Занятие 1

Основные понятия. Описательная статистика.

Нина Александровна Васильева
ninavasilieva@gmail.com

Познание мира

Дедуктивные умозаключения – из общих суждений (аксиом) к частным.

Индуктивные – от частных суждений к общим.

С помощью индукции мы:

1) описываем явления;

Ночью все кошки серы

2) устанавливаем связи между ними (много снега – много хлеба).

Как проверить истинность индуктивных суждений?

отправиться в поле, мерить снежный покров, регистрировать урожай пшеницы; ловить кошек.

собирать ДАННЫЕ (data) – результаты измерений какой-либо ПЕРЕМЕННОЙ – variable. Например: вес, длина тела, пол, окрас, температура .....

Статистика – инструмент для количественного анализа и интерпретации данных – для проверки ИСТИННОСТИ индуктивных суждений.

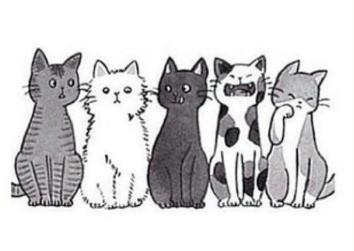
Обычно невозможно доказать истинность индуктивного суждения, рассматривая все объекты подряд. Хорошо бы ограничиться конечным числом объектов!

Генеральная совокупность (population) – совокупность всех интересующих нас объектов

Выборка (sample)

Наблюдение (observation= object)

Описательная (descriptive) статистика:
ОПИСЫВАЕМ ВЫБОРКУ



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

Если мы на основе выборки хотим судить о популяции, она должна хорошо отражать свойства популяции (быть репрезентативной).
Проще всего этого достичь – сделать её **СЛУЧАЙНОЙ (RANDOM)**, чтобы все объекты имели равные шансы в неё попасть.
(это действительно важно, и удаётся не всем!)

Два слова о вероятности

Пусть A – некоторое событие (event);
N – общее число наблюдений;
n - количество наступлений события A, тогда

Вероятность события A:

$$P(A) = \lim_{N \to \infty} \frac{n}{N},$$

$0 \le P(A) \le 1$

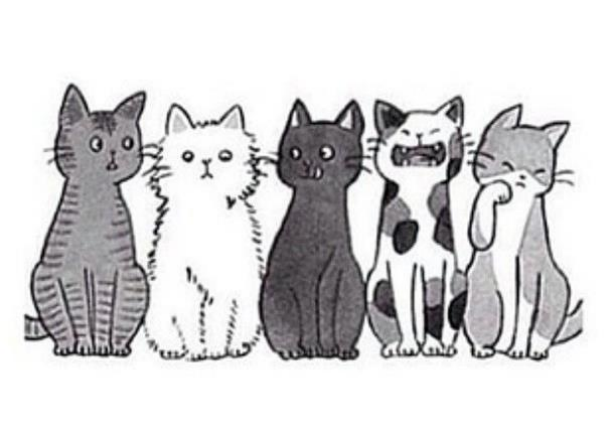
Т.е., если мы много раз бросаем монетку, примерно в половине случаев выпадет «орёл», т.е., P(орла) = 0,5.

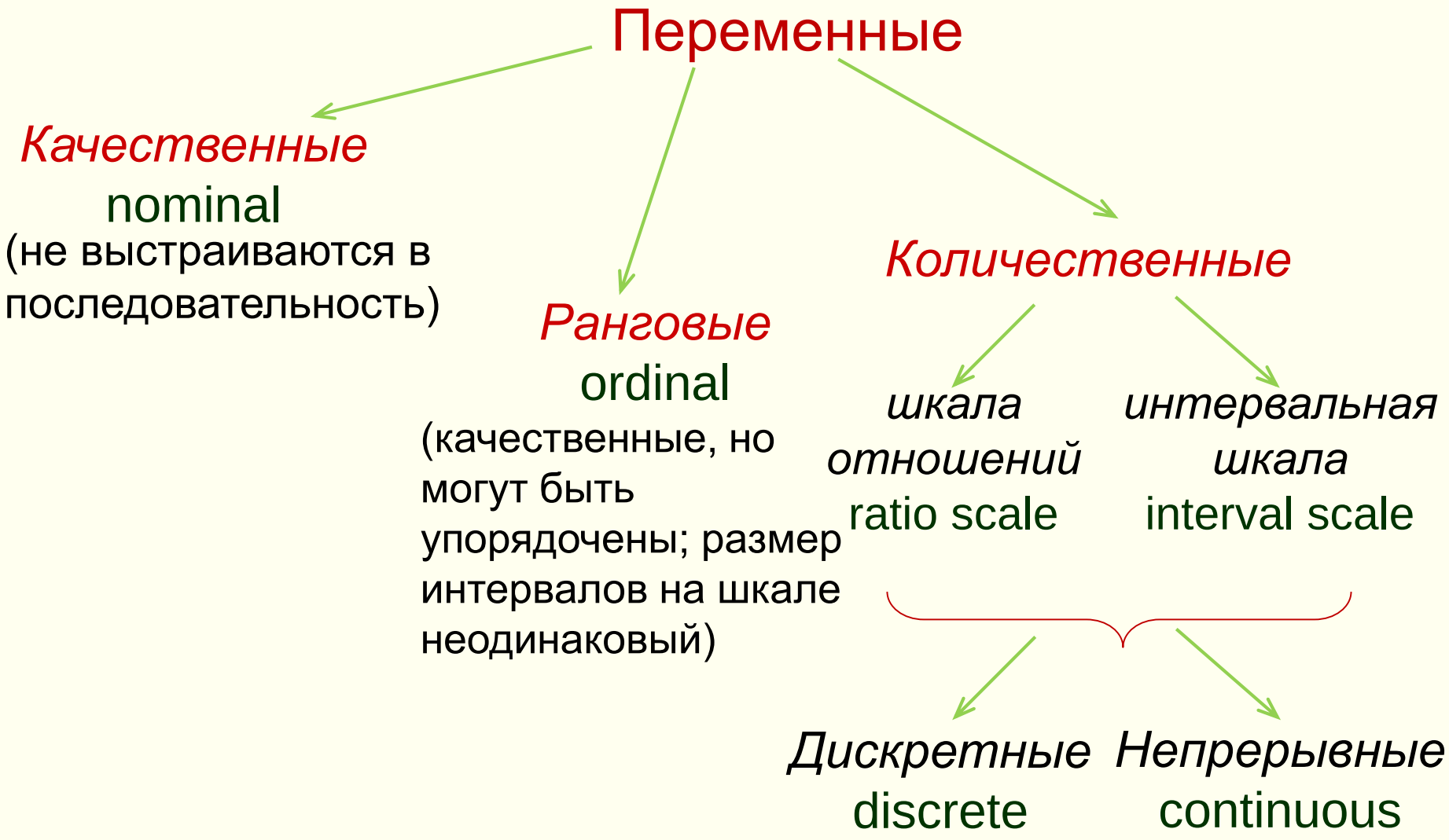
К чему всё это?

Классический подход (frequentist approach)

Мы хотим по **случайной выборке** судить о ГС!
Мы поймали 3 серых кошки и 1 черную. Что теперь нам известно о цвете кошек вообще?

Получаем представления о **значениях переменных** в **популяции**, измеряя эти переменные (цвет кошек) в **выборке**.




Переменные
Качественные
nominal
(не выстраиваются 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;

• возможно посчитать отношение значений;

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;

