Статистические тесты

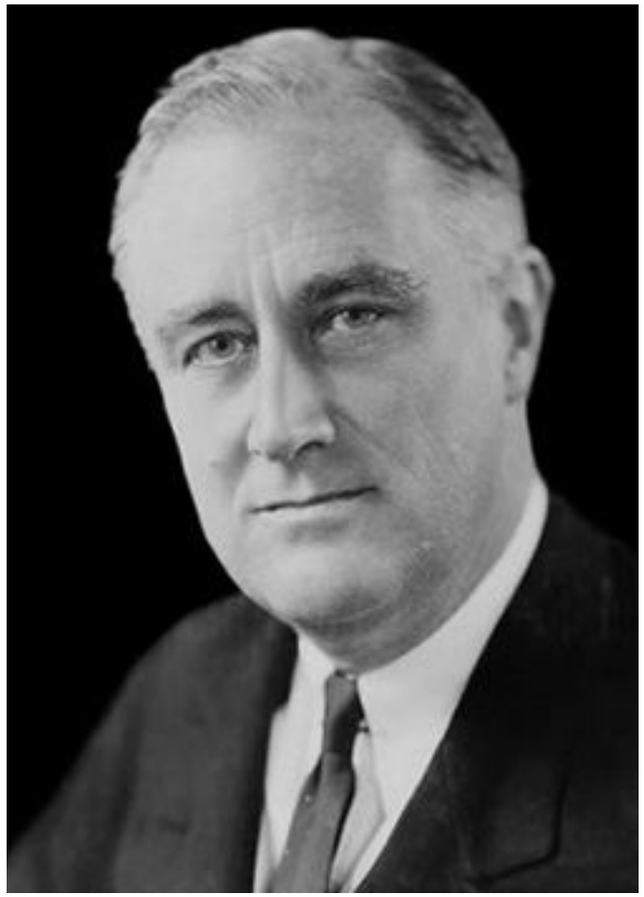
Понятие выборки

Генеральная совокупность (в англ. — *population*) — совокупность всех объектов (единиц), относительно которых учёный намерен делать выводы при изучении конкретной проблемы.

Выборка или **выборочная совокупность** — множество случаев (испытуемых, объектов, событий, образцов), с помощью определённой процедуры выбранных из генеральной совокупности для участия в исследовании.

Репрезентативность - выборка может рассматриваться в качестве репрезентативной или нерепрезентативной

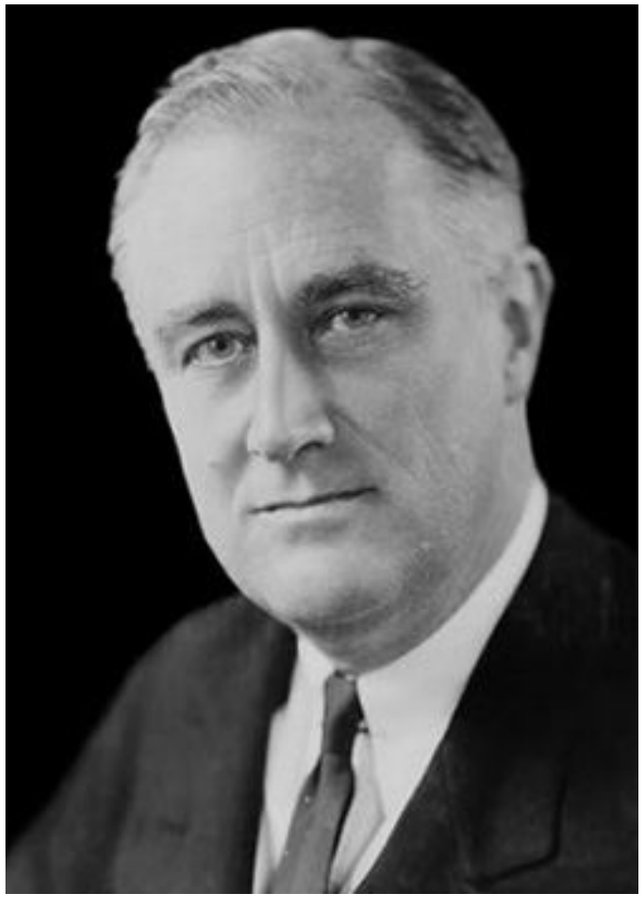
Примеры нерепрезентативных выборок



Предсказание результатов выборов «Литрери Дайджест»

