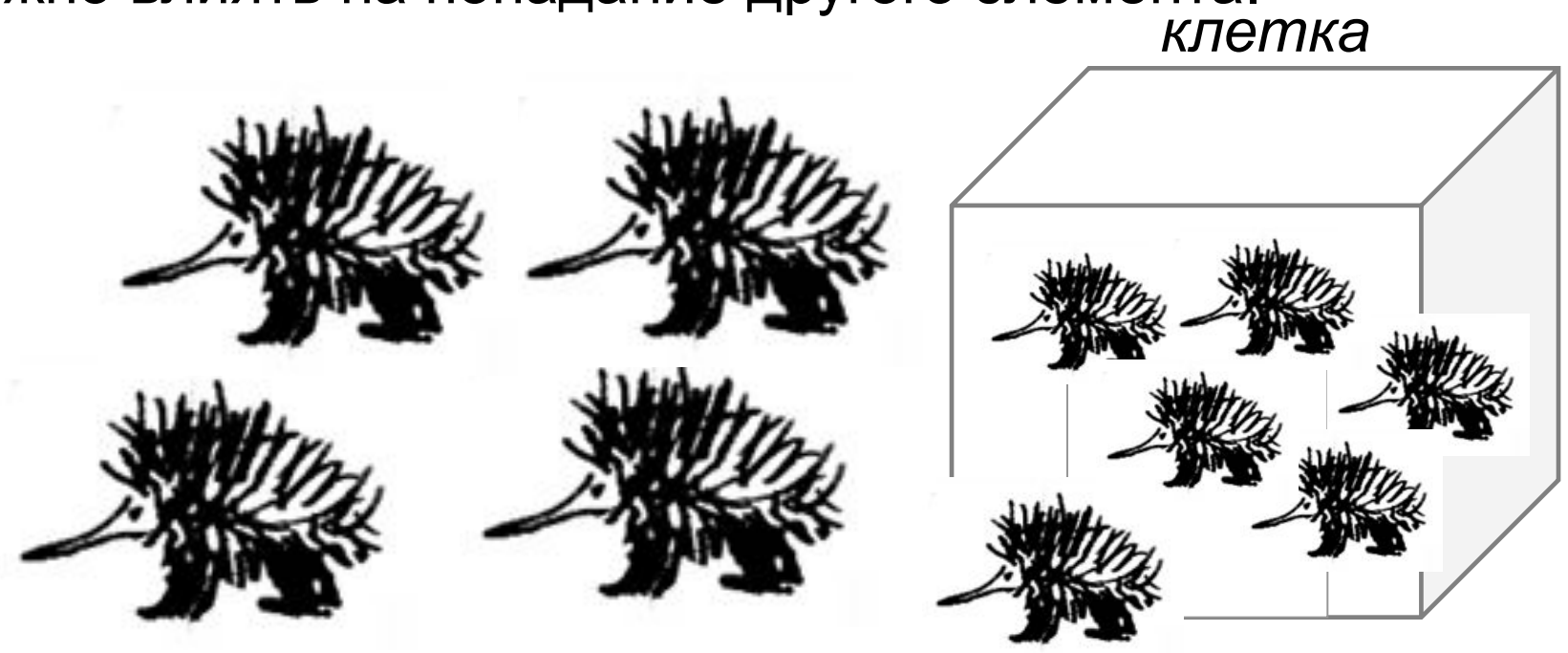
популяция – совокупность всех интересующих нас объектов



Популяция может быть воображаемой (гипотетической).

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.



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

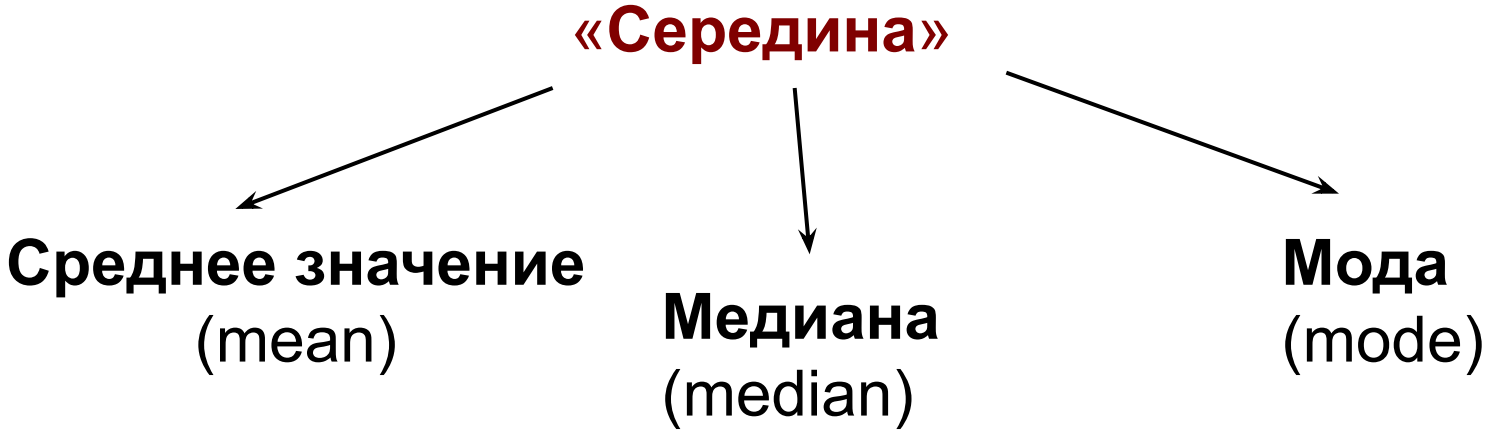
Как описать частотное распределение переменной?

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

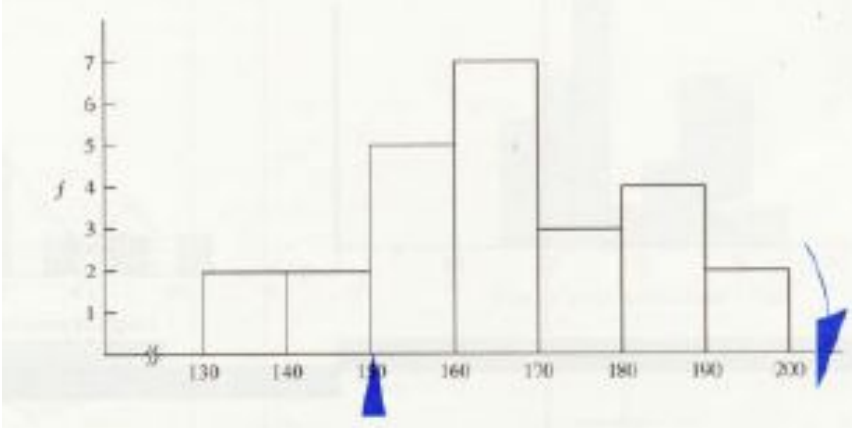


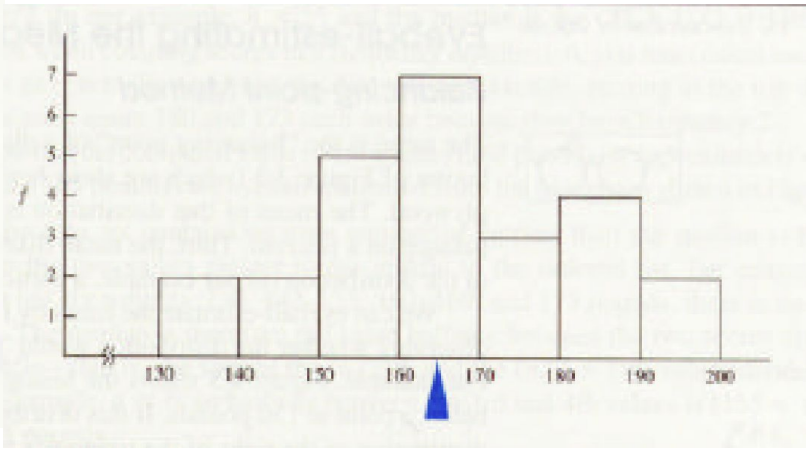
Все они могут служить оценками популяционного среднего.
Среднее в выборке – наиболее эффективная и несмещённая оценка.

Частотное распределение переменной (frequency distribution)

«**Середина**» распределения

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**

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

Частотное распределение переменной (frequency distribution)

«**Середина**» распределения

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 ранговых! (не для качественных).

Частотное распределение переменной (frequency distribution)

- ✔ 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).