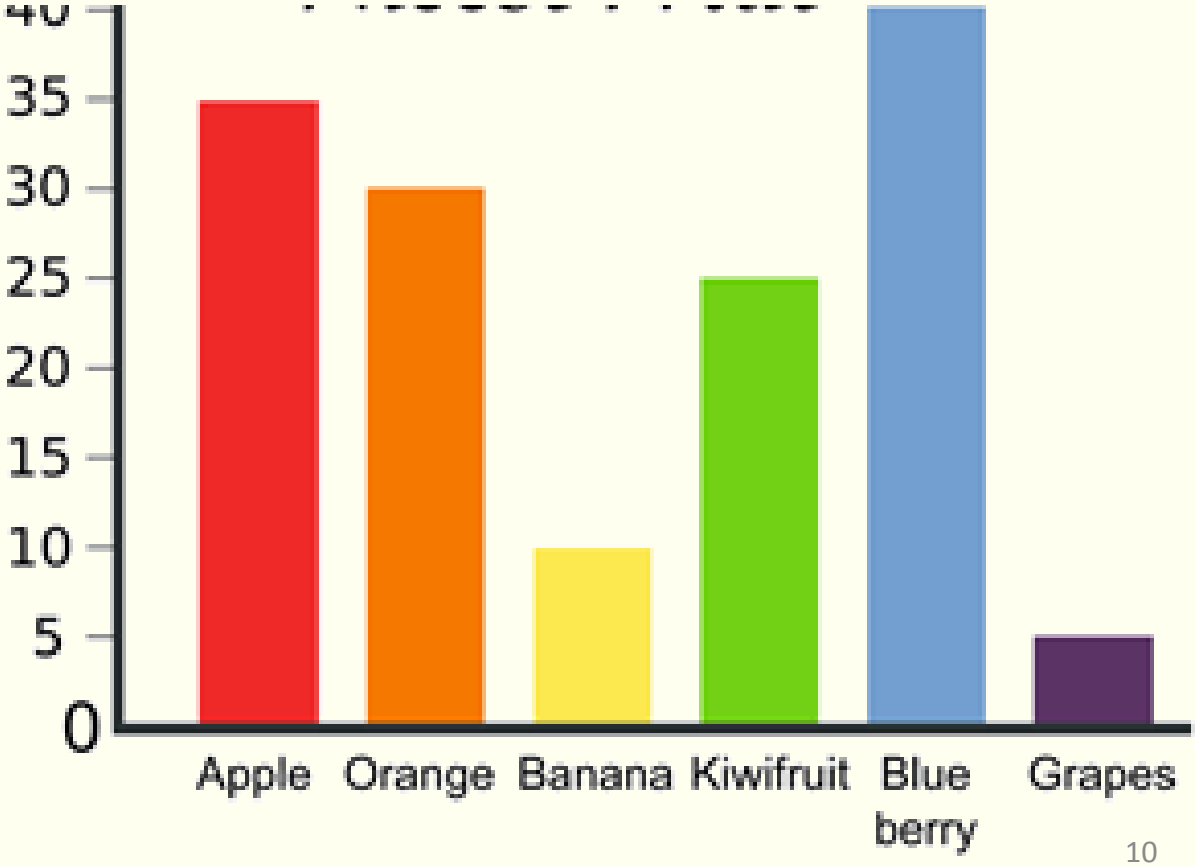
•Отношение значений не имеет смысла

Примеры: температура по Цельсию, время дня, дата

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

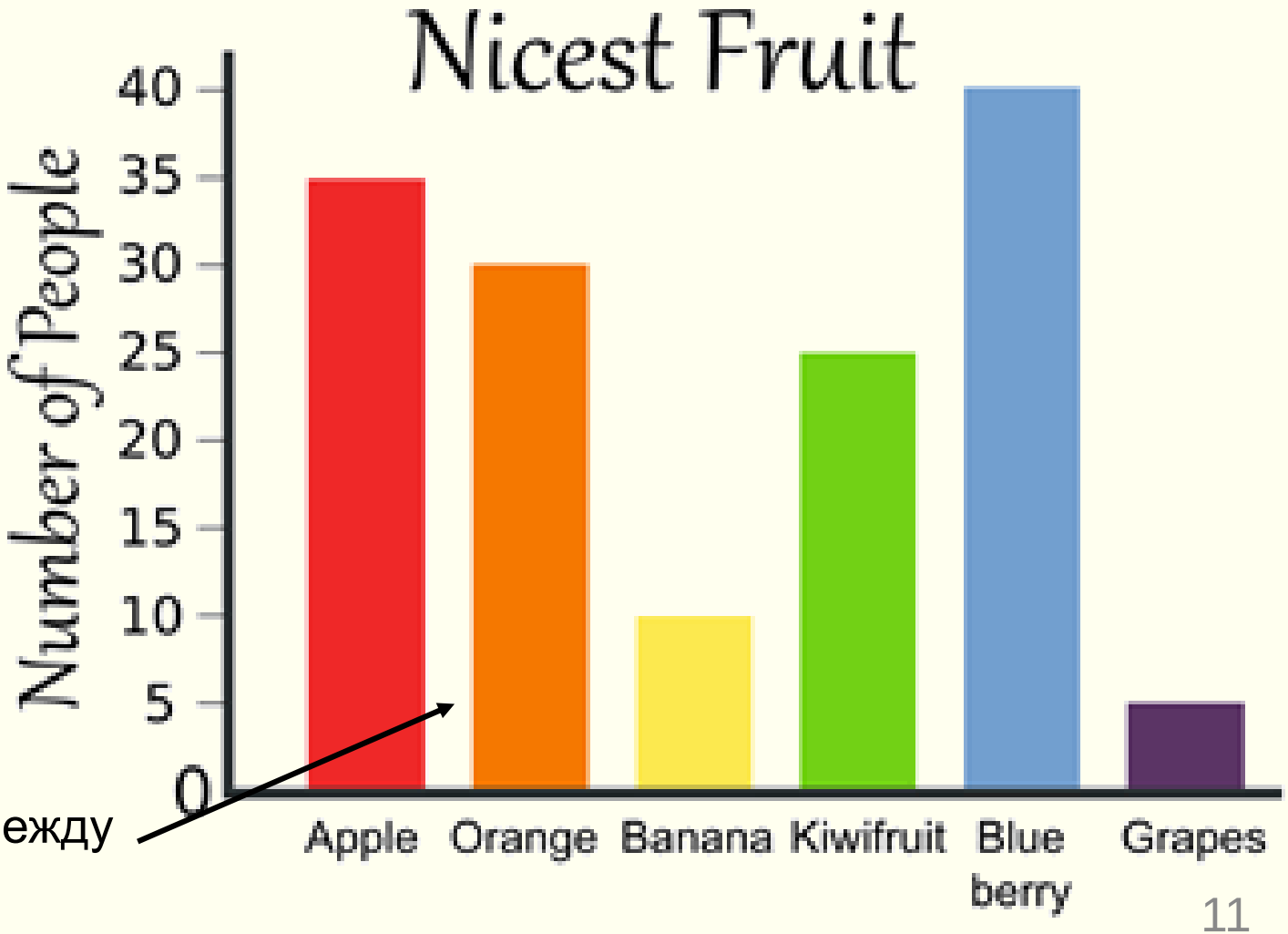
Рассмотрение частотного распределения облегчает обдумывание и обсуждение данных

Можно представить в виде таблички или картинки.



Частотное распределение переменной

Для **КАЧЕСТВЕННЫХ** или **ранговых** переменных - **bar graph** = столбчатая диаграмма («гистограмма» - не совсем верно).

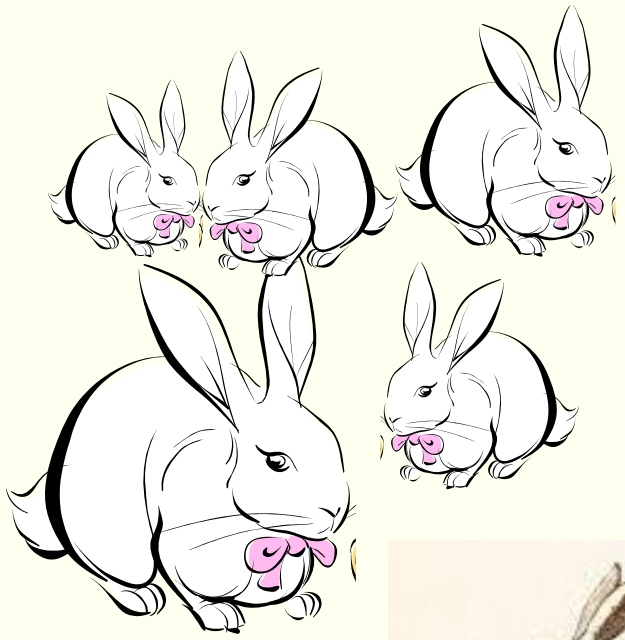


промежутки между столбиками

Частотное распределение переменной (frequency distribution)

Для **КОЛИЧЕСТВЕННЫХ** переменных строим картинку, разделив наблюдения на группы:

1. Получили значения переменной – взвесили ***n*** кроликов;
2. Разбили весь диапазон масс **на равные промежутки**;
3. **Сгруппировали** кроликов по этим промежуткам и посчитали, в какую группу сколько попало.



Частотное распределение переменной (frequency distribution)

Частота – то, сколько кроликов попало в каждую группу.

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

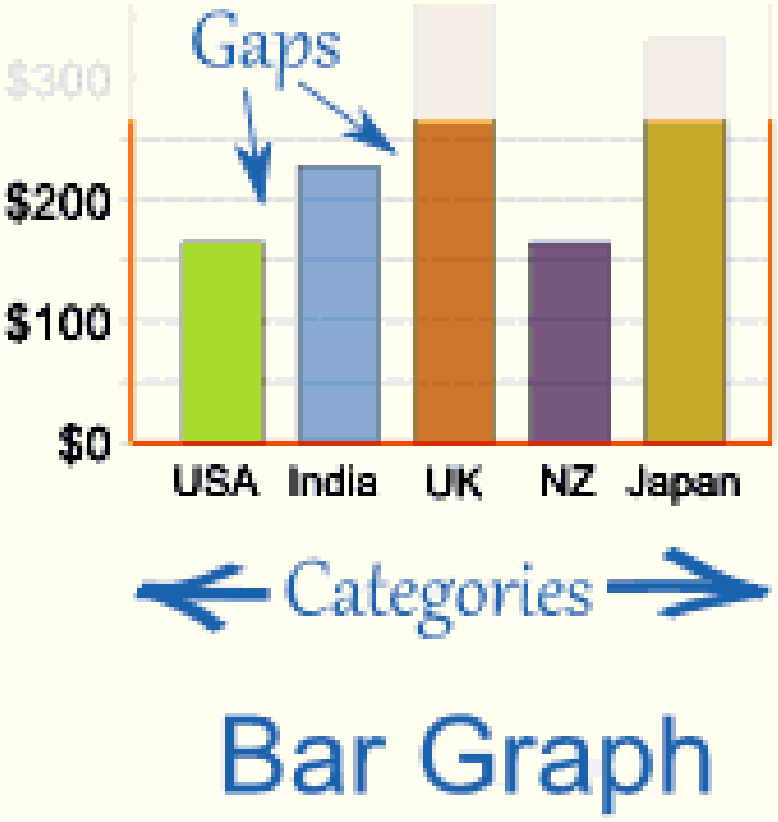
Интервалы должны быть:

•одного размера,

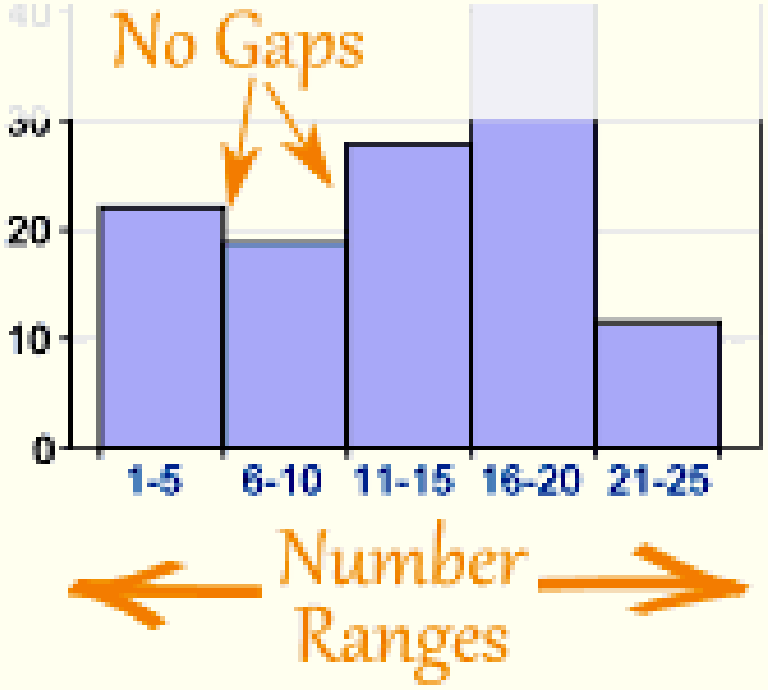
•не должны иметь общих точек,

•для биологических данных – 10-20 интервалов

Частотное распределение переменной



Bar Graph



Histogram

Как описать частотное распределение переменной?

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

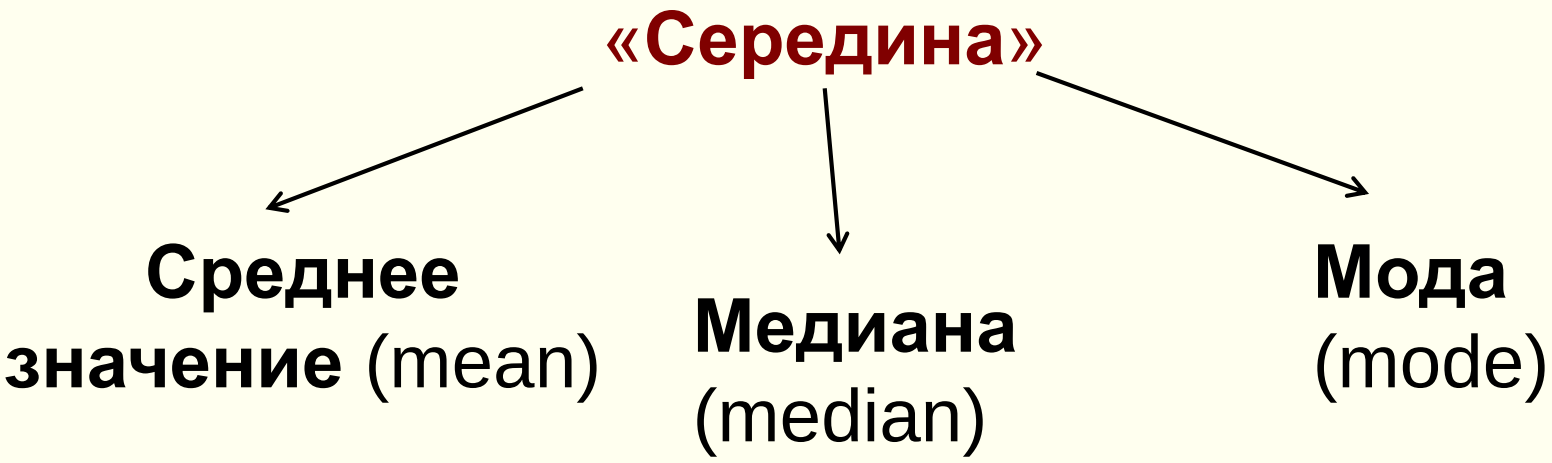
1. «**Середина**» распределения - center;
2. «**Ширина**» распределения - spread;
3. **Форма** распределения