Предсказали победу Альфа Лондона

Примеры нерепрезентативных выборок



Предсказание результатов выборов «Литрери Дайджест»

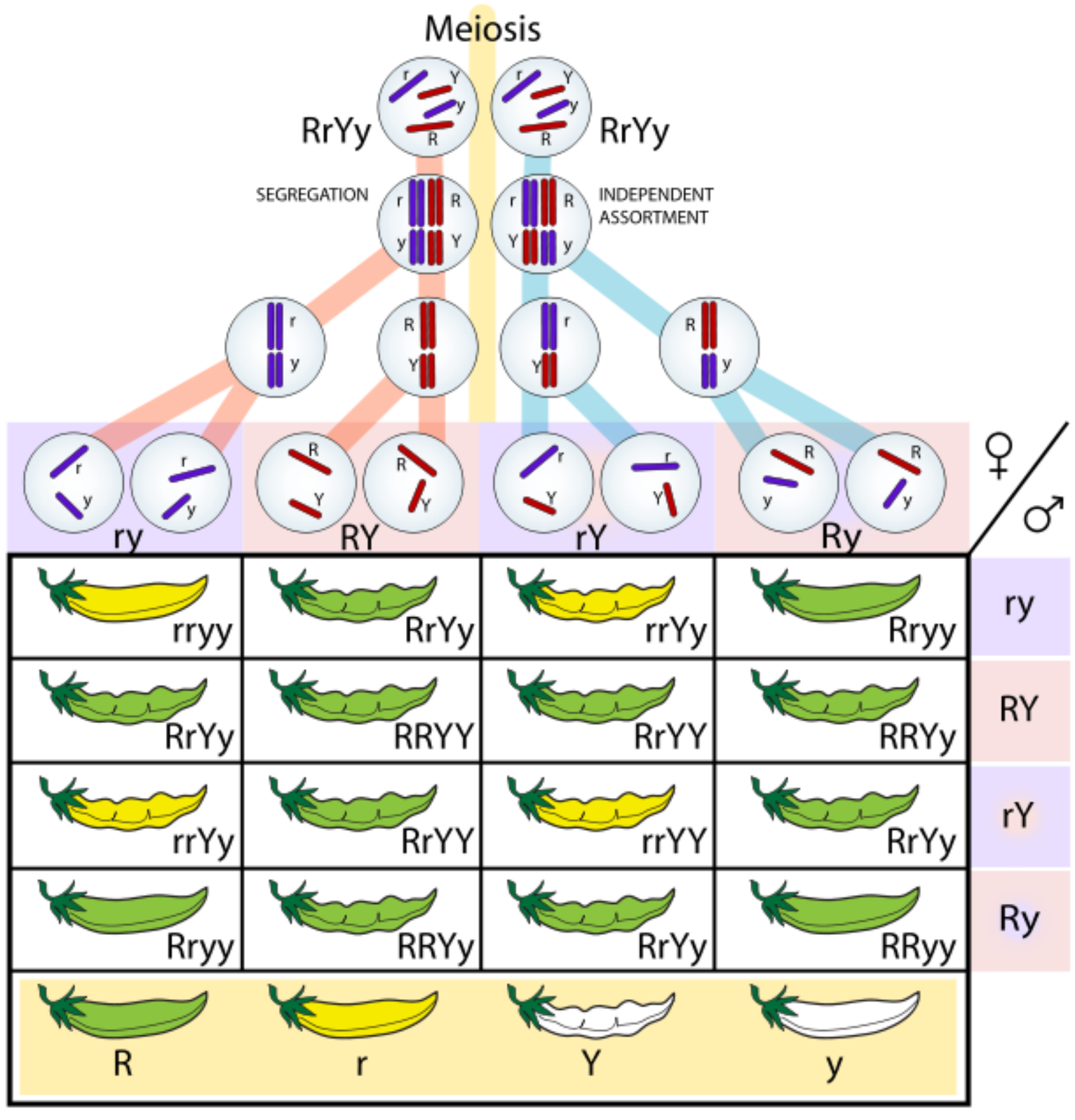
Предсказали победу Альфа Лондона
Победил Франклин Рузвельт