Частотное распределение переменной (frequency distribution)

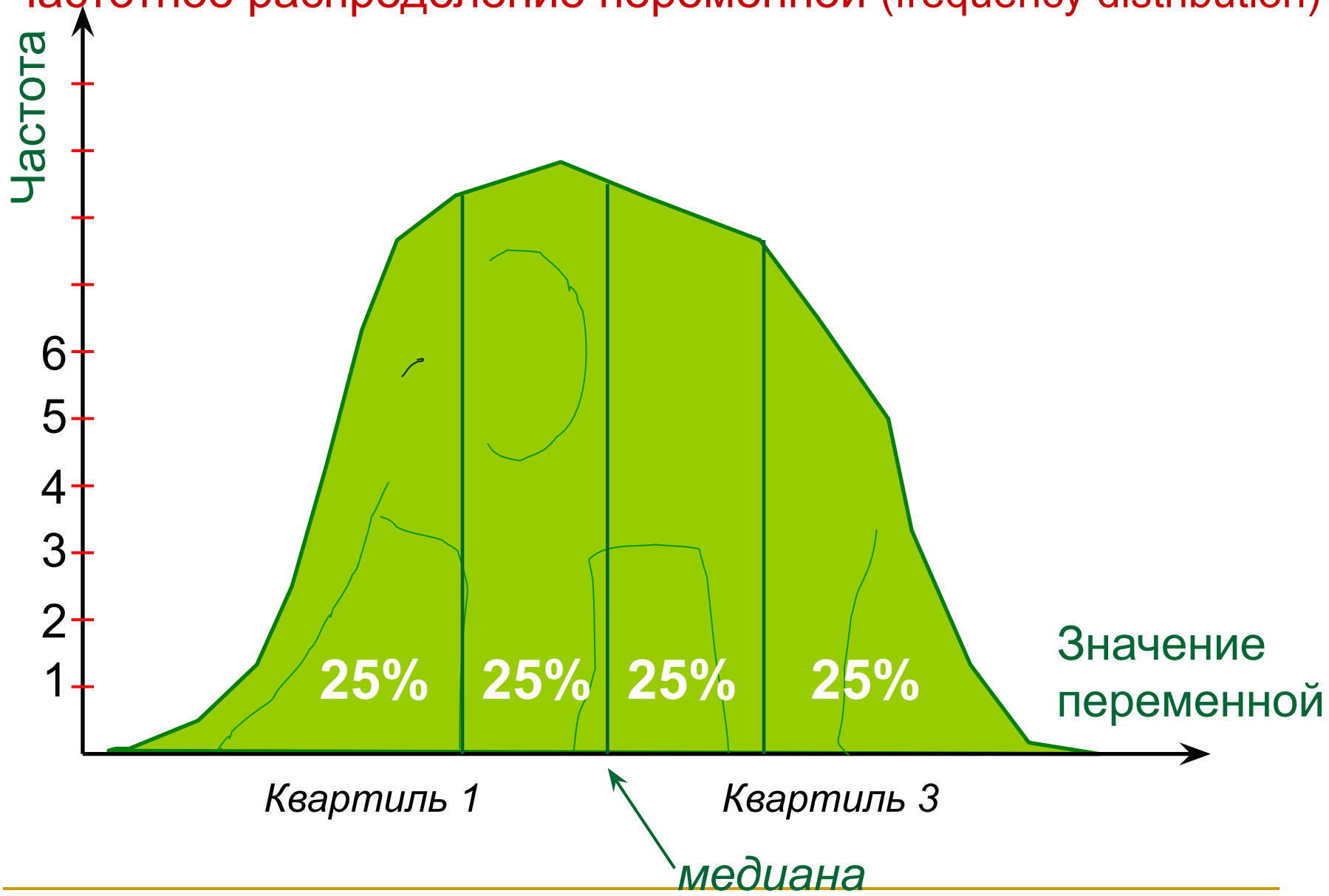
Квартили (quartiles) делят распределение на четыре части так, что в каждой из них оказывается поровну значений (2-я квартиль = медиана).

1-я квартиль = 25% процентиль

3-я квартиль = 75% процентиль

Интерквартильный размах – разница между третьей и первой квартилями.

Частотное распределение переменной (frequency distribution)

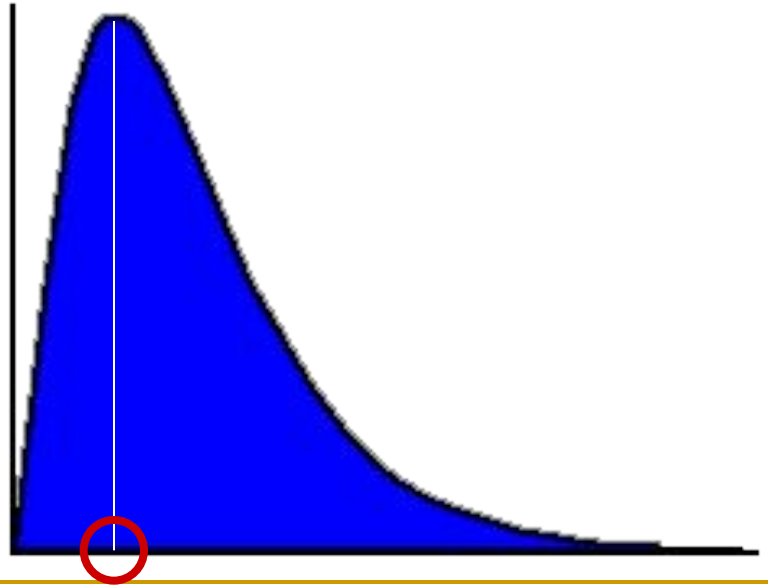


Частотное распределение переменной (frequency distribution)

«**Середина**» распределения

Мода (mode) – наиболее часто встречающееся значение

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



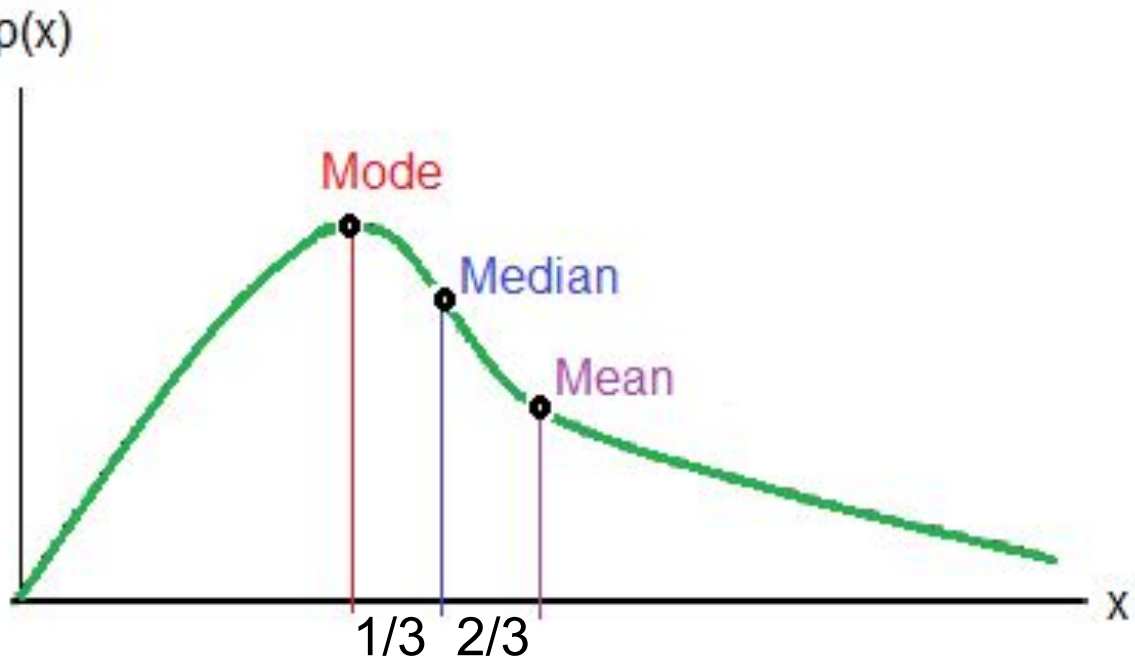
В первую очередь биолога интересует количество мод в распределении, а не мода как таковая

Частотное распределение переменной (frequency distribution)

«**Середина**» распределения

Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения

ЗАРПЛАТА, $	ЧАСТОТА
200000	1
20000	1
19000	1
14000	3



К появлению перекоса чувствительнее всего среднее значение

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

Частотное распределение переменной (frequency distribution)

Разброс распределения

Стандартное отклонение (standard deviation)

Для **выборки**:

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

Для популяции:

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

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

Сумма квадратов (sum of squares = SS)

Стандартное отклонение зависит ото всех значений переменной.

Измеряется в тех же единицах, что и переменная!

Частотное распределение переменной (frequency distribution)
Разброс распределения

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, возведённых в квадрат (что не всегда удобно).

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

Частотное распределение переменной (frequency distribution)
Разброс распределения

Коэффициент вариации
(Coefficient of variation)

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

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.)

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

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

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

$J = \frac{H}{\log k}$ Нормированный индекс Шеннона ($\in [0;1]$)

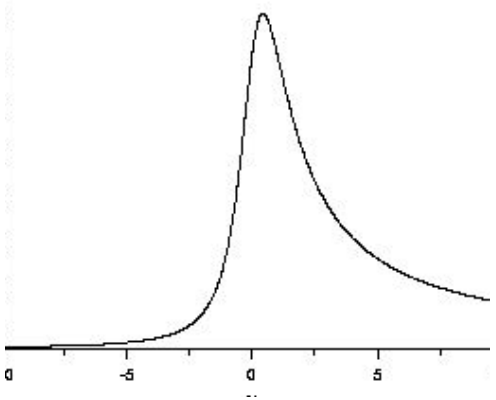
Этих индексов много для разных целей; это показатели ОПИСАТЕЛЬНОЙ статистики!

Частотное распределение переменной (frequency distribution)

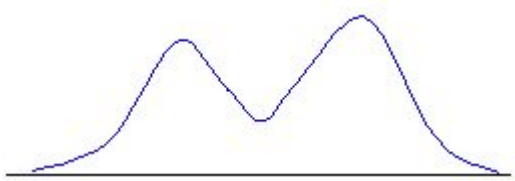
Как описать непрерывное распределение?