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

Parameter estimation - важный раздел анализа данных

Речь идёт и о количественных данных, и о качественных

«Середина» распределения (central tendency)



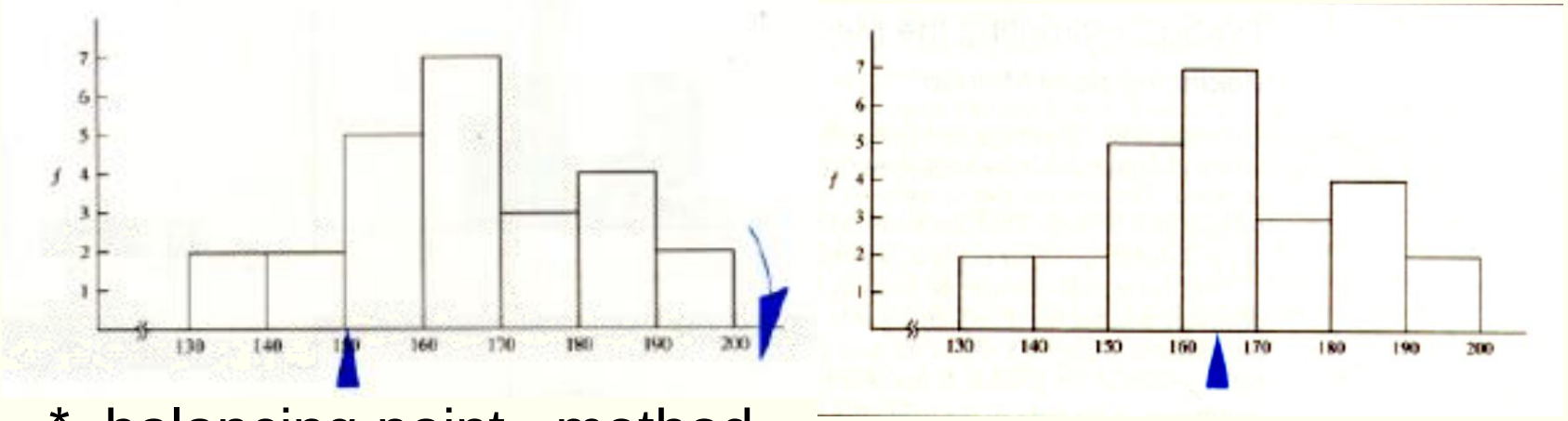
Все они могут служить оценками популяционного среднего.

Среднее в выборке – наиболее эффективная и несмещённая оценка (т.е., взяв много выборок, мы получим поровну средних значений больше и меньше реального среднего).

Частотное распределение переменной

«**Середина**» распределения

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



*«balancing point» method

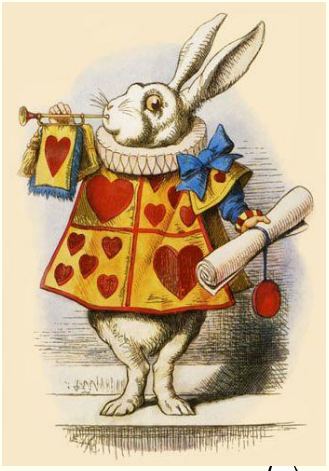
Среднее для **выборки**

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

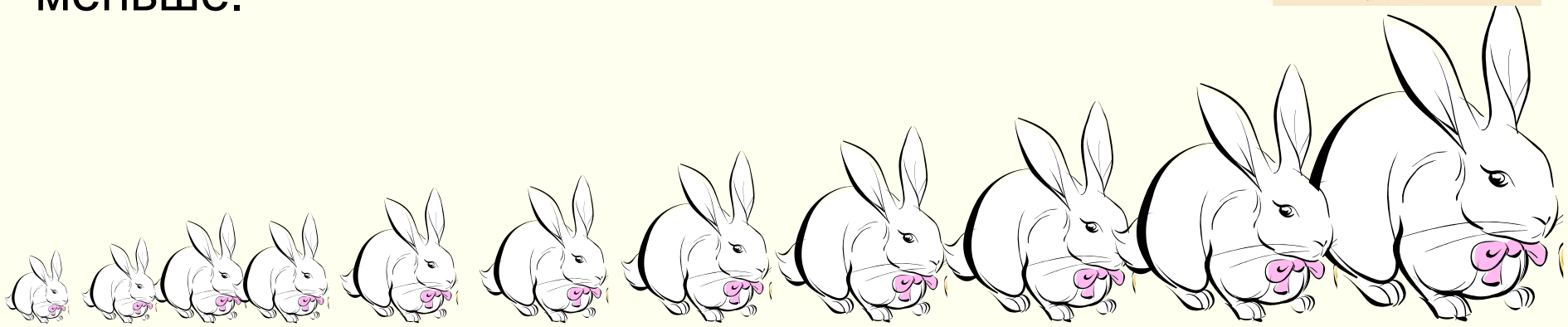
Среднее для **популяции**

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

Частотное распределение переменной «**Середина**» распределения



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



1,0 1,5 3,2 4,1 5,7 **6,0** 7,1 7,9 9,5 10,4 11,0

Медиана

Масса тела, кг

Имеет смысл не только для количественных переменных, но и для ранговых! (не для качественных).

Частотное распределение переменной

- ✓ Если распределение не симметричное, **медиана** лучше характеризует **центр** распределения (но не среднее).
- ✓ 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.

Данные можно ранжировать и поделить распределение не только на ДВЕ равные части, но и на:

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

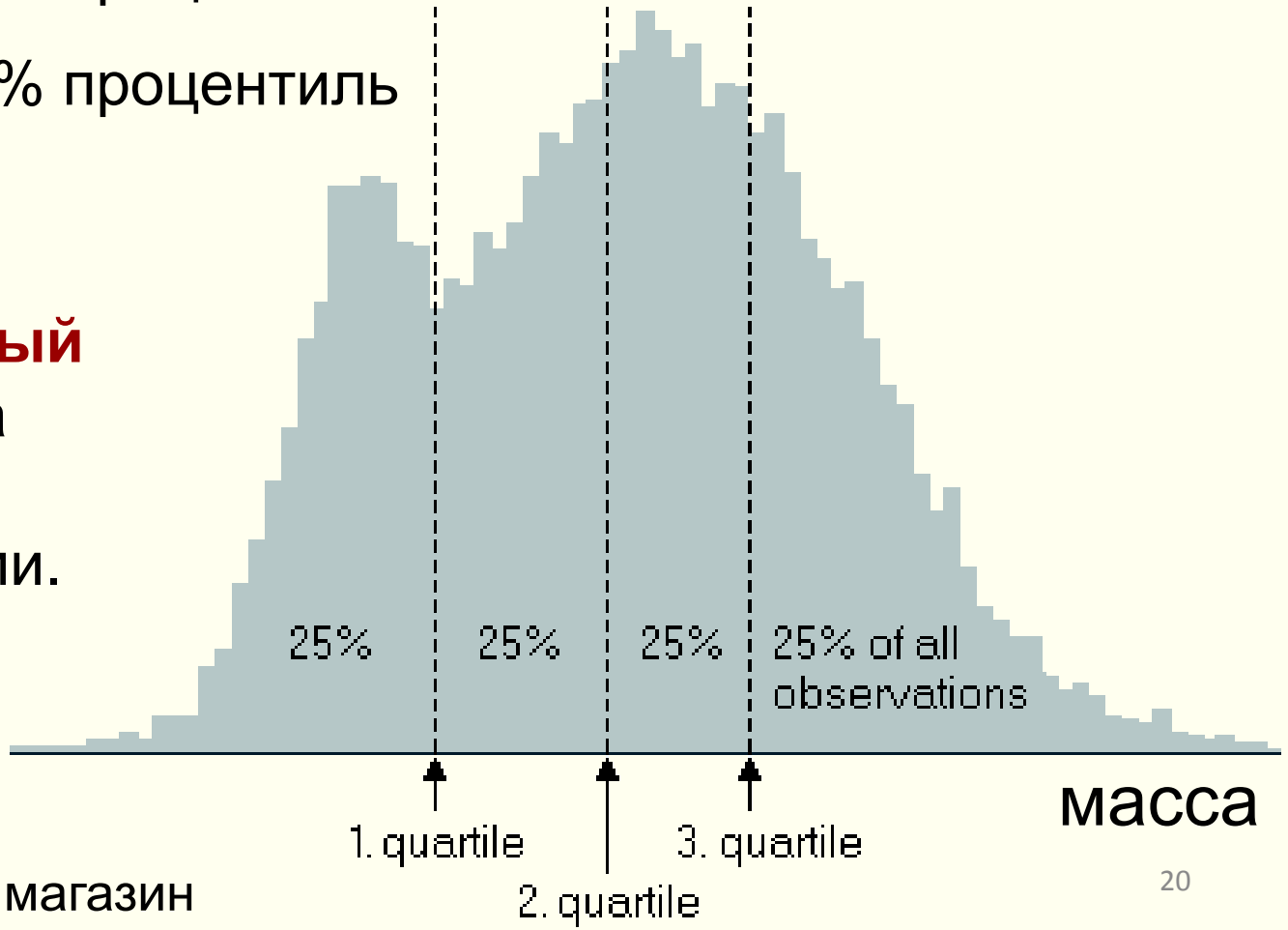
Частотное распределение переменной

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.

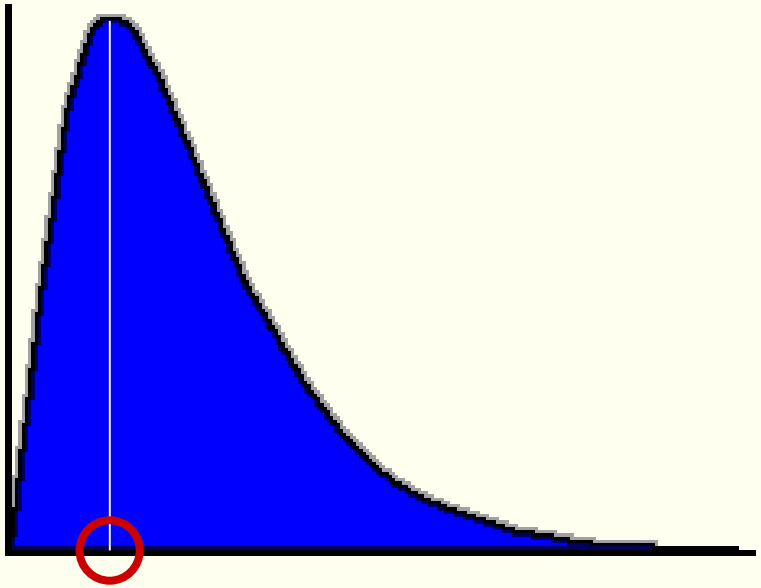


Пример про 500 р и магазин

Частотное распределение переменной «**Середина**» распределения

Мода (mode) – наиболее часто встречающееся значение, локальный максимум.