Ошибка «Литрери Дайджест» заключалась в следующем: желая увеличить репрезентативность выборки, — так как им было известно, что большинство их подписчиков считают себя республиканцами, — они расширили выборку за счет людей, выбранных из телефонных книг и регистрационных списков автомобилей. Однако они не учли современных им реалий и в действительности набрали еще больше республиканцев: во время Великой депрессии обладать телефонами и автомобилями могли себе позволить в основном представители среднего и верхнего класса (то есть большинство республиканцев, а не демократов)

Тренировка

1. Необходимо собрать репрезентативную выборку людей по весу и по размеру используемой посуды;
2. Необходимо собрать выборку людей по политическому спектру (левые-правые);
3. Необходимо собрать выборку птиц по месту гнездования;
4. Необходимо собрать выборку людей по их отношению к курению;
5. Необходимо собрать выборку людей по употреблению ими легких наркотиков.

Законы Менделя



Boxplot

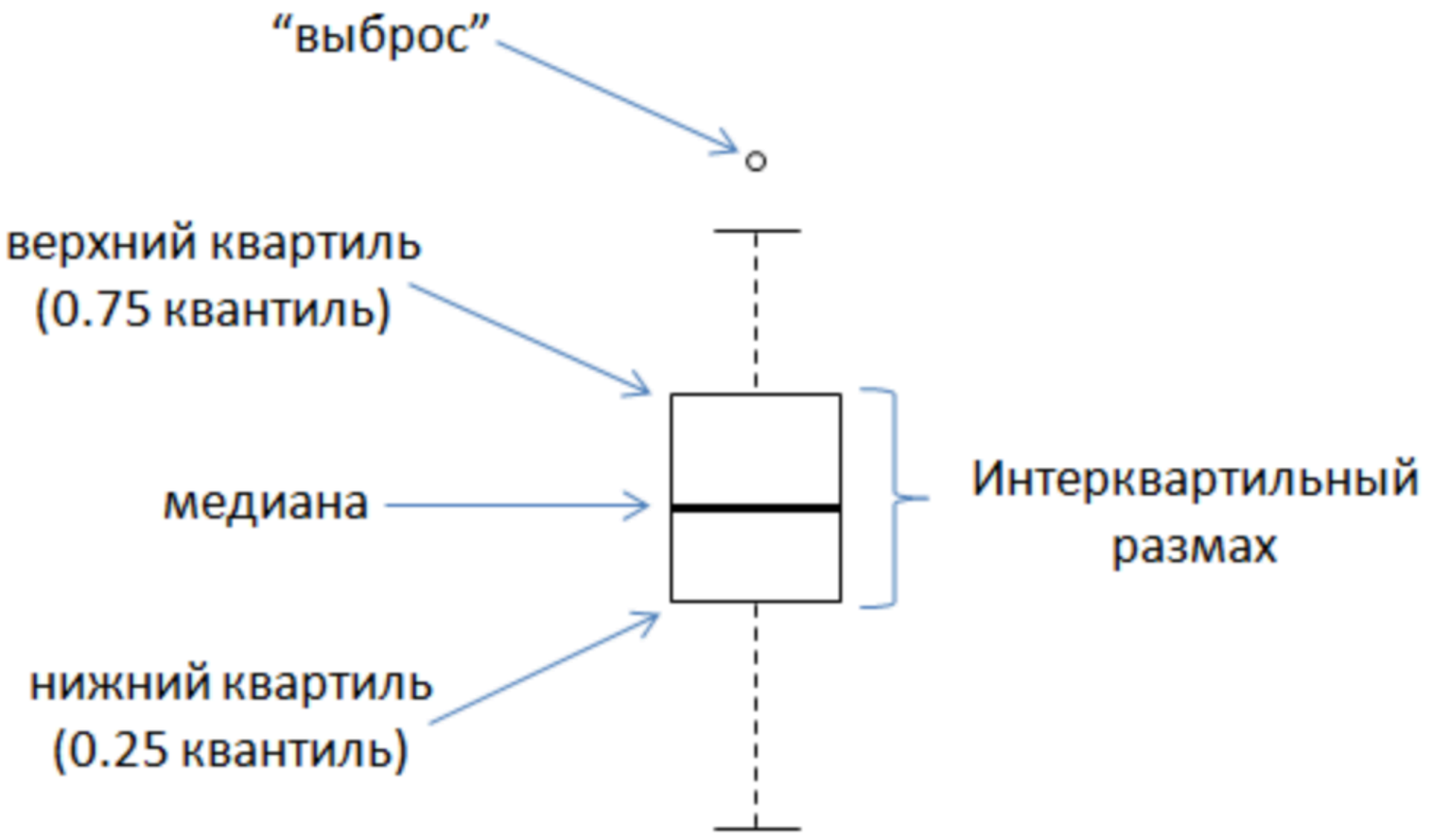


Рисунок 1. Строение диаграммы размахов.