1. По количеству «максимумов» (мод):

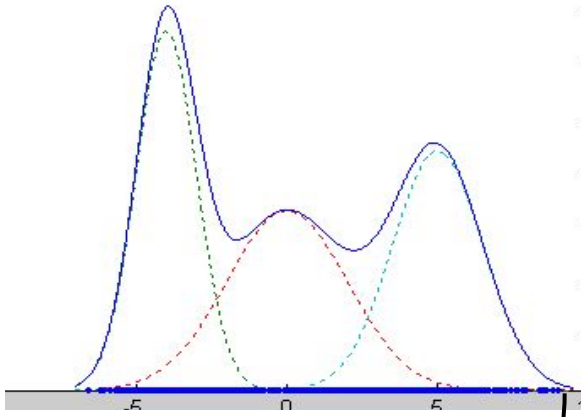
унимодальное



бимодальное



мультимодальное



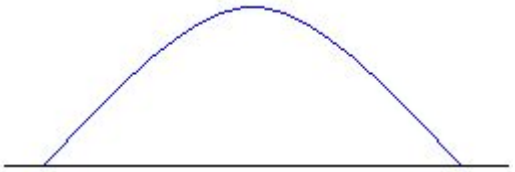
обычно возникают, если популяция имеет естественные обособленные подгруппы

Частотное распределение переменной (frequency distribution)

Как описать непрерывное распределение?

2. По признаку симметрии:

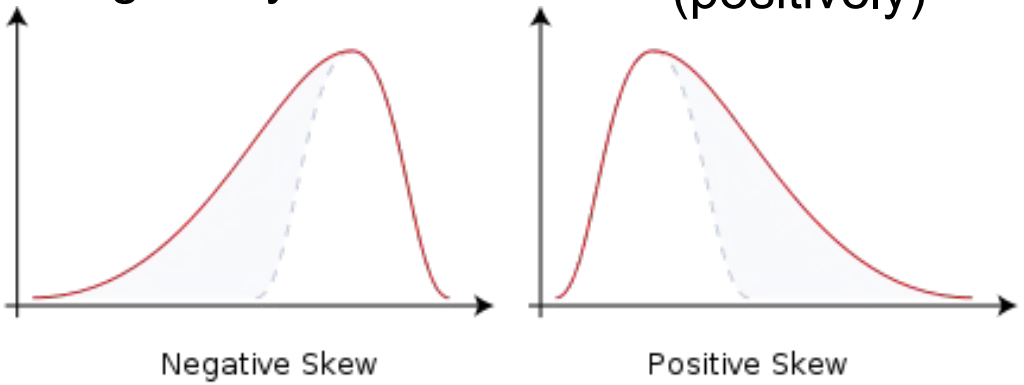
Симметричное



Скошенное (skewed)

влево
negatively

вправо
(positively)



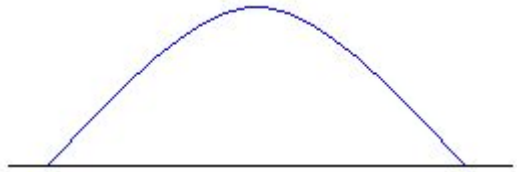
Частотное распределение переменной (frequency distribution)

Как описать непрерывное распределение?

3. распределение

асимптотическое

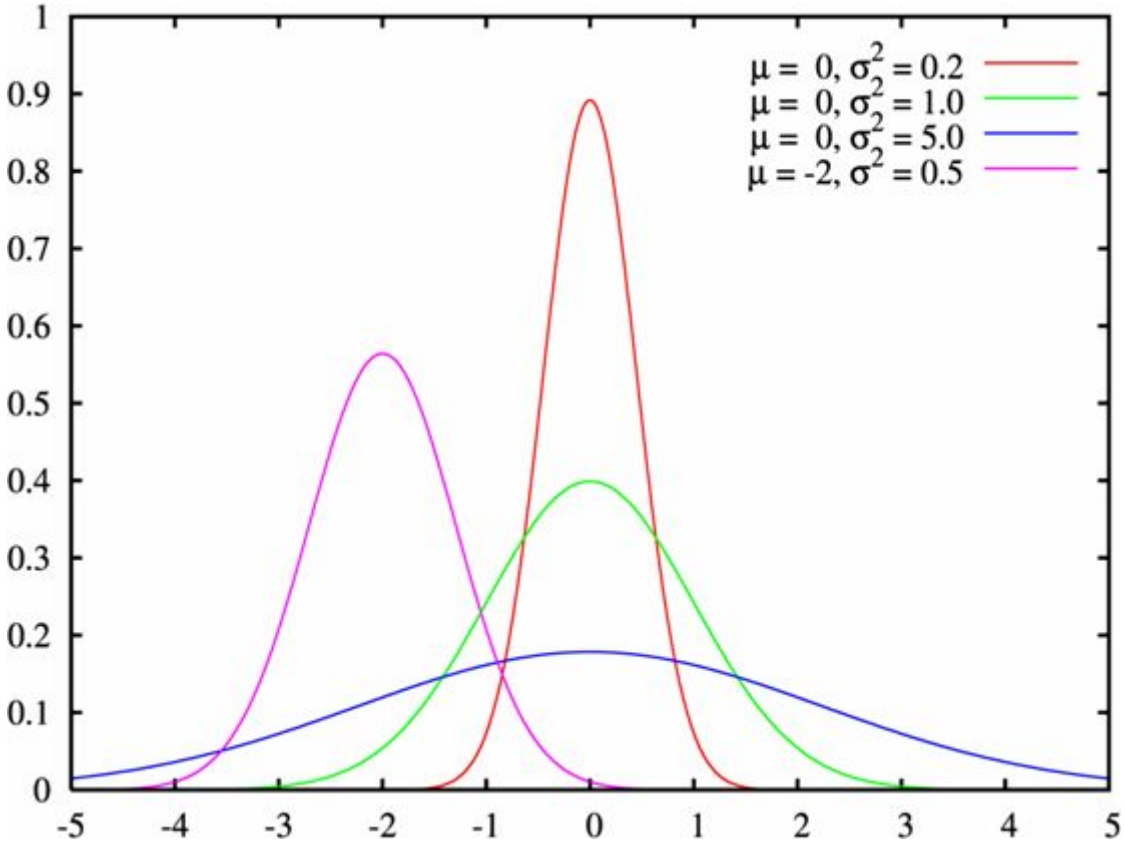
не асимптотическое



Частотное распределение переменной (frequency distribution)

Нормальное распределение (Гауссово):
первое знакомство

- ✔ Унимодальное
- ✔ Симметричное
- ✔ Асимптотическое



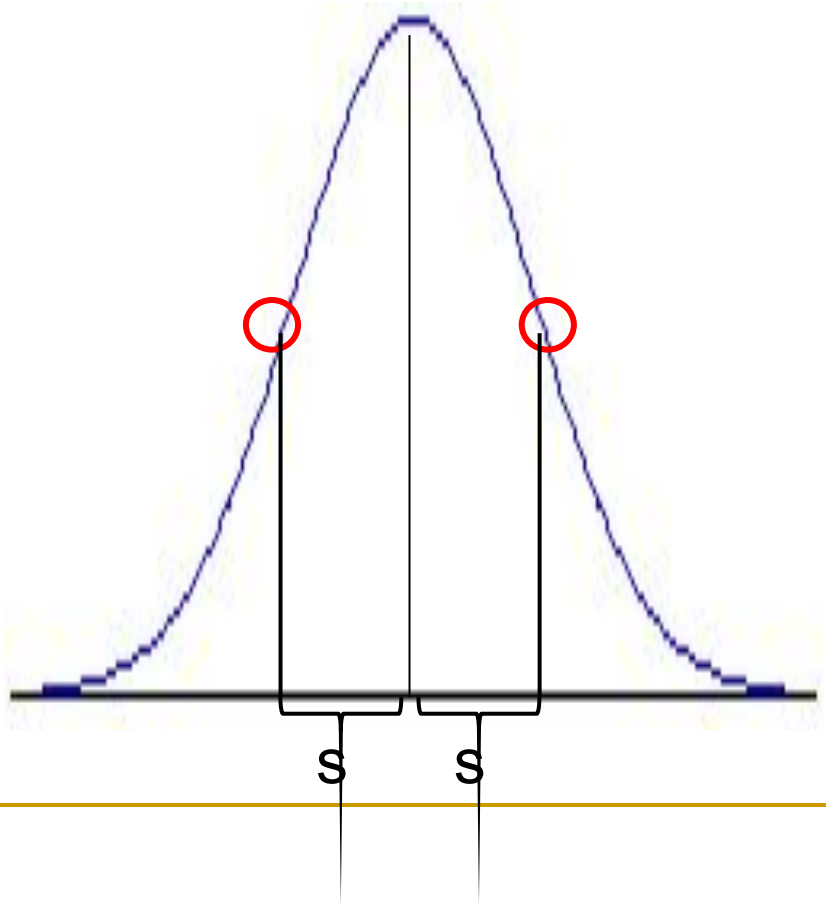
Это
<u>непрерывное</u>
распределение

Высота деревьев, масса тела новорожденных, IQ, скорость прохождения лабиринта крысами и многие, многие другие переменные

Название в честь Гаусса не совсем справедливо – первым его описал вовсе не он.
Симметрия и эксцесс.