Существует для количественных, для ранговых и для качественных переменных

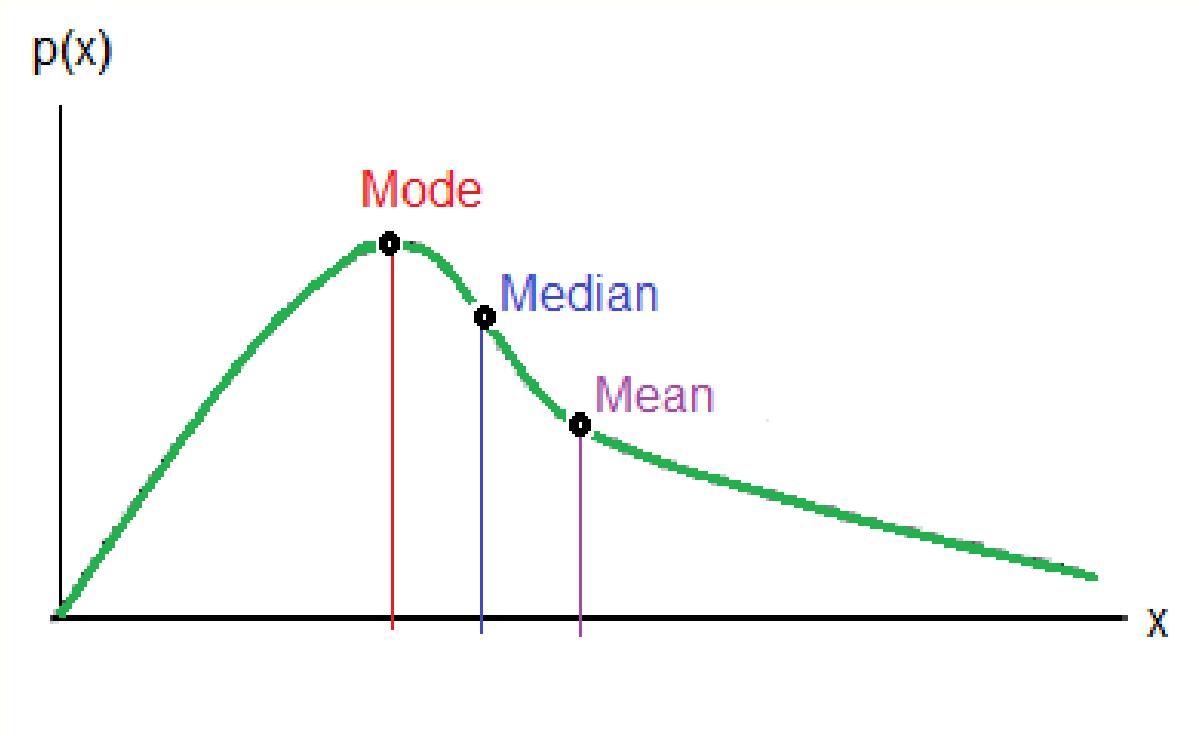


В первую очередь биолога интересует количество мод в распределении, а не мода как таковая. Если мода не одна, наверняка выборка может быть поделена на группы

Частотное распределение переменной «**Середина**» распределения

Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения

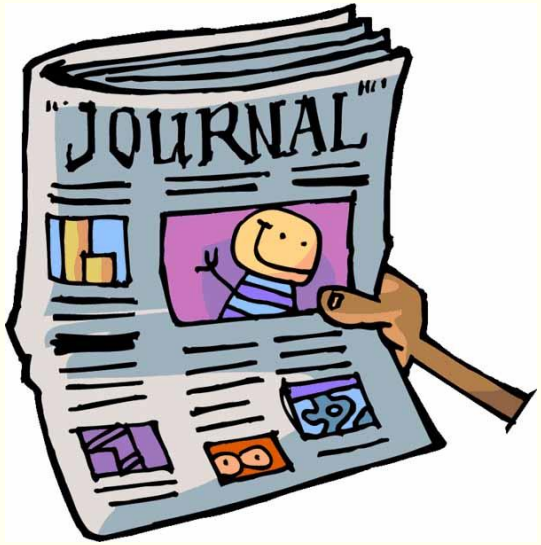
ЗАРПЛАТА, руб	ЧАСТОТА
200000	2
20000	10
19000	20
10000	100



К появлению перекоса чувствительнее всего среднее значение

Для публикаций

- ✓Традиционно, для выборки приводят **среднее значение** (mean) – удобно для сравнения с литературой и пр.;
- ✓Если распределение скошенное, дополнительно приводят медиану (M);
- ✓Моду обычно не приводят; на количество мод мы смотрим, когда обдумываем данные.



Частотное распределение переменной

«**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.

* Это лишь основные параметры разброса

Дисперсия (variance)

Для **выборки**:

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

Для популяции:

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

Сумма квадратов *отклонений (sum of squares = SS)*

- ✓Зависит ото всех значений переменной.
- ✓Измеряется в единицах переменной, возведённых в квадрат (что не всегда удобно).
- ✓Дисперсия используется больше как составляющая в анализе данных, а не сама по себе в описательной статистике

Для популяции чтобы получить несмещённую оценку дисперсии, мы ставим в знаменатель n-1, иначе она получится меньше, чем нужно

Частотное распределение переменной

Разброс распределения

Стандартное отклонение (standard deviation)

Для **выборки**:

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

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

Для популяции:

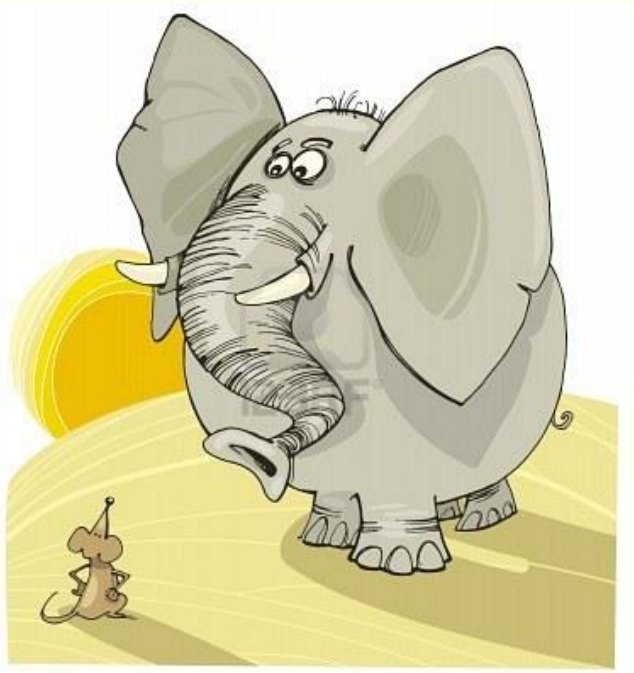
$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$$

Сумма квадратов *отклонений (sum of squares = SS)*

- ✓Это корень из дисперсии.
- ✓Стандартное отклонение зависит ото всех значений переменной.
- ✓Измеряется в тех же единицах, что и переменная!

Разброс распределения

Если надо сравнить изменчивость длины уха слонов и мышей:



Коэффициент вариации
(Coefficient of variation)

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

Даёт понять, насколько на самом деле велик разброс в данных, независимо от единиц и масштаба измерений (маленький разброс – меньше 5%)

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

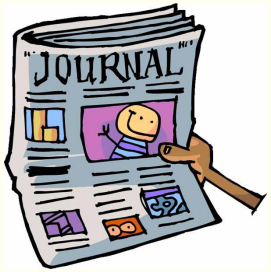
$$H = -\sum_{i=1}^{k} p_i \log p_i$$

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!

Для публикаций



- ✓ Традиционно, вместе со **средним** значением приводят **стандартное отклонение** (±SD);
- ✓ Иногда в статье приводится размах, но в дополнение следует привести ещё какую-нибудь характеристику разброса;
- ✓ Коэффициент вариации приводят, если хотят сравнить разброс в разных по характеру данных.

«…if not noted otherwise, values in the table are means ± SD (range; N)»

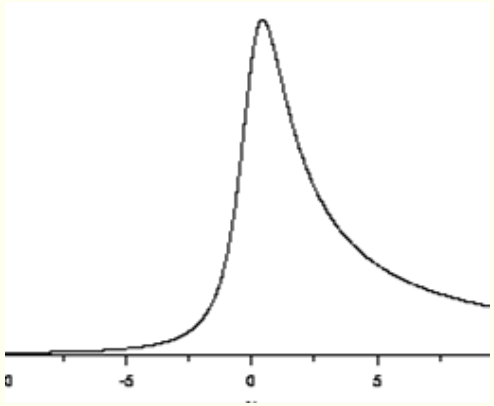
«…The difference between the dates was 55.8 ± 3.1 days (mean ± SD) and varied only slightly among the females (coefficient of variation = 5.5)»

Частотное распределение переменной

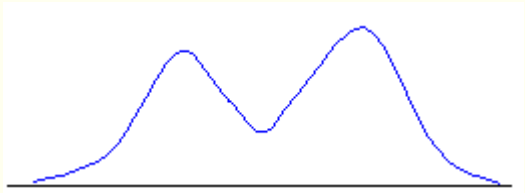
По ФОРМЕ распределения различаются:

1. По количеству «максимумов» (мод):

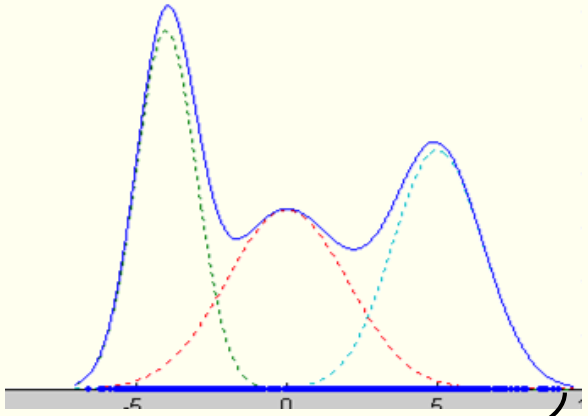
унимодальное



бимодальное



мультимодальное



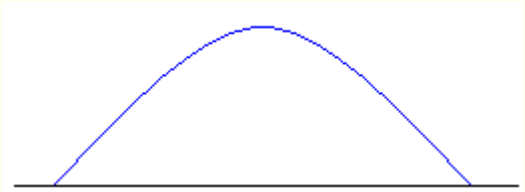
обычно возникают, если популяция имеет естественные обособленные подгруппы

Частотное распределение переменной

По ФОРМЕ распределения различаются:

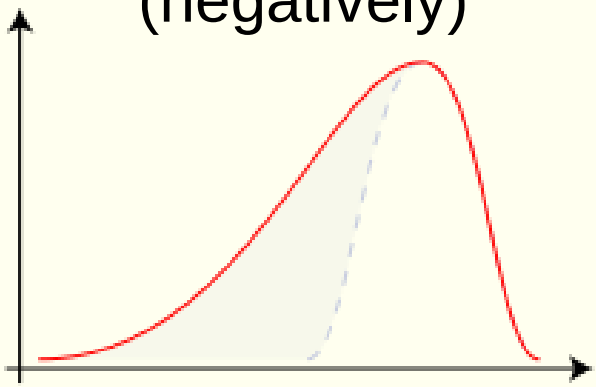
2. По признаку симметрии:

Симметричное



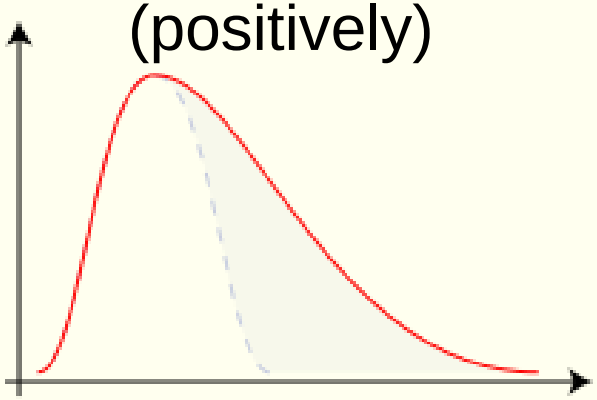
Скошенное (skewed)