Диаграмма Кливленда

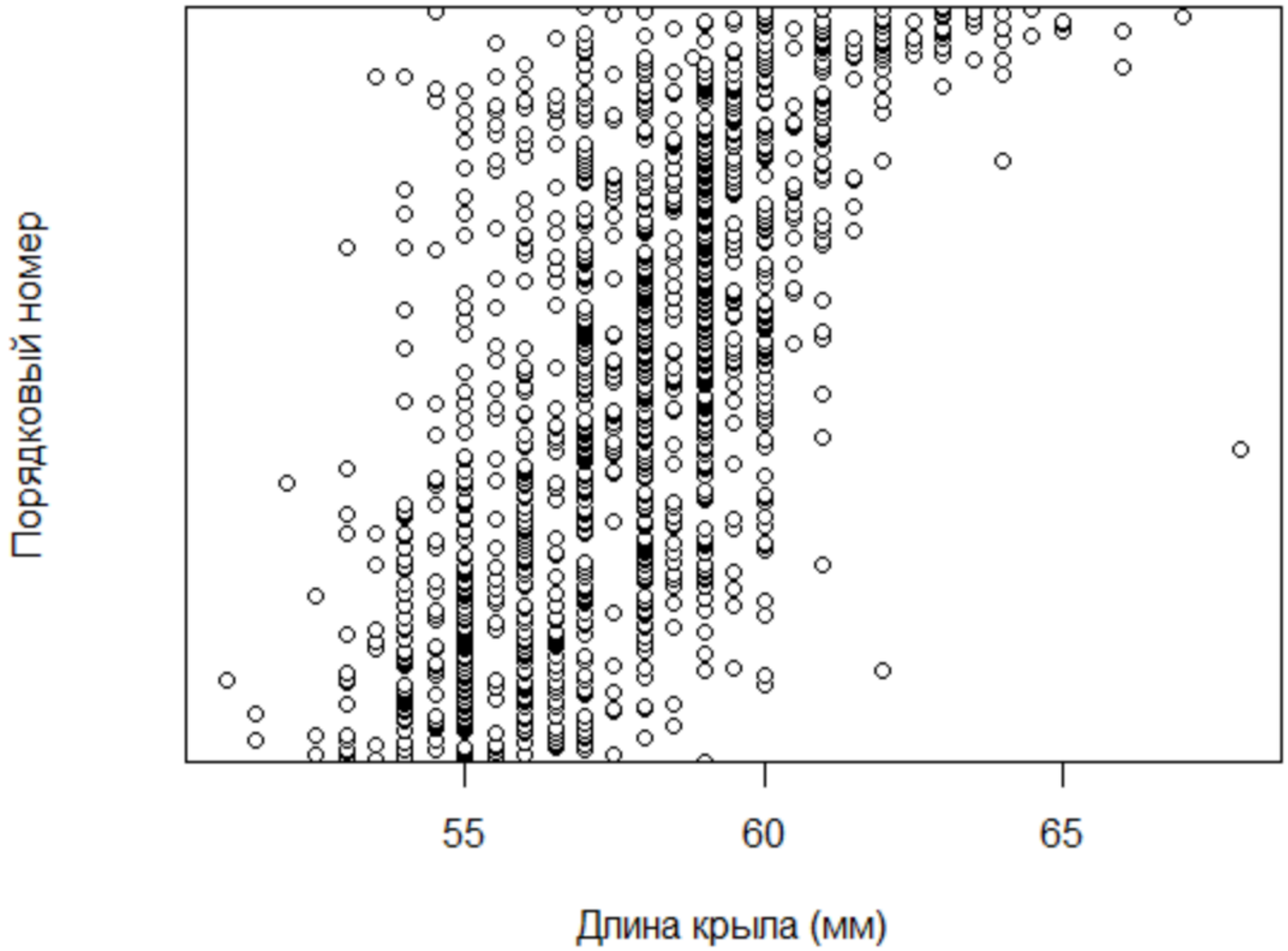


Рисунок 2. Точечная диаграмма Кливленда, изображающая данные о длине крыла у 1295 воробьев (Zuur et al. 2010). В этом примере данные предварительно были упорядочены в соответствии с весом птиц, и поэтому облако точек имеет примерно S-образную форму.