Частотное распределение переменной (frequency distribution)
Разброс распределения

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

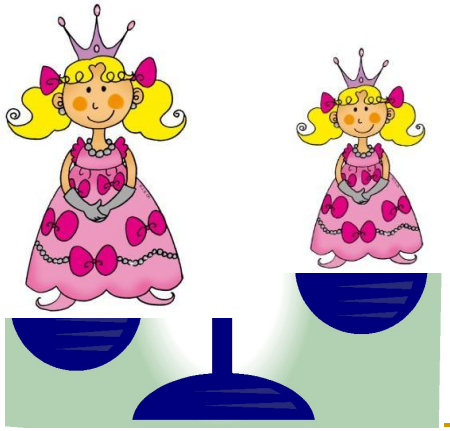


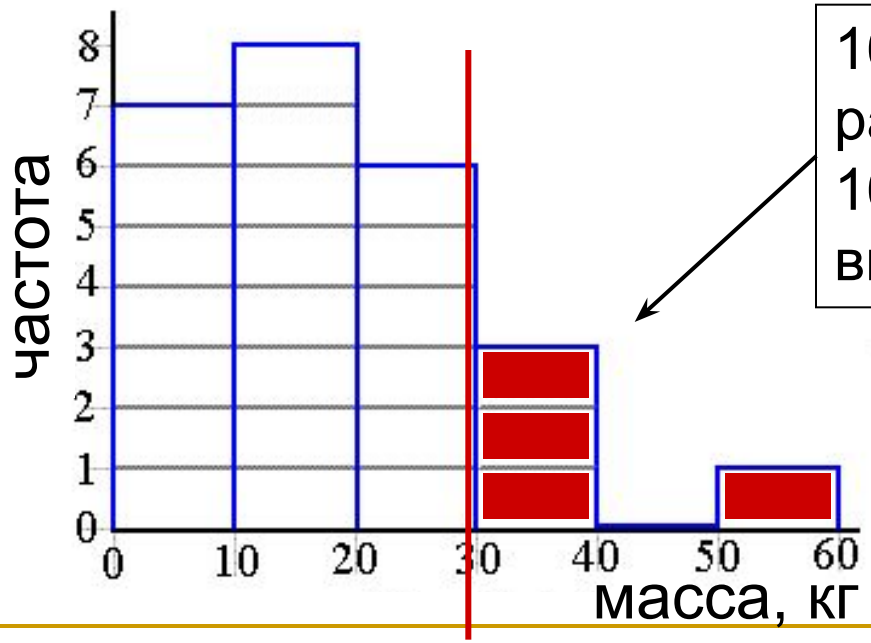
Частотное распределение переменной (frequency distribution)

«Площадь распределения»

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

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



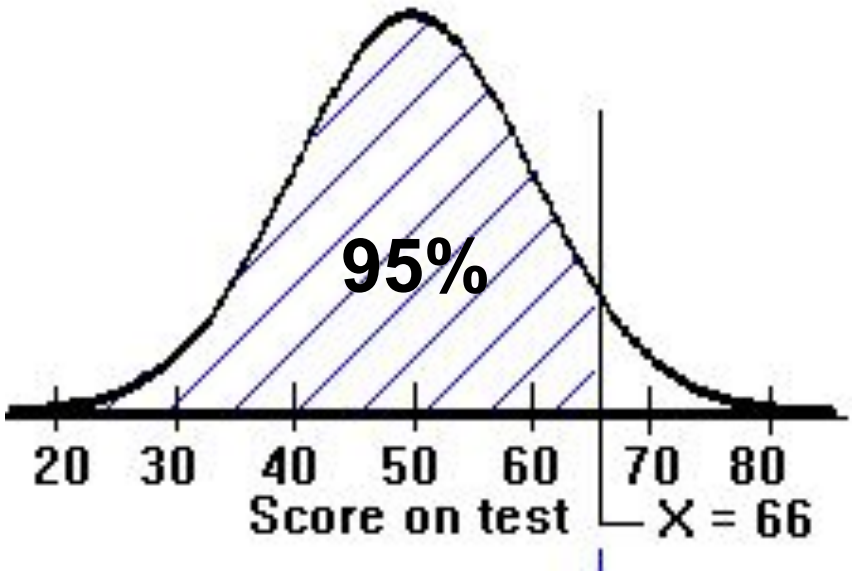


16% площади распределения ~ 16% объёма выборки

Частотное распределение переменной (frequency distribution)

Процентили и z-оценка

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

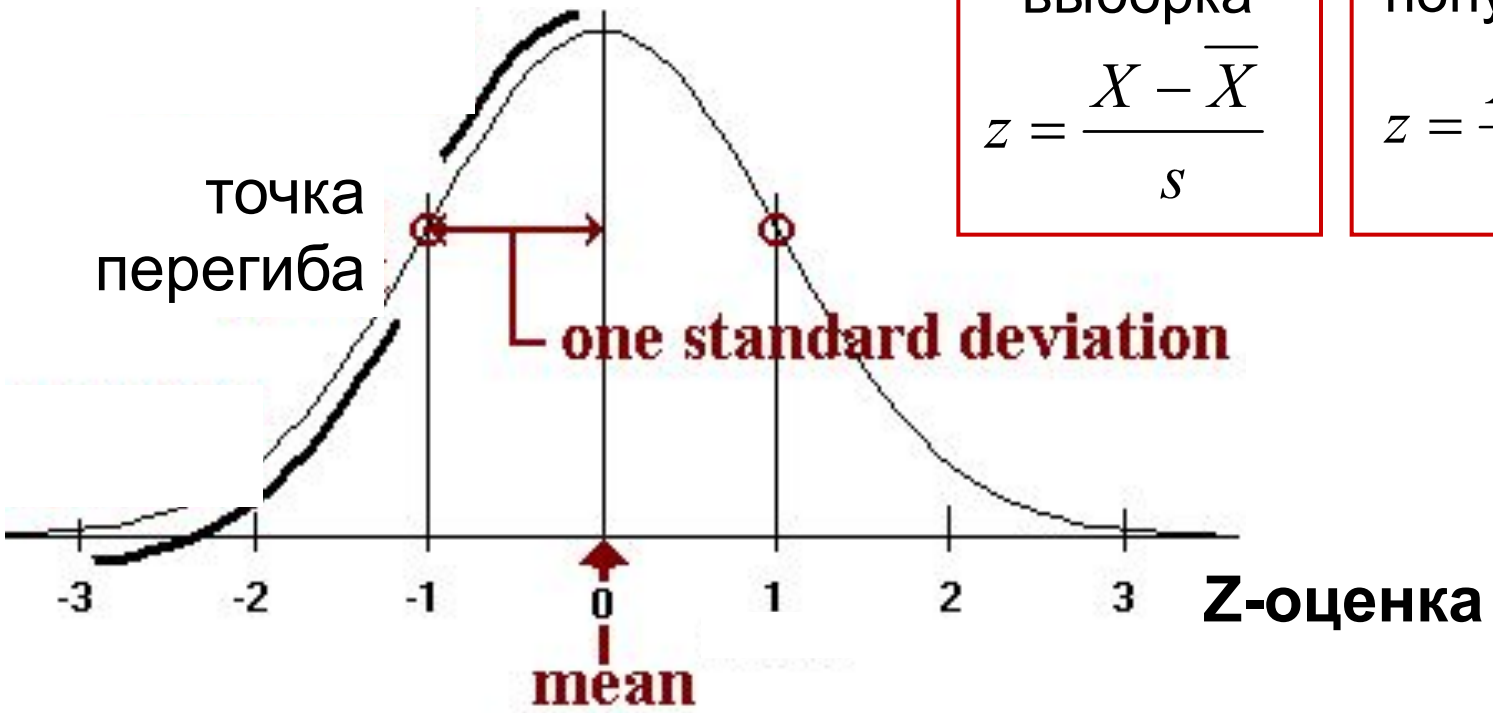


Частотное распределение переменной (frequency distribution)