влево (negatively)



Negative Skew

вправо (positively)



Positive Skew

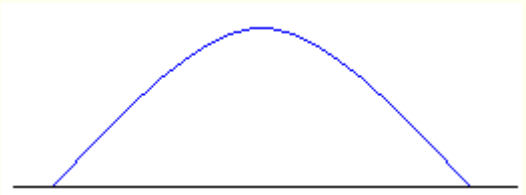
Частотное распределение переменной

По ФОРМЕ распределения различаются:

3. распределение

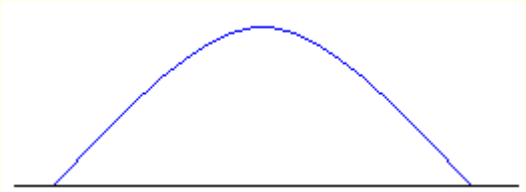
асимптотическое

не асимптотическое

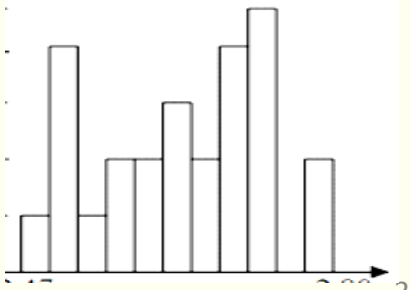


4. распределение

непрерывное

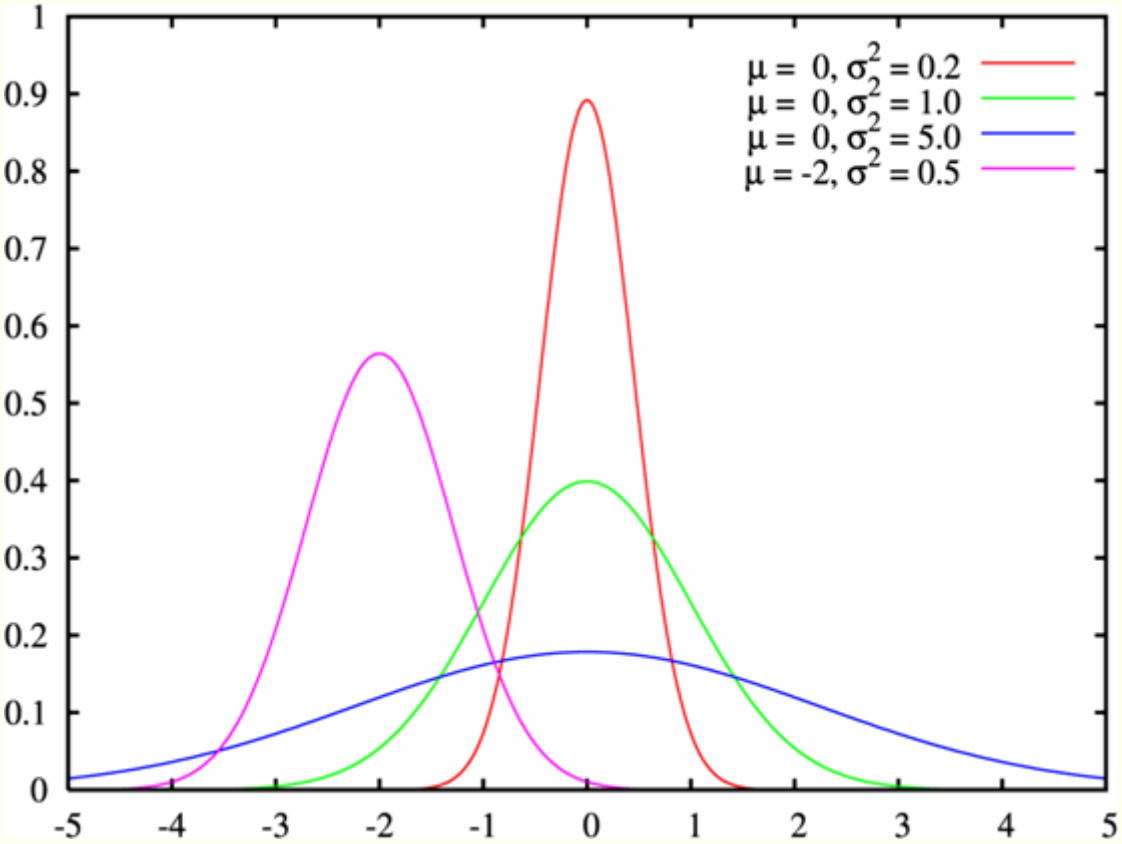


дискретное



Нормальное распределение (Гауссово): первое знакомство

- ✓ Унимодальное
- ✓ Симметричное
- ✓ Асимптотическое



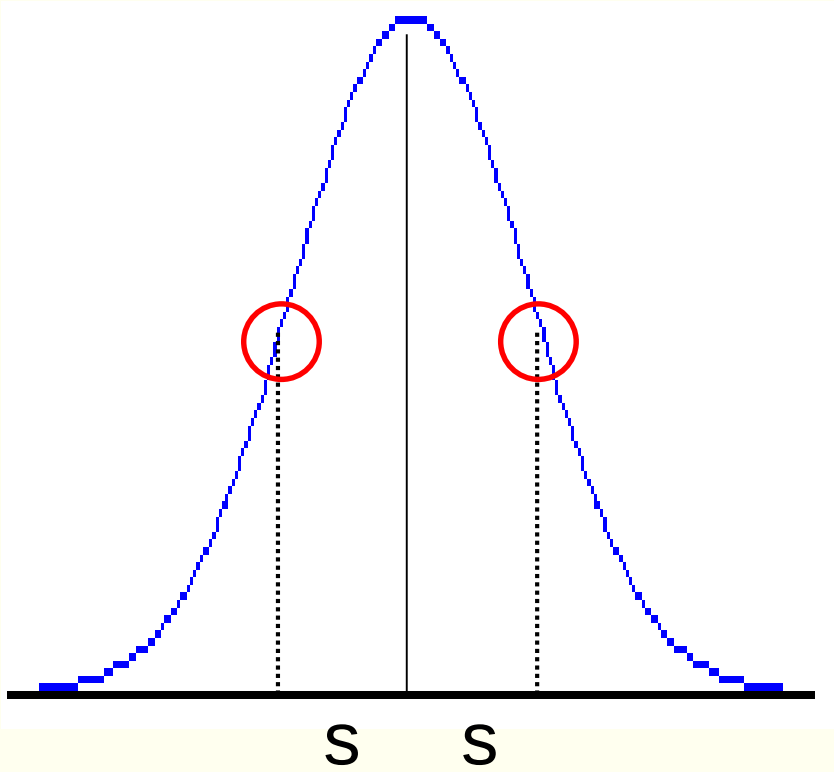
Это непрерывное распределение

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.

Волшебные свойства
нормального распределения

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



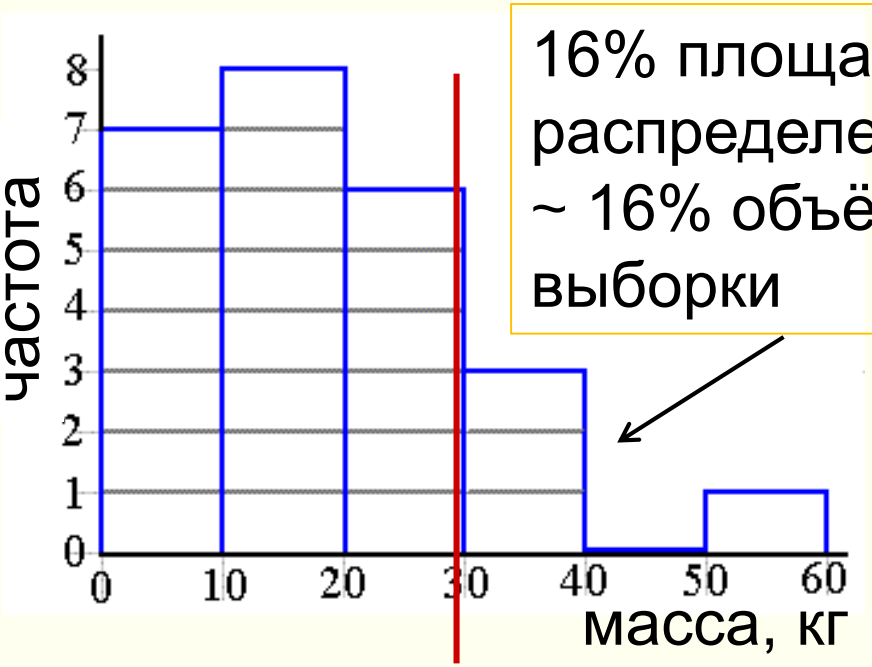
Частотное распределение переменной

«Площадь распределения»

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

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

То же – для генеральной совокупности



16% площади распределения ~ 16% объёма выборки

Обсуждали это в рассказе о медиане и квартилях

Площадь нормального распределения

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

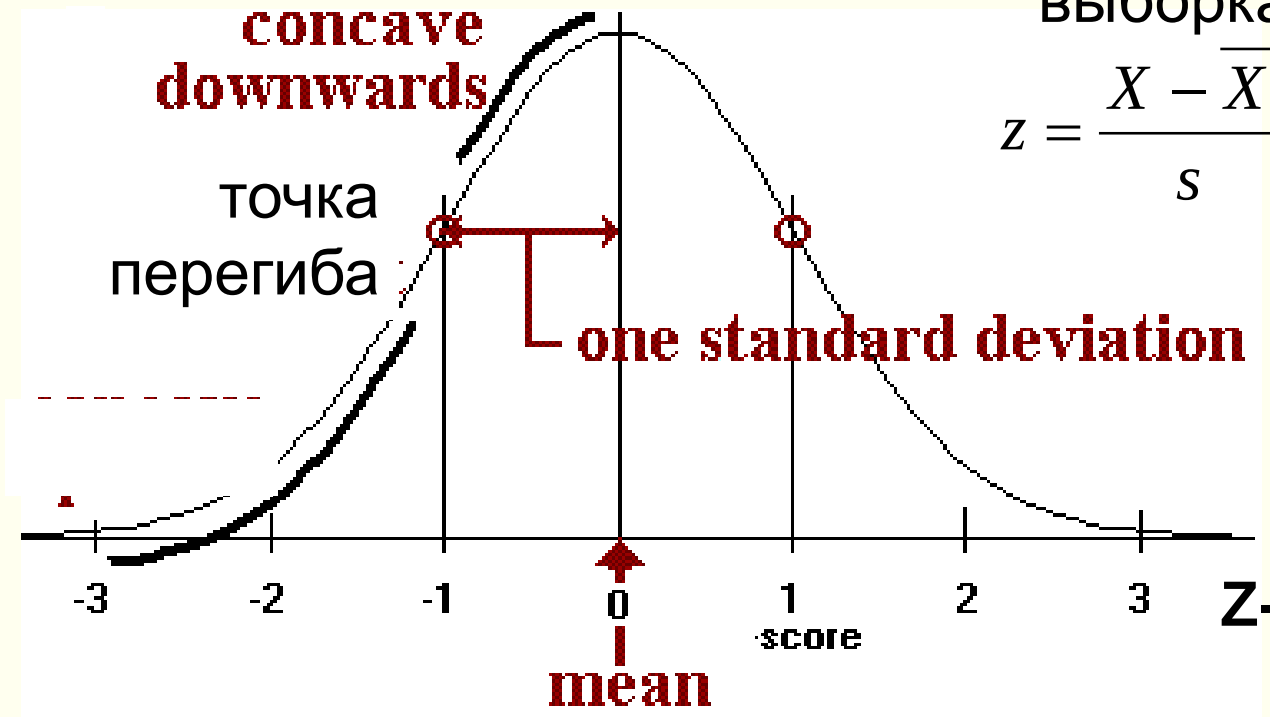
$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Волшебное свойство:

Относительные площади нормального распределения над одинаковым количеством стандартных отклонений всегда одинаковы!

z-оценка (standard score)

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



выборка

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

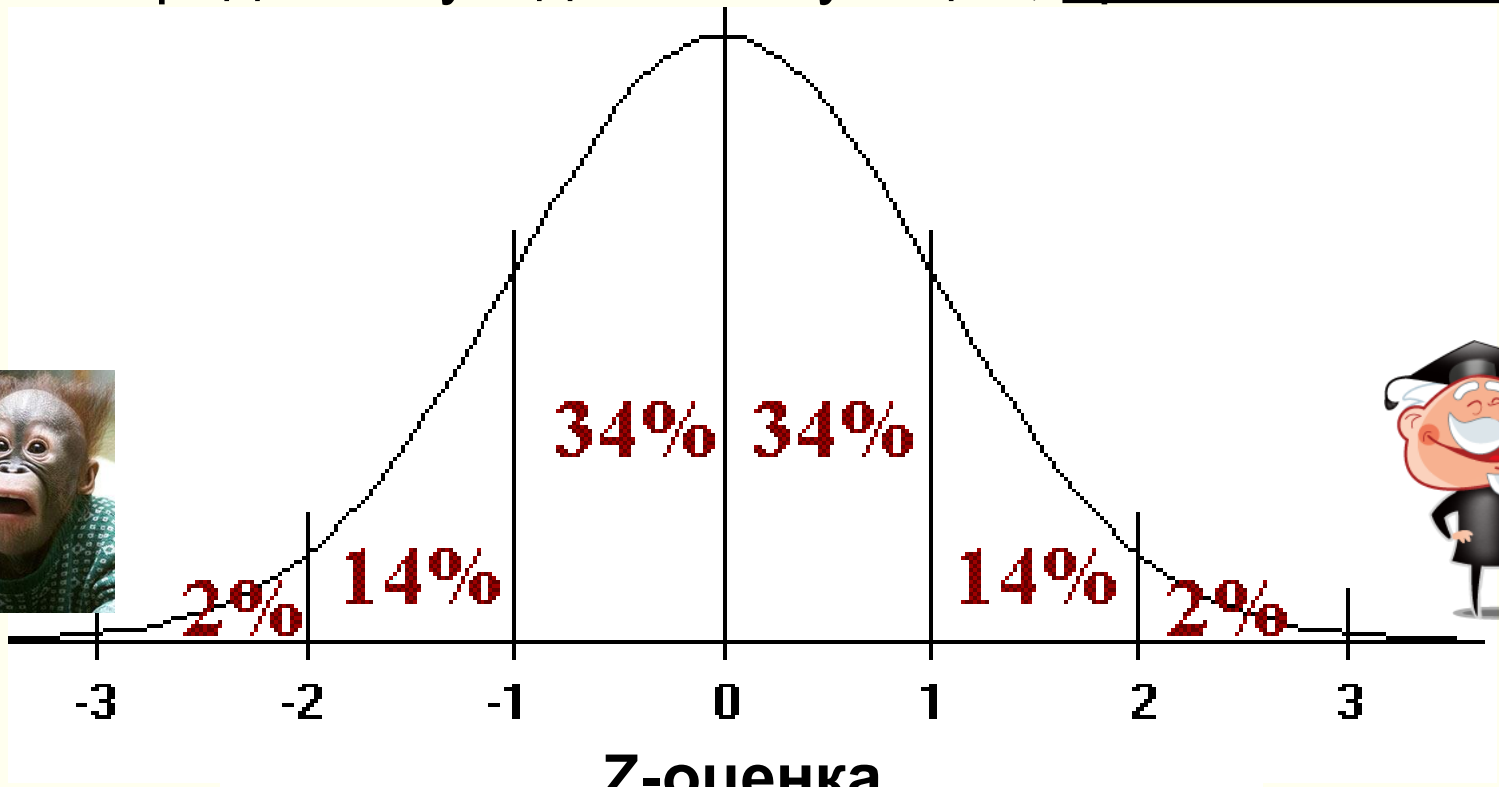
популяция

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

Z-оценка

Площадь нормального распределения

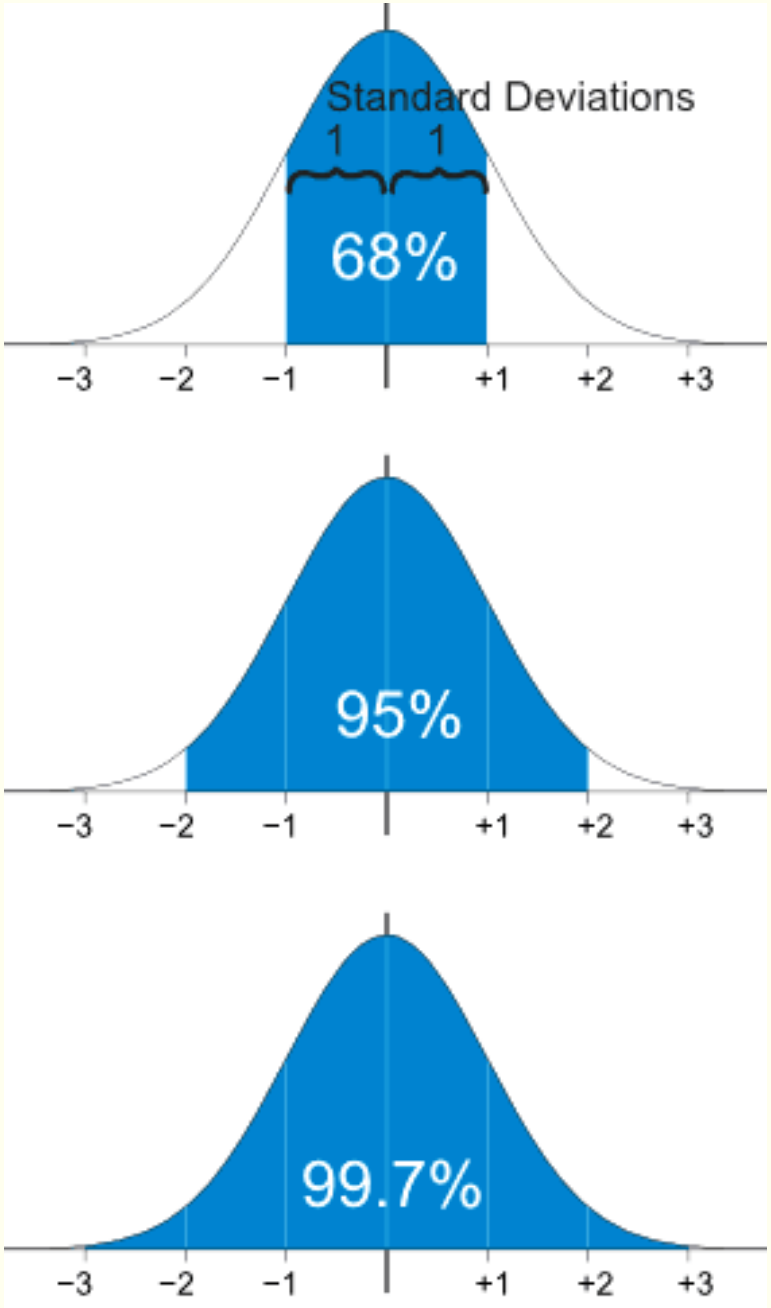
Откладывая от среднего значения стандартное отклонение (в ту или другую сторону) мы всегда отрезаем строго определённую долю популяции, приблизительно:



Z-оценка
(количество стандартных отклонений)

Пример с IQ (μ=100, σ=15)

Площадь нормального распределения



А площадям под частотными распределениями на самом деле соответствуют вероятности попасть в данных интервал.

Распределение выборочных средних и стандартная ошибка среднего

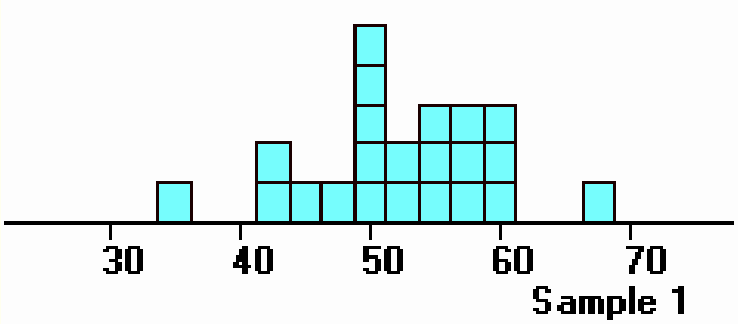
Обычно у нас в руках лишь одна выборка из популяции. Но никто не запрещает нам сделать несколько случайных выборок (пусть они будут одного размера)!

В магазин привезли апельсины в 25 ящиках, по 22 штуки в каждом. На этикетках написана масса ящика, в среднем 1 апельсин весит 50 г. Но апельсины **разные**, и **средние массы** апельсинов в ящиках будут немного **разные**!

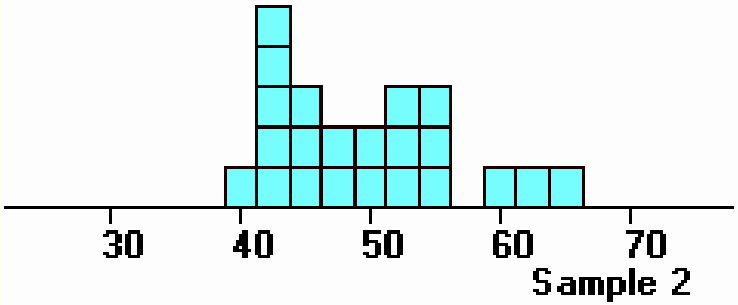
Распределение выборочных средних

Построим частотные распределения массы апельсинов для каждого ящика:

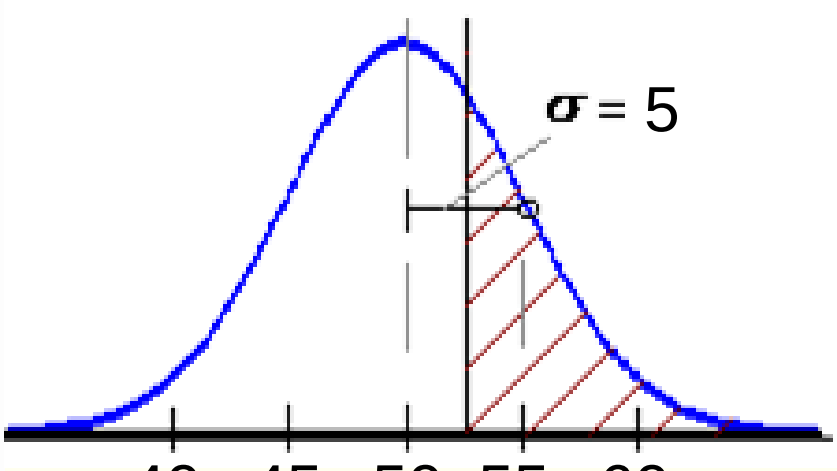
Ящик 1:



Ящик 2:



…

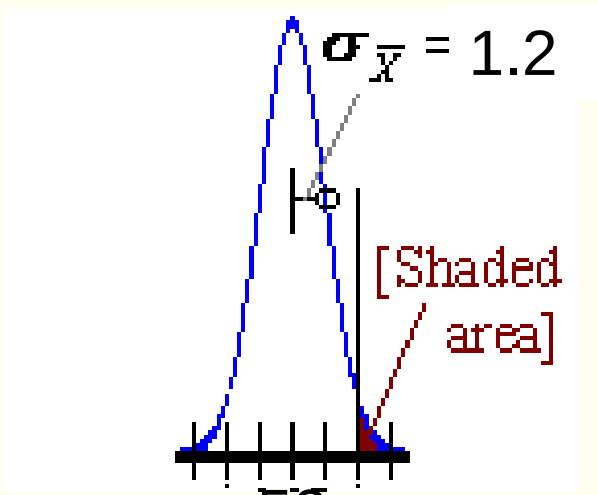


Это распределение всех апельсинов вообще (ГС)

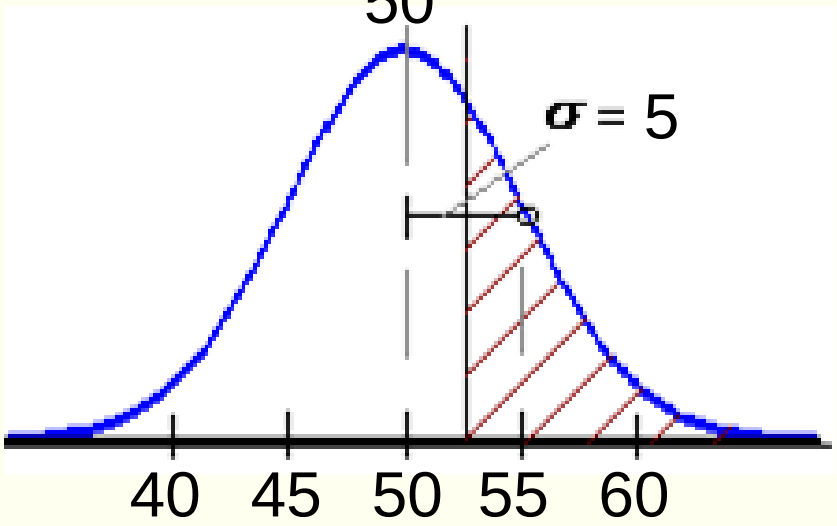
Получится 25 распределений, и 25 **средних масс**!

Распределение выборочных средних

А теперь построим **распределение** из этих СРЕДНИХ значений масс



Среднее в этом распределении будет близко популяционному среднему, и оно будет намного **УЖЕ** распределения всех апельсинов, и УЖЕ, чем каждое из распределений выборок



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

Распределение выборочных средних

	Популяция (все апельсины)		Выборка (ящик)		Распределение выборочных средних
среднее	μ	$\approx$	$\overline{X}$	$\approx$	$\mu_{\overline{X}}$
стандартное отклонение	σ	$\approx$	s	>>	$\sigma_{\overline{X}}$

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

Стандартная ошибка среднего (Standard error = SE)

Распределение выборочных средних

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

Определяет форму, среднее и разброс в распределении выборочных средних

• **Форма:** с увеличение размера выборок 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.

Распределение выборочных средних

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!

Распределение выборочных средних

Масса кролика определяет многими факторами:

Генотип – 7 кг

Внутриутробные условия – 5 кг

Качество вскармливания мамой – 8 кг

Уход и любовь хозяина – 25 кг

Питание – 20 кг

Т.е., масса кролика – среднее по выборке многих гипотетических масс. А массы нескольких кроликов – выборочные средние

Распределение выборочных средних

Пусть у нас есть одна выборка (один ящик!). Для неё мы посчитали среднее значение $\overline{X}$.

Насколько оно близко среднему значению в популяции *(μ)?*

Пусть нам известно *μ*, найдём, какими вообще могут быть $\overline{X}$

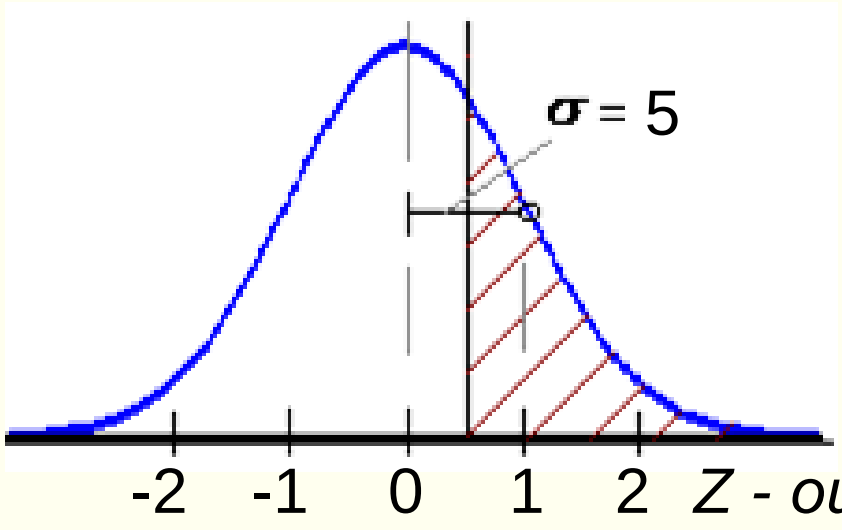
Мы знаем, что для нормального распределения есть **z-оценка**, значениям которой соответствуют **определённые площади** распределения, а они соответствуют **вероятностям** попасть в заданный интервал.

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

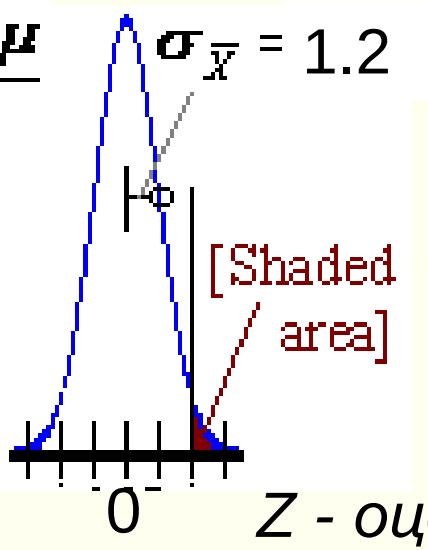
Это значит, что, зная среднее в популяции, мы можем рассчитать интервал, в который попадёт выборочное среднее с вероятностью, скажем, в 95% (или 99%).

Распределение выборочных средних

«Стандартизируем» наши распределения – вычтем среднее и поделим на стандартное отклонение – получим распределения z-оценок.



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



Вопрос: какая доля всех АПЕЛЬСИНОВ имеет массу больше 55 г?

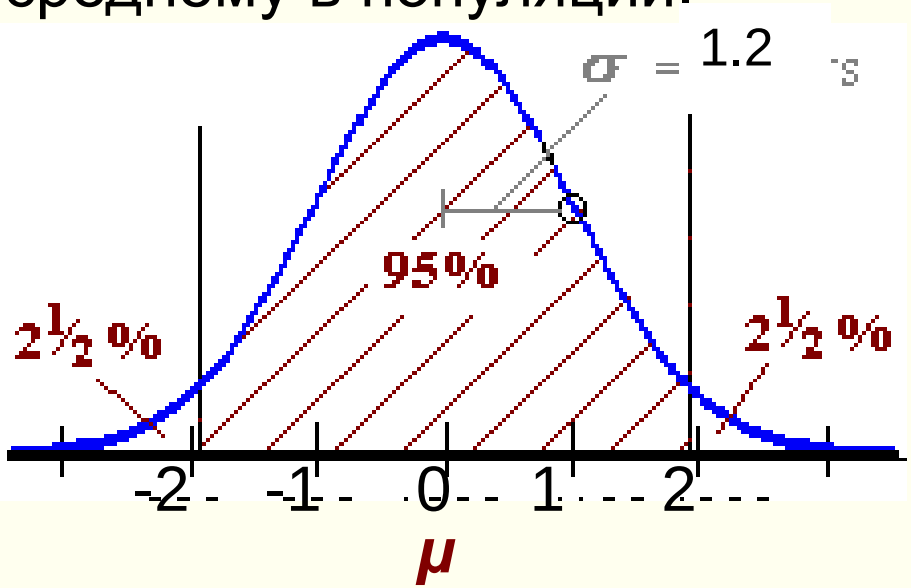
Другой вопрос: какая часть ЯЩИКОВ имеет СРЕДНЮЮ массу апельсинов больше 55 г?

С z-оценками ответить на оба вопроса – легко!

Оценка параметров популяции на основе свойств выборки (parameter estimates)

Оцениваем среднее значение

Пусть мы знаем среднюю массу апельсина и стандартное отклонение в популяции (на ящиках написано). Как оценить среднюю массу в ящике, не взвешивая апельсины? 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 ящиках.

Оценка параметров популяции

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 среднего для 95% выборок **не больше 1.96 SE**

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, граница интервала.

Оценка параметров популяции

Вопрос: чему равно ***μ?***

Ответ: я точно не знаю, но наиболее вероятно – лежит в пределах ± **1.96** 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: чему равно ***μ?***

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

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

Оценка параметров популяции

Мы не знаем стандартное отклонение в популяции, и оцениваем его через стандартное отклонение в выборке (не точно!) – 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

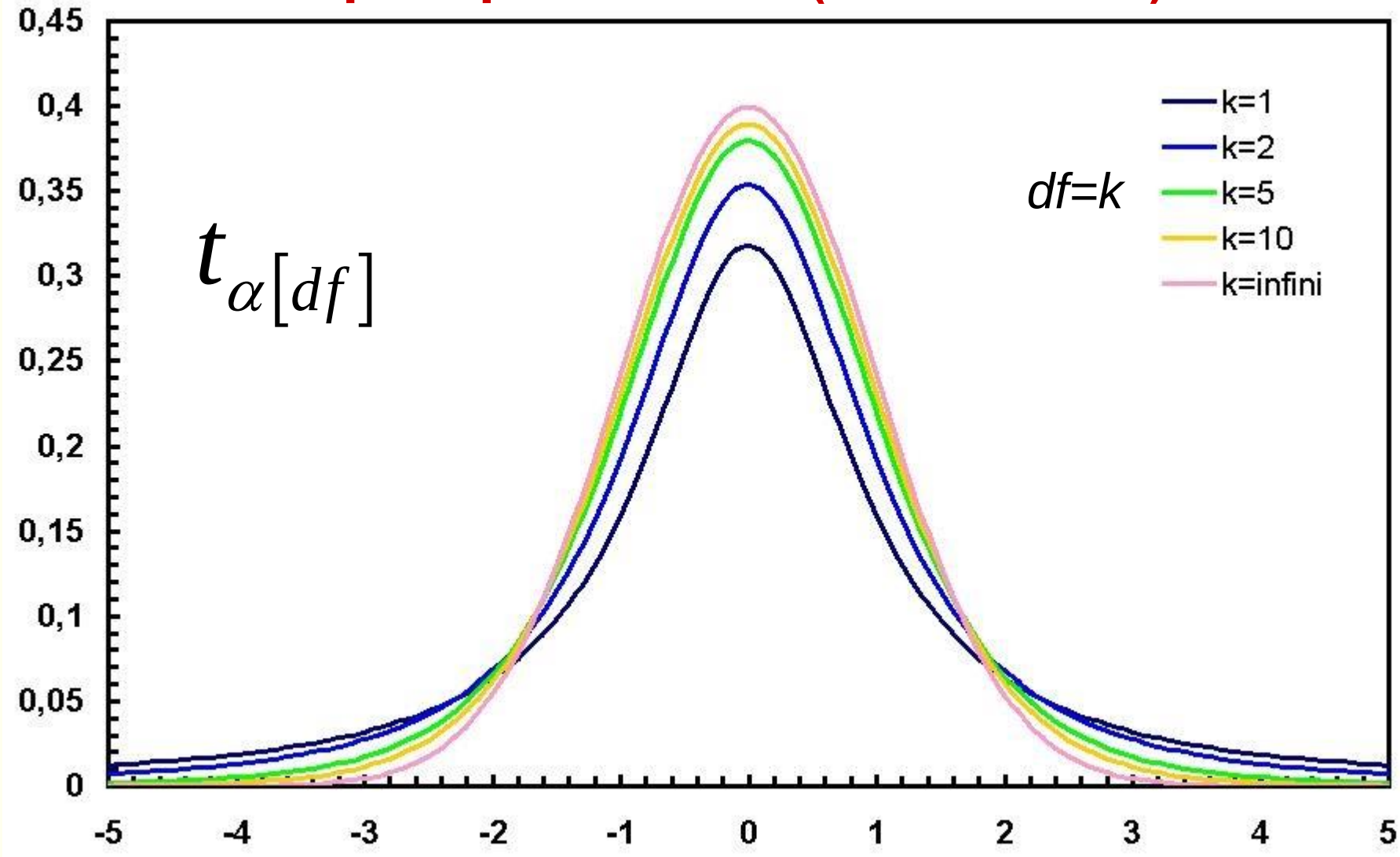
$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$$