Выбрасывать ли “выбросы”?

Выбрасывать ли “выбросы”?

1) Получаем более робастные статистики и оценки параметров

Выбрасывать ли “выбросы”?

1) Получаем более робастные статистики и оценки параметров

2) Уменьшаем статзначимость в случае малого числа наблюдений

Выбрасывать ли “выбросы”?

1) Получаем более робастные статистики и оценки параметров

2) Уменьшаем статзначимость в случае малого числа наблюдений

3) Можем просто не учесть значения в областях, которые нам интересны, но данных в них было мало и мы решили, что они выбросы

Убираем коллинеарность

Под *коллинеарностью* (англ. *collinearity*) понимают наличие линейной зависимости между двумя предикторами. В задачах с несколькими предикторами (например, при выполнении *множественного регрессионного анализа*) говорят также о *мультиколлинеарности* (англ. *multicollinearity*), т.е. наличии линейной зависимости между несколькими переменными.

Такие переменные могут значимо навредить вашему исследованию (большое число таких переменных приводит к неустойчивости любой построенной модели и т.д)

Как с этим бороться? Узнаете на лекции про линейные модели

Тестирование гипотез

Нулевая гипотеза (H0) – это *о*сновное проверяемое предположение, которое обычно формулируется как отсутствие различий, отсутствие влияние фактора, отсутствие эффекта, равенство нулю значений выборочных характеристик и т.п. Примером нулевой гипотезы в педагогике является утверждение о том, что различие в результатах выполнения двумя группами учащихся одной и той же контрольной работы вызвано лишь случайными причинами.

Другое проверяемое предположение (не всегда строго противоположное или обратное первому) называется **конкурирующей** или **альтернативной** гипотезой (**H1**). Обычно она соответствует предположению, что мы нашли значимое воздействие какого-то фактора

Тренировка

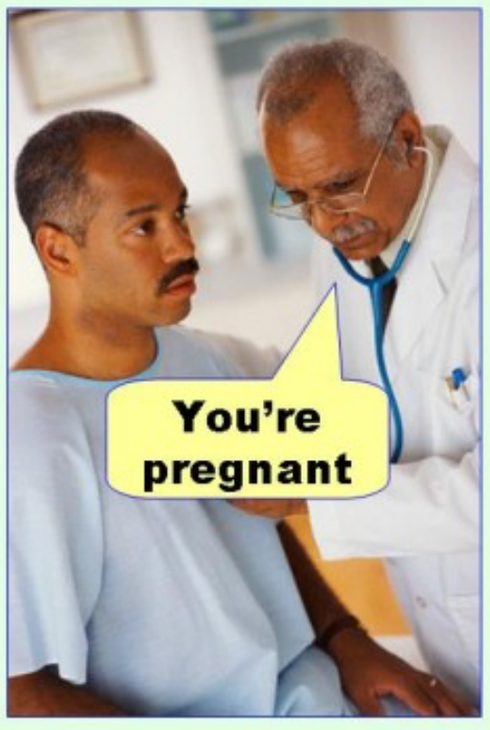
Сформулируйте нулевую и альтернативную гипотезу в данных случаях:

1) Есть две группы студентов, одна приходит всегда раньше первой и слушает лекцию раньше. Вторая слушает ту же лекцию, но позже. Известны оценки первой и второй группы студентов;
2) Есть две группы людей, одни в числе близких друзей имеют людей с повышенным весом, вторые нет. Известны веса людей в первой и второй группах;
3) Есть две группы студентов. Одна работает весь семестр, вторая учит предмет в последнюю неделю. Известны оценки по этому предмету для обеих групп;
4) Есть две группы людей. Одни слушают смотрят телеканал “Домашний” в течении не менее одного часа, вторые - нет. Известны результаты теста (в баллах) на критическое мышление для обеих групп;
5) Есть два чата в Телеграм - флудящий и нефлудящий. Известно количество годноты, кинутой в данный чат в течении месяца (расписано по дням);
6) Есть две группы программистов - пишущие только на Python, и пишущие только на C/C++ . Известна статистика самоубийств среди первых и вторых в течении года;
7) Есть две группы людей - одна ест всю еду палочками, вторая есть наиболее удобным способом. В течении месяца с помощью тестов (дающих число в заданном диапазоне) измерялся уровень агрессивности данных людей.