Процентили и z-оценка

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

выборка	популяция
$z = \frac{X - \overline{X}}{s}$	$z = \frac{X - \mu}{\sigma}$



Площадь нормального распределения

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

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

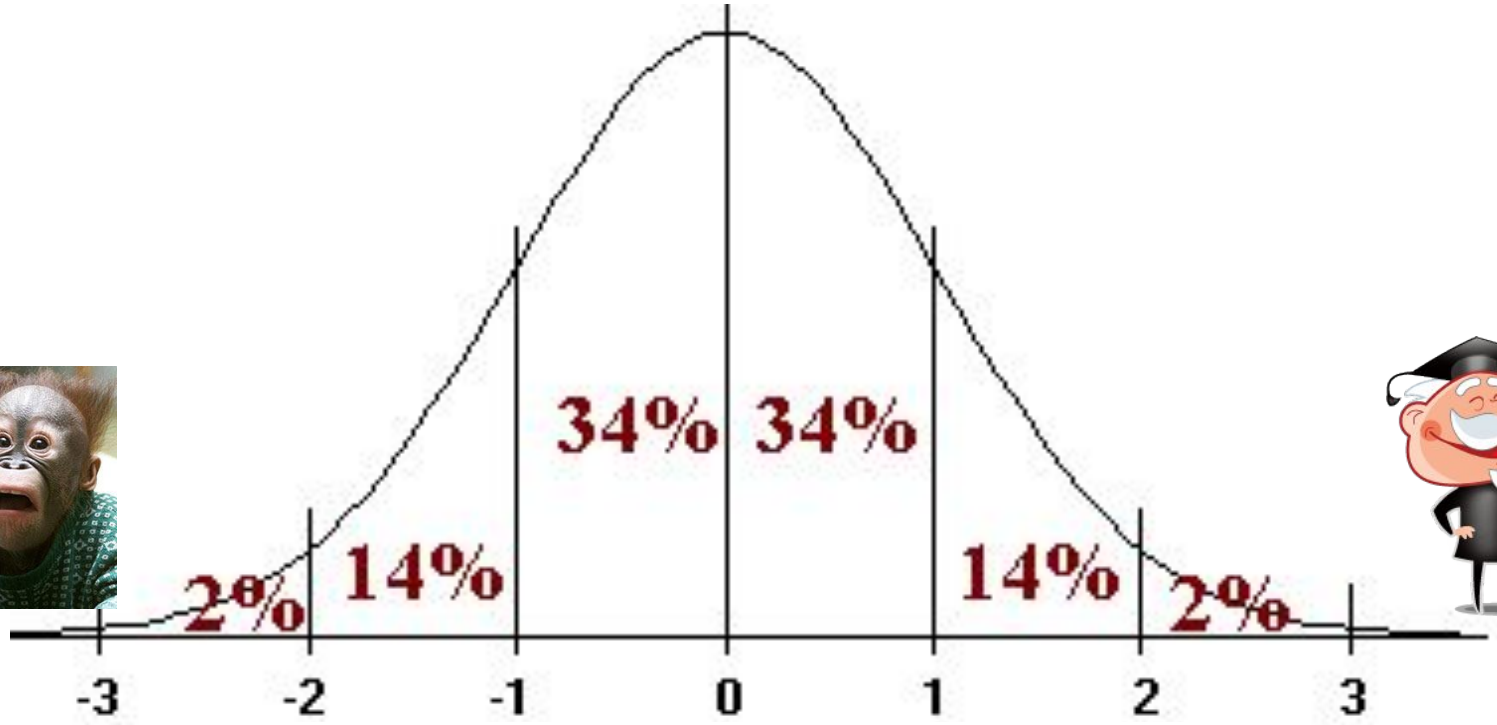
Необыкновенное свойство:

Относительные площади под участками нормального распределения всегда одинаковы!

Частотное распределение переменной (frequency distribution)

Площадь нормального распределения

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



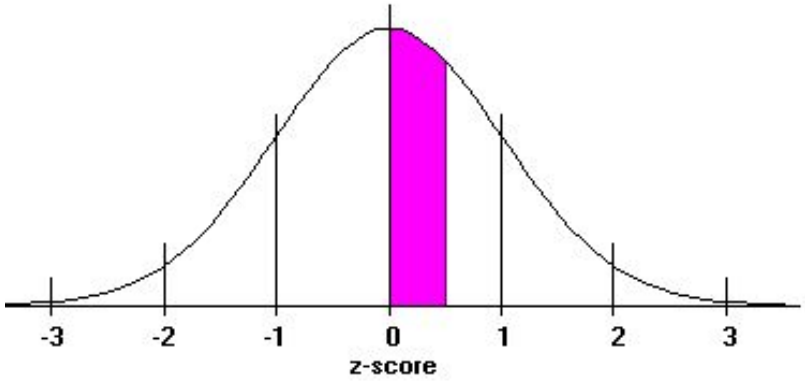
Z-оценка

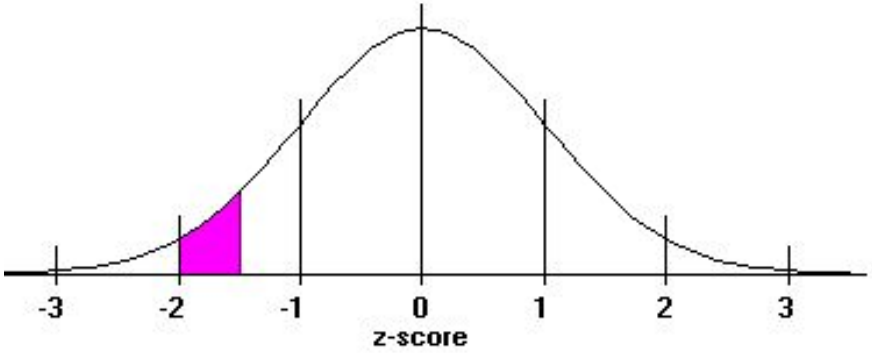
(количество стандартных отклонений)

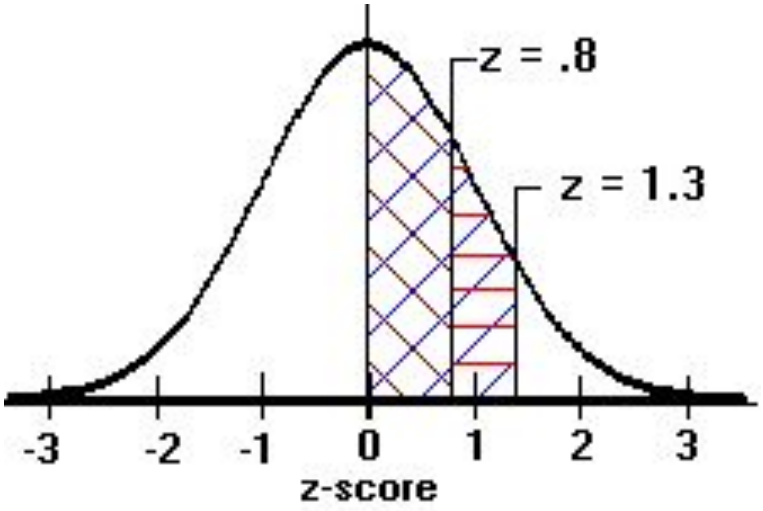
Пример с IQ (μ=100, σ=15)

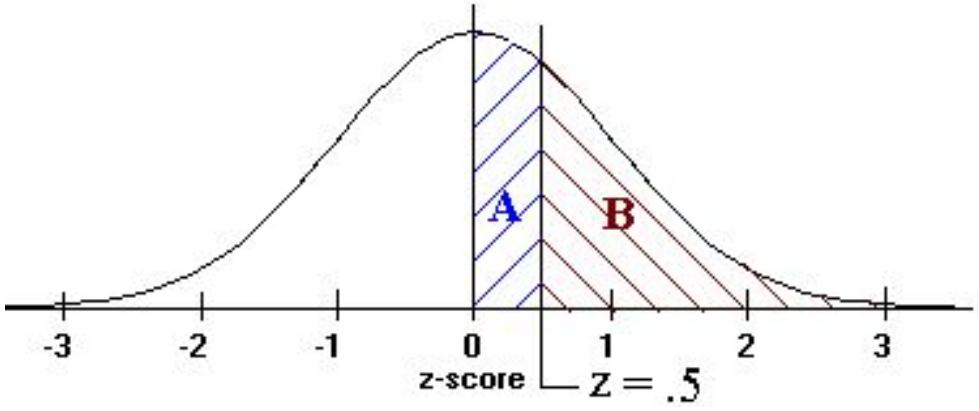
Частотное распределение переменной (frequency distribution)