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

Но теперь мы будем строить интервал не для нормального распределения (z-оценок), а для особого *t*-распределения.

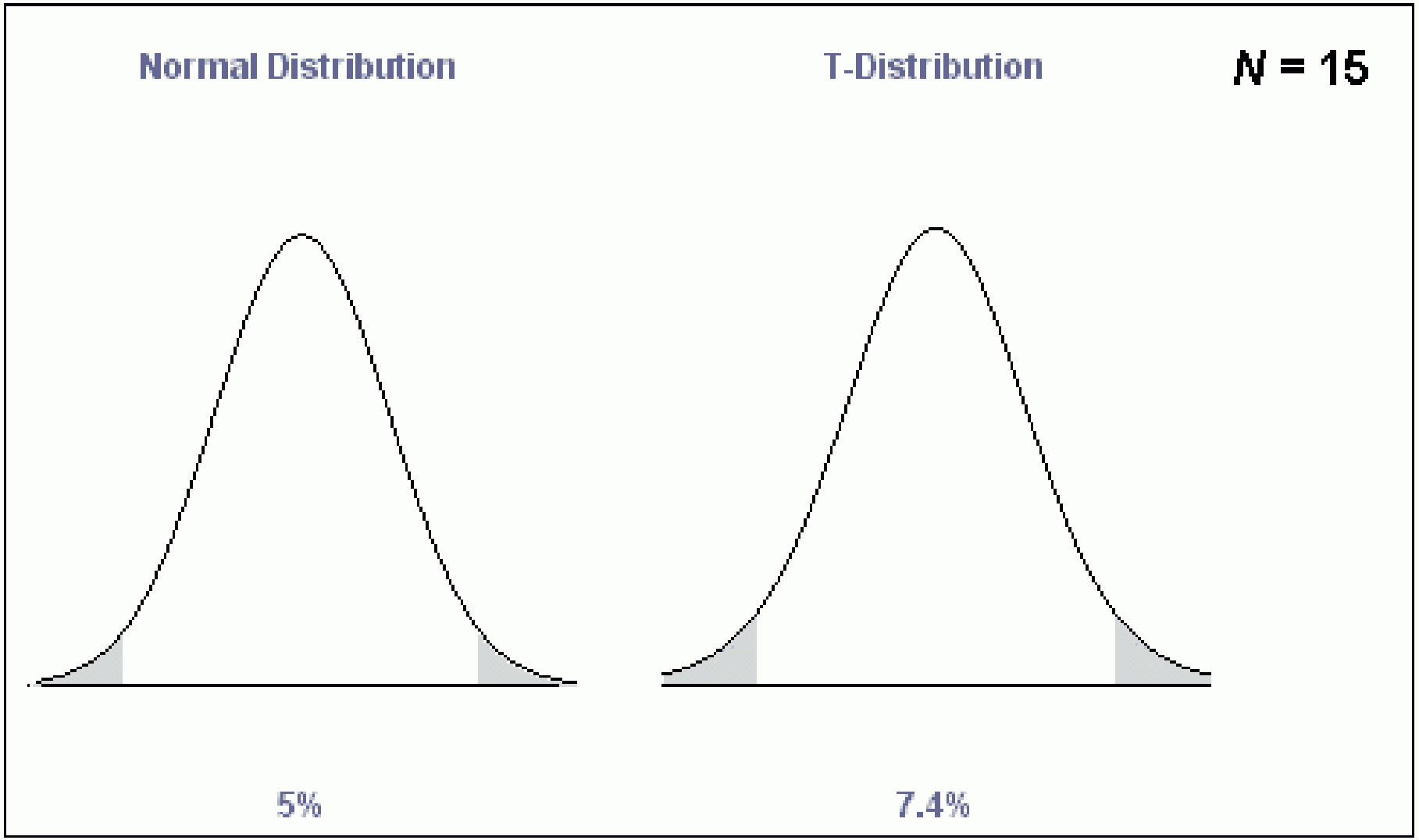
Оценка параметров популяции

t-распределение (Стьюдента)



При больших (>30) размерах выборок приближается к **нормальному**

Оценка параметров популяции



У него больше площадь «хвостов», и доверительный интервал для заданного уровня достоверности получится ШИРЕ, чем при нормальном распределении

Оценка параметров популяции

Выборочное среднее – точечная оценка (point estimation) среднего значения в популяции.

Доверительный интервал – интервальная оценка (interval estimate) среднего значения в популяции.

Точечные и интервальные оценки получают для самых разных параметров и это важнейший раздел анализа данных!

Для публикаций

✓ иногда стандартную ошибку среднего приводят как показатель разброса в выборке (±SE); это не очень корректно, т.к. это характеристика не выборки, а выборки выборочных средних;

✓ зато в публикациях нередко используют доверительный интервал (95% CI), ведь он показывает местонахождение среднего в популяции: если мы многократно поучаем из популяции выборки заданного размера, в 95% этих выборок попадёт среднее ***μ***.

«…mean risk was 1.13, 95% CI = 0.99 to 1.29»
«… the univariate RRs for triglyceride were 1.32 (95% CI 1.26–1.39)»

Есть и другие подходы к оценке среднего в ГС

Метод максимального правдоподобия (maximum likelihood, ML)

✓Мы представляли параметр (среднее значение) как неизвестное, но вполне постоянное и определённое значение.

✓Теперь представим, что мы имеем дело с целым распределением этих параметров.

✓Всего-навсего из них нужно выбрать тот, который максимизирует правдоподобие (likelihood) наших данных (шансы получить такую выборку, как у нас).

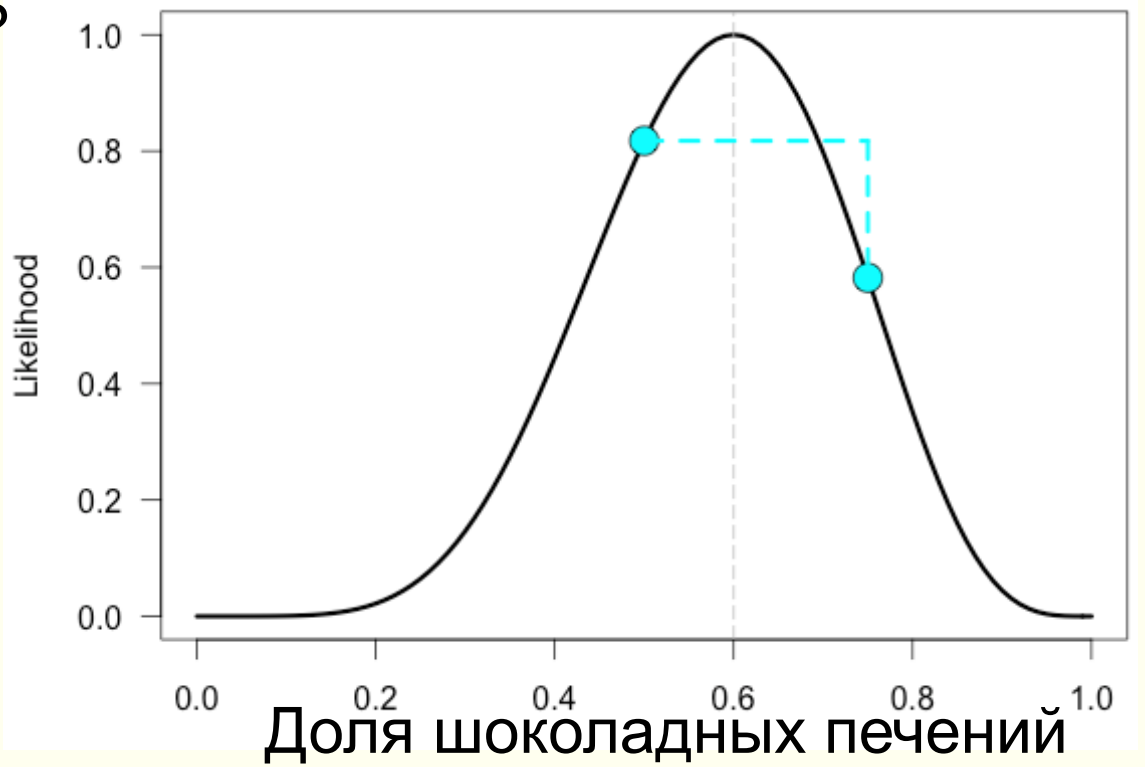
✓То есть, наши данные (выборка) – постоянны, меняются возможные параметры.

Оценка параметров популяции

ML: представим несколько коробок с печеньями – шоколадными и ванильными (N печений в коробках одинаково). Вытащим наугад 10 печений. Получилось 6 шоколадных и 4 ванильных - выборка.

Какой может быть доля шоколадных печений в коробках (популяционный параметр)?

Отложим по одной оси возможные значения параметра, а по другой – шансы получить такое соотношение печений, как у нас (likelihood), при разных значениях параметра.
Это – likelihood function.



Оценка нашего параметра – то значение, при котором функция достигает максимума.

Resampling methods: bootstrap and jackknife.

Bootstrap: много раз (1000 и более) мы случайным образом выбираем из нашей выборки наблюдения так, чтобы каждый раз получалась новая выборка такого же размера, как и исходная (наблюдения могут повторяться), и считаем каждый раз значение нужного показателя (напр., среднее). Среднее этих показателей и будет бутстреп-оценкой искомого параметра.

Jackknife – считаем нужный показатель в выборке, поочерёдно исключая одно из значений, на основе этих чисел оцениваем параметр. Менее точный и популярный метод.

Оценка параметров популяции

Байесовский подход (baesian approach)

Основывается на оценке максимального правдоподобия, но ещё учитывает наши предварительные знания о популяции.

Теорема Байеса