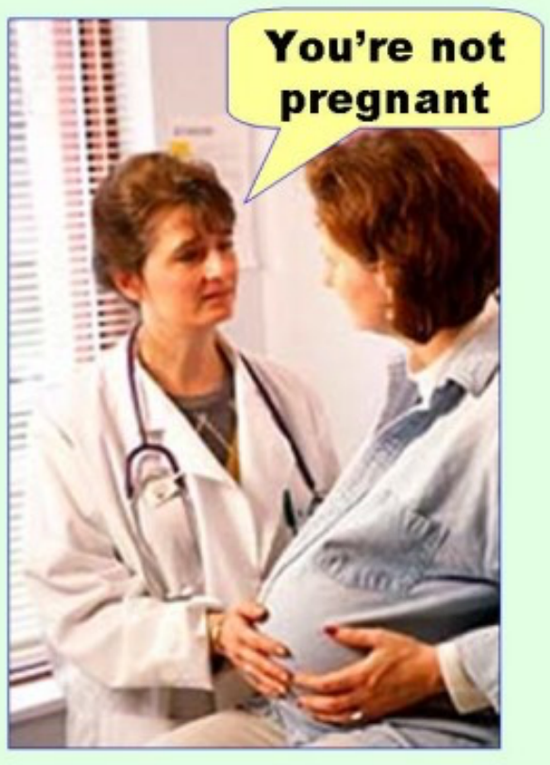
Ошибки первого и второго рода

H_0	верная	ложная
отклоняется	ошибка первого рода	решение верное
не отклоняется	решение верное	ошибка второго рода

Type I error
(false positive)



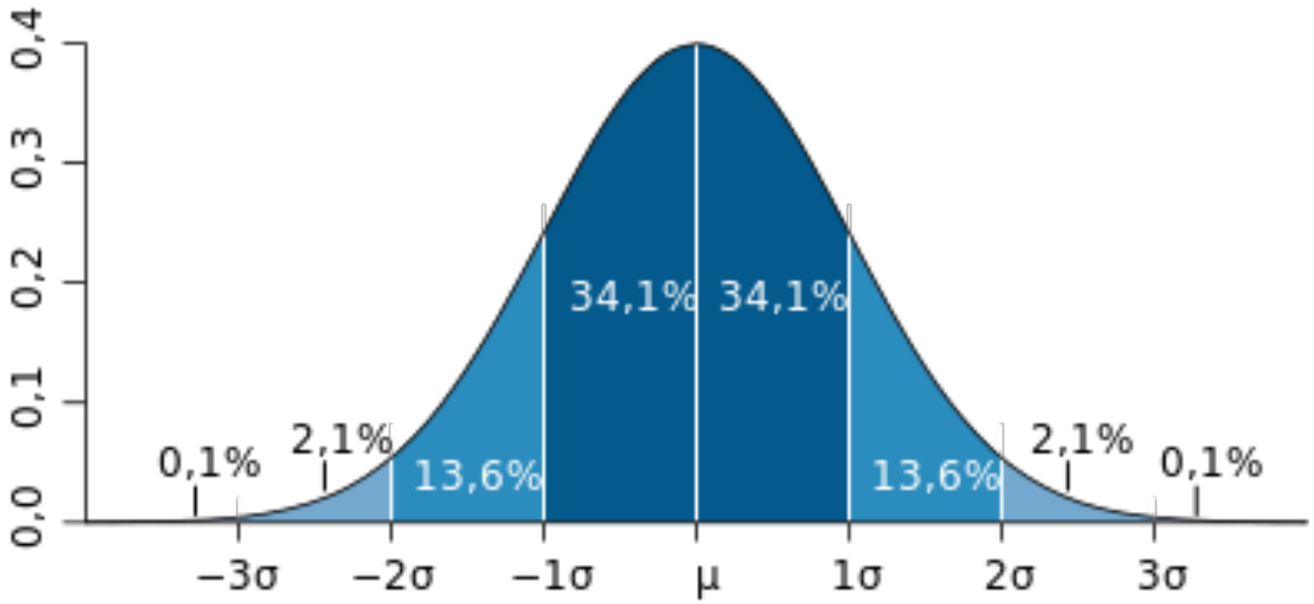
Type II error
(false negative)



Тесты

1) Параметрические

Критерий различия называют ***параметрическим,*** если он основан на конкретном типе распределения генеральной совокупности (как правило, нормальном) или использует параметры этой совокупности (средние, дисперсии и т.д.).