Площадь нормального распределения

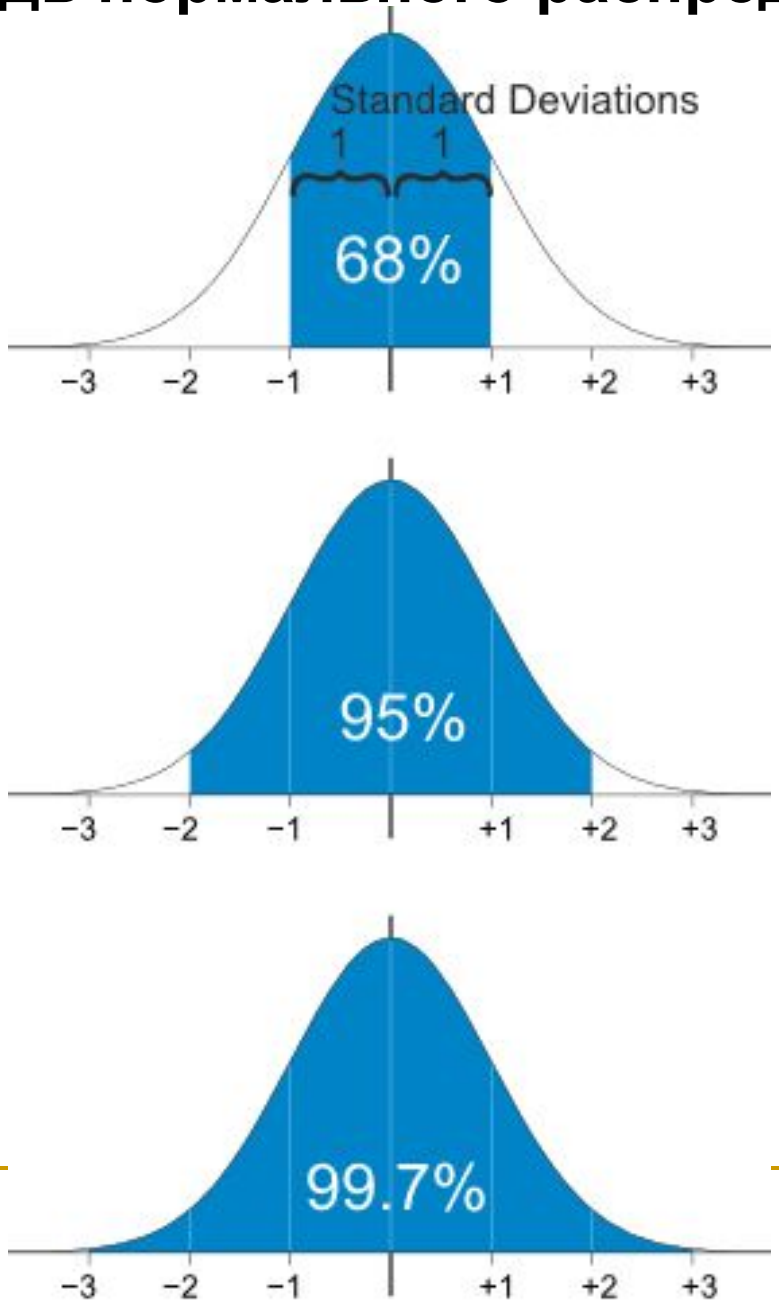








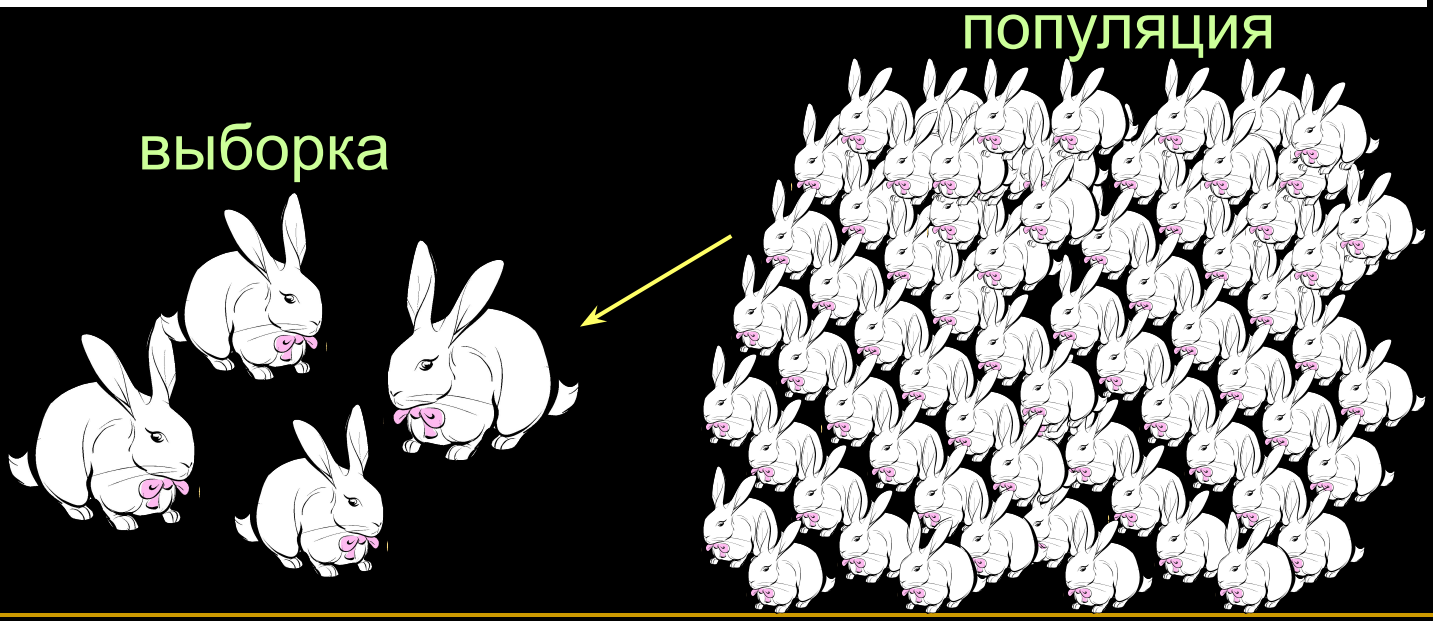
Площадь нормального распределения



Распределение выборочных средних (sampling distribution of the means)

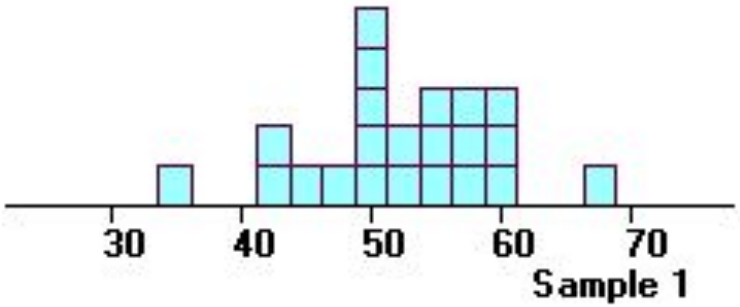
Три основные концепции в анализе данных:

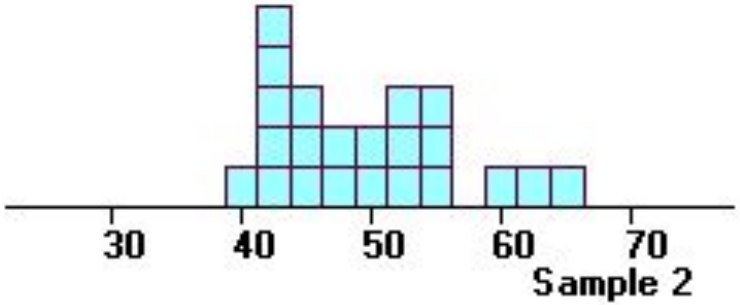
1. Что такое РАСПРЕДЕЛЕНИЕ переменной и как его описывать
2. Что такое распределение ВЫБОРОЧНЫХ СРЕДНИХ и как оно связано с распределением переменной
3. Что такое СТАТИСТИКА КРИТЕРИЯ



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.

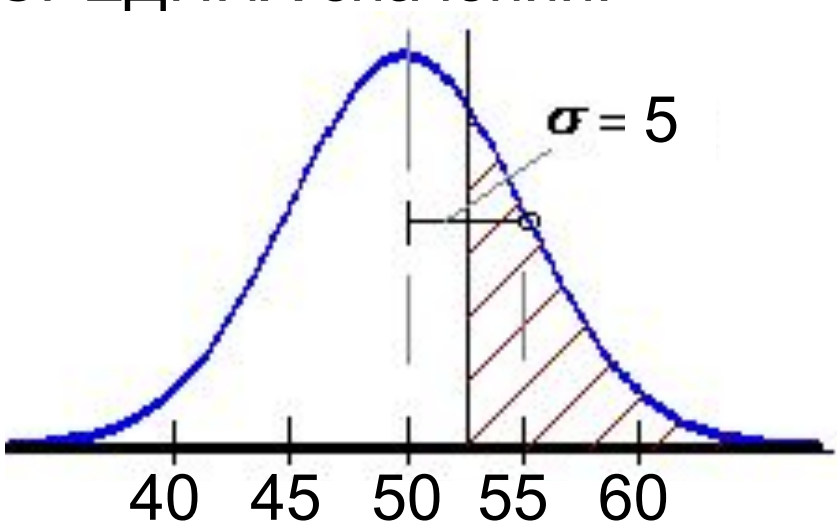
Средняя масса студента – μ=50 кг, σ = 4 кг.

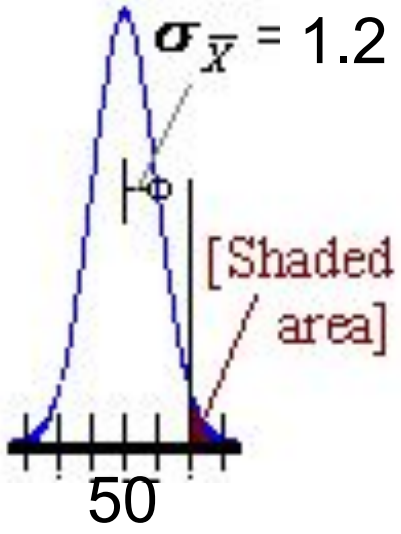
Посчитаем средние массы для каждой группы!

Форма распределений маленьких выборок не обязательна 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)

Пример про бутылки с кока-колой

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)

Распределение выборочных средних (sampling distribution of the means)

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

Определяет форму, среднее и разброс в распределении выборочных средних

- **Форма:** с увеличение 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.

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

Пример с монеткой

Распределение выборочных средних (sampling distribution of the means)

Следствие:

если некоторая величина отклоняется от среднего под воздействием слабых, независимых друг от друга факторов, она имеет нормальное распределение. Поэтому оно так широко распространено в природе!

Пример про высоту деревьев в лесу

Распределение выборочных средних (sampling distribution of the means)

У нас есть только одна выборка. Из неё мы получили среднее значение $\overline{X}$

Насколько оно близко среднему значению в популяции (μ)?

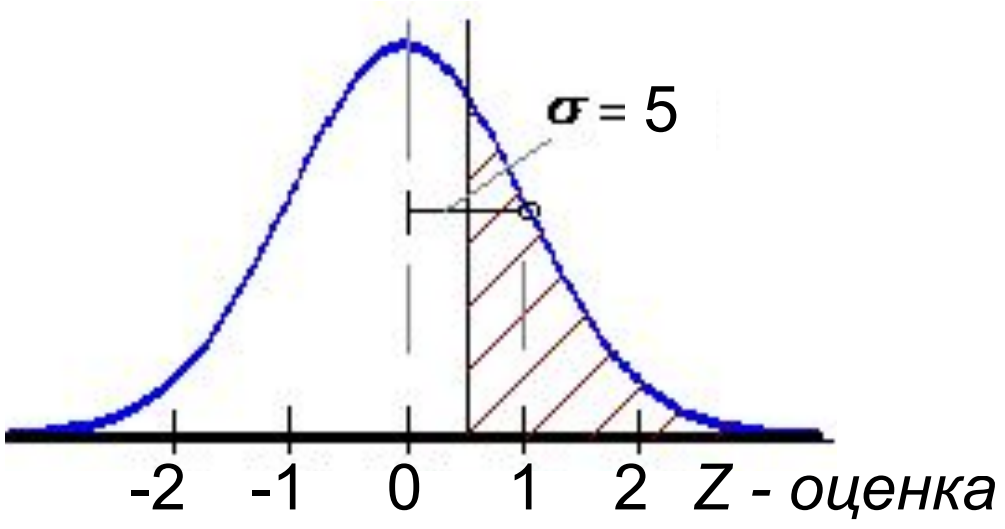
Решим обратную задачу. Пусть нам известно μ, найдём $\overline{X}$

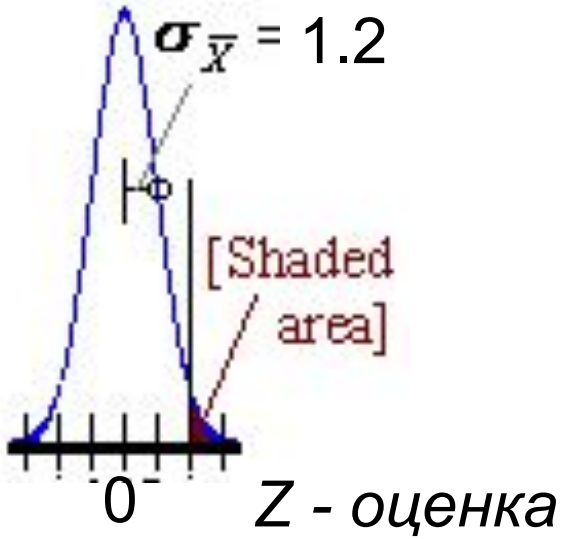
Мы знаем, что для нормального распределения есть **z-оценка**, значениям которой соответствуют **определённые площади** распределения.

Но мы также знаем, что **выборочные средние** образуют **нормальное** распределение!!

Это значит, что, зная среднее в популяции, мы можем предсказать (с … вероятностью) интервал, в который попадёт выборочное среднее.

Распределение выборочных средних (sampling distribution of the means)





Вопрос: какая часть ОСОБЕЙ имеет массу больше 55 кг?

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

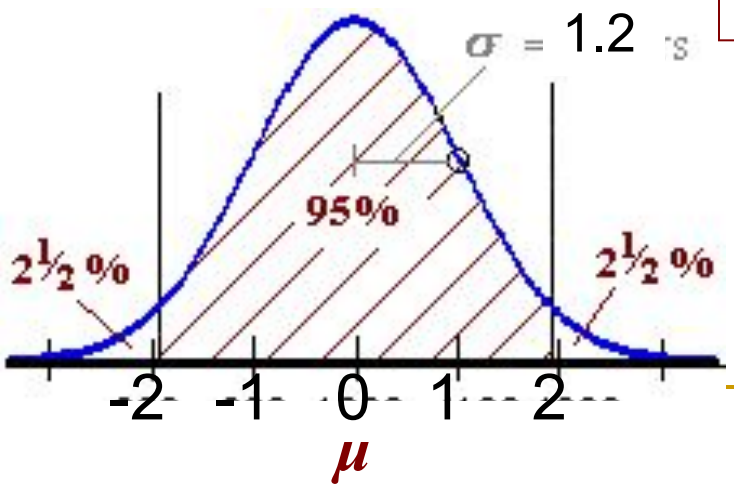
Другой вопрос: какая часть ВЫБОРОК имеет СРЕДНЮЮ массу больше 55 кг?

Оценка **параметров популяции** на основе свойств выборки

Пусть мы <u>изначально знаем</u> среднюю массу студентов 1-го курса и стандартное отклонение в популяции. Как оценить среднюю массу в одной из групп?

Построим распределение выборочных средних! Вспомним, что оно – нормальное, а его среднее значение соответствует среднему в популяции.

Зная стандартное отклонение в нем (=**SE!!**) можем рассчитать **интервал**, в который попадёт 95% (99%) всех средних масс в группах:



Оценка **параметров популяции** на основе свойств выборки

95% доверительный интервал (95% confidence interval): интервал значений переменной, который с вероятностью 95% содержит нужный параметр.

Т.е., расстояние от среднего значения в популяции до выборочного среднего для 95% выборок **не больше 1.96 SE**

Вернёмся к исходной задаче:
Как оценить среднюю массу в популяции, если нам известно среднее в выборке??

Расстояние от среднего в выборке до (неизвестного) среднего в популяции с вероятностью 95% **не больше 1.96 SE**

$z_{cv_{0.05}} = 1.96$

cv – critical value, критическое значение статистики (в данном случае, Z)

Оценка **параметров популяции** на основе свойств выборки

Вопрос: где расположено ***μ?***

Ответ: я точно не знаю, но наиболее вероятно – в пределах ± 2-х стандартных ошибок среднего (SE)

$$\overline{X} - z_{cv_{0.05}} SE < \mu < \overline{X} + z_{cv_{0.05}} SE$$

Чем больше уровень достоверности – 99%, 99,9%... (= доверительный уровень) тем ШИРЕ будет интервал

Вопрос: где расположено ***μ?***

Ответ: я совершенно уверен, что оно лежит в пределах... от $-\infty$ до $+\infty$

В примере нам было известно ***σ***, но на практике оно обычно неизвестно!

Оценка **параметров популяции** на основе свойств выборки

Мы не знаем стандартное отклонение в популяции, и оцениваем его через стандартное отклонение в выборке – поэтому, доверительный интервал должен быть **ШИРЕ**, чем при известном σ.

Насколько шире? Это будет зависеть от **РАЗМЕРА ВЫБОРКИ** (от числа **степеней свободы** df = n-1)

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

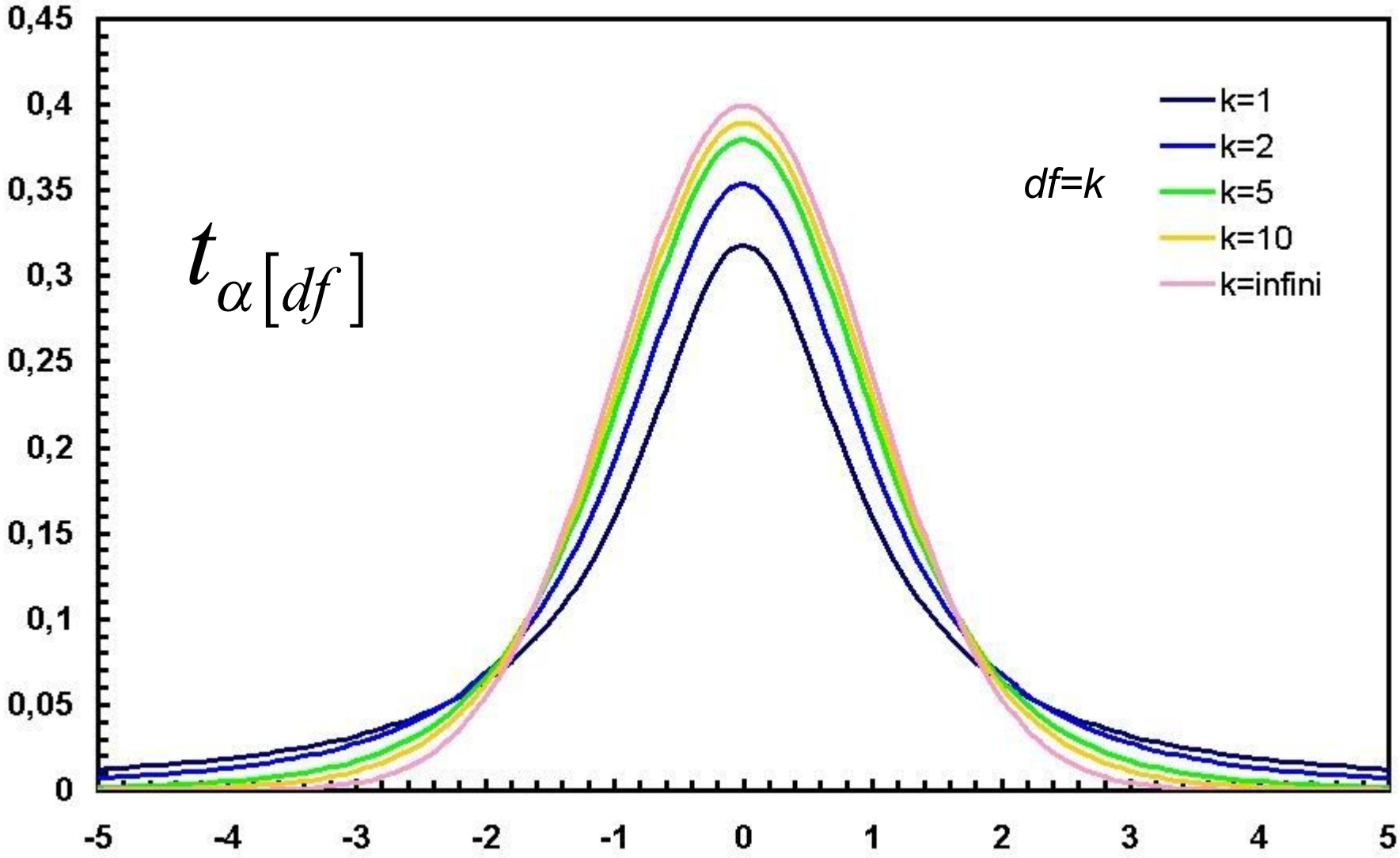
df

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

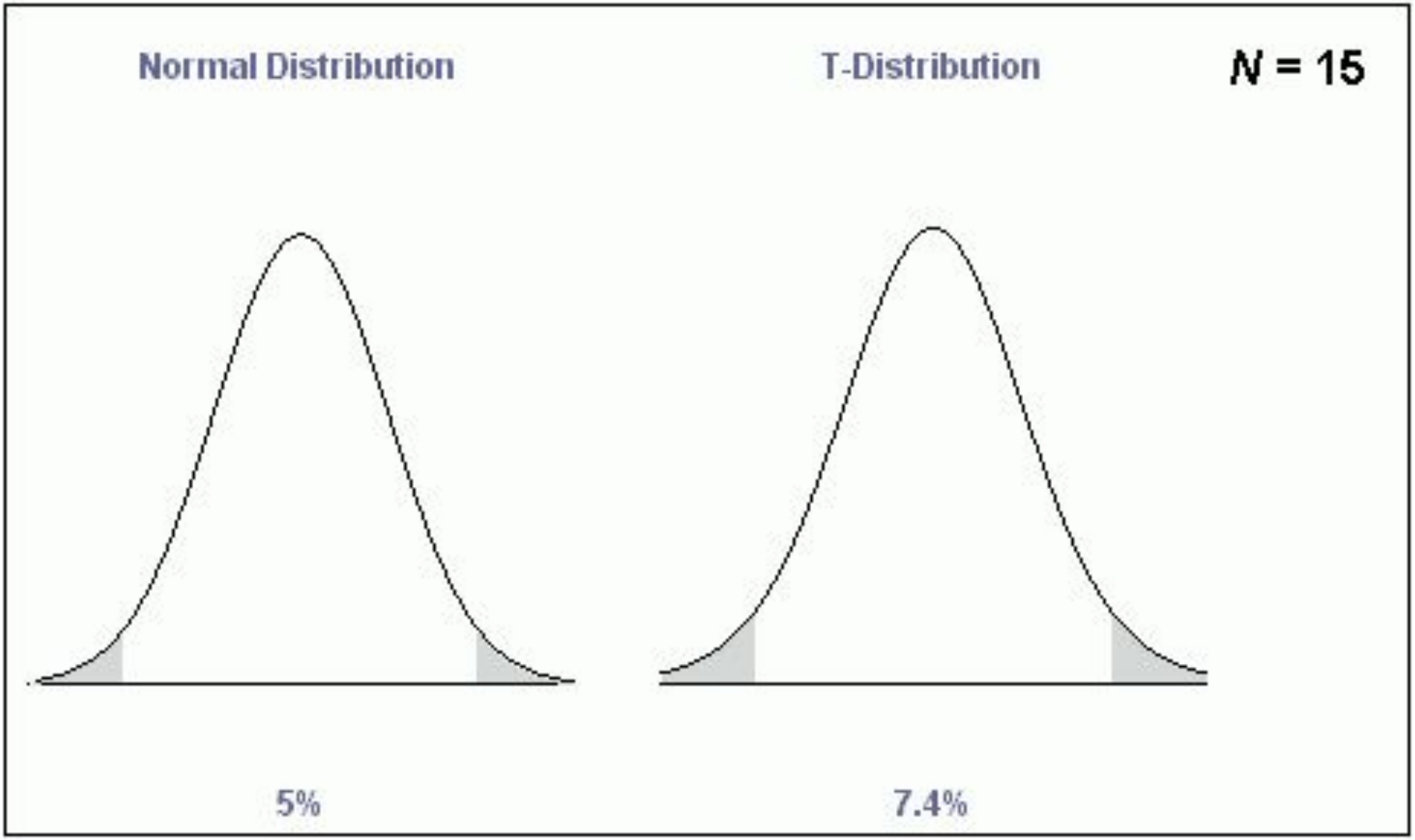
$$SE = s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Пояснить про число степеней свободы

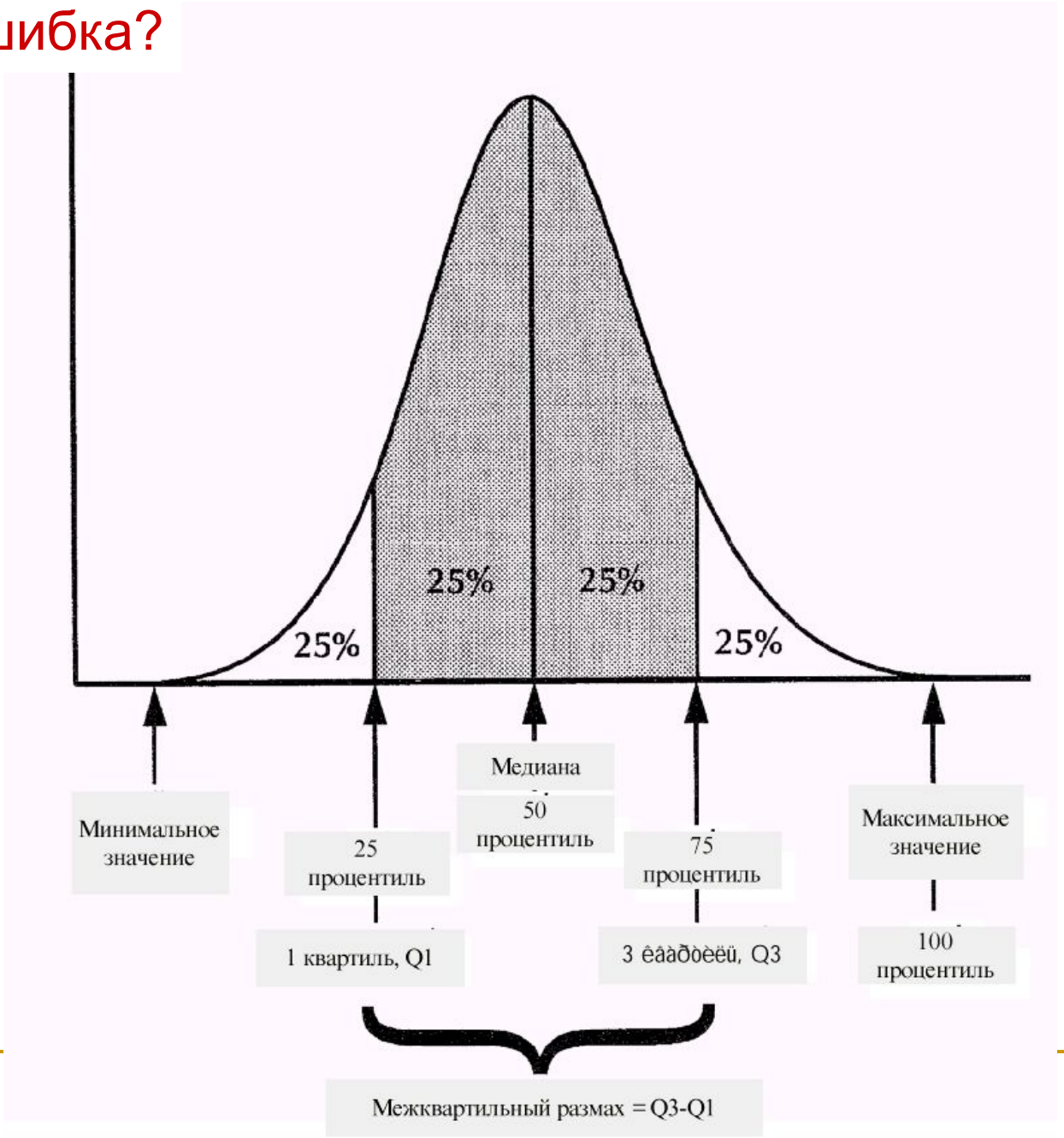
t-распределение (Стьюдента)



При больших (>30) размерах выборок приближается к нормальному


Normal Distribution
T-Distribution
N = 15
5%
7.4%

В чём ошибка?


25%
25%
25%
25%
Минимальное значение
25 процентиль
1 квартиль, Q1
Медиана
50 процентиль
75 процентиль
3 êâàðòèëü, Q3
Максимальное значение
100 процентиль
Межквартильный размах = Q3-Q1