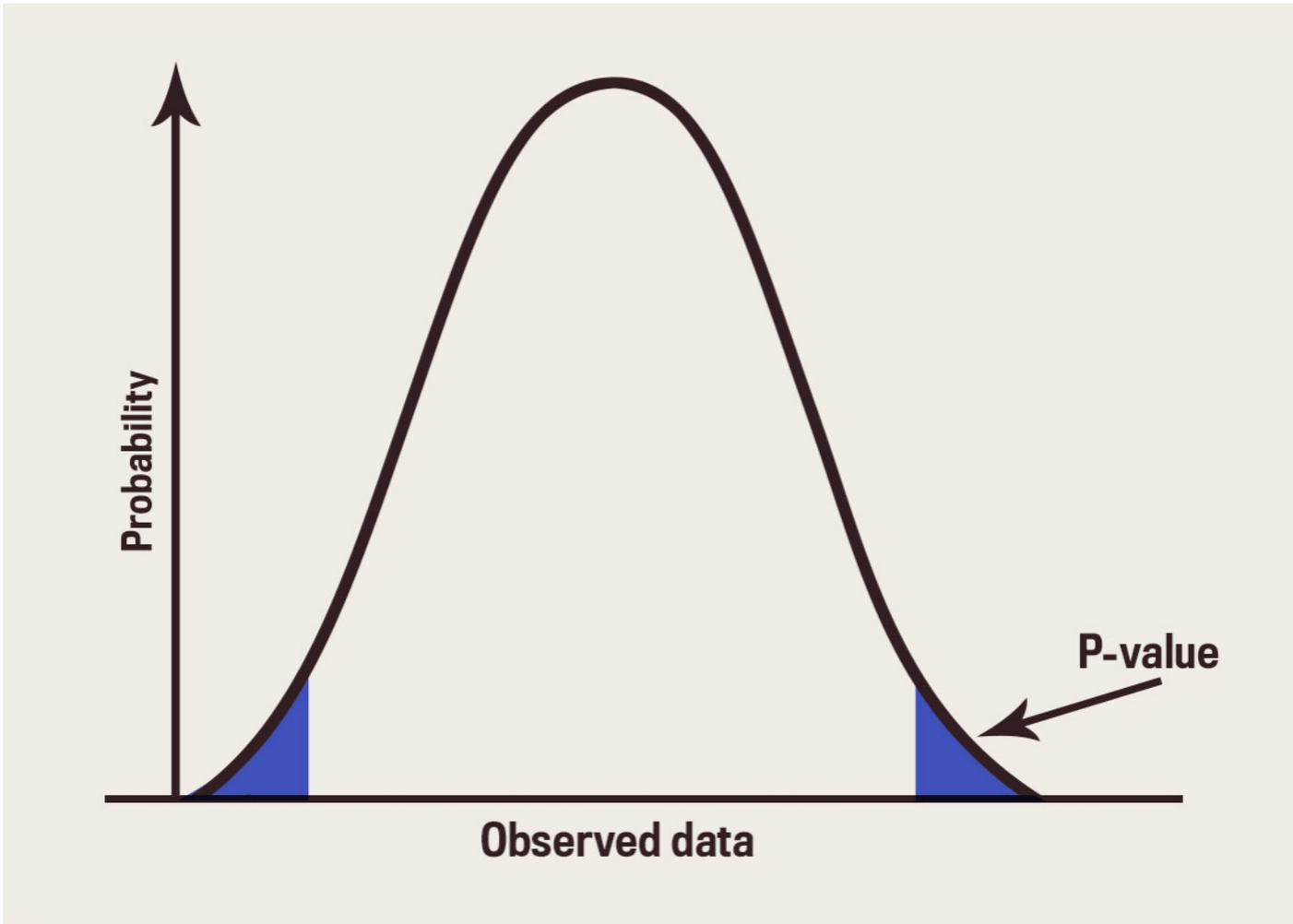
2) Непараметрические

Критерий различия называют ***непараметрическим,*** если он не базируется на предположении о типе распределения генеральной совокупности и не использует параметры этой совокупности. Поэтому для непараметрических критериев предлагается также использовать такой термин как «критерий, свободный от распределения».

P-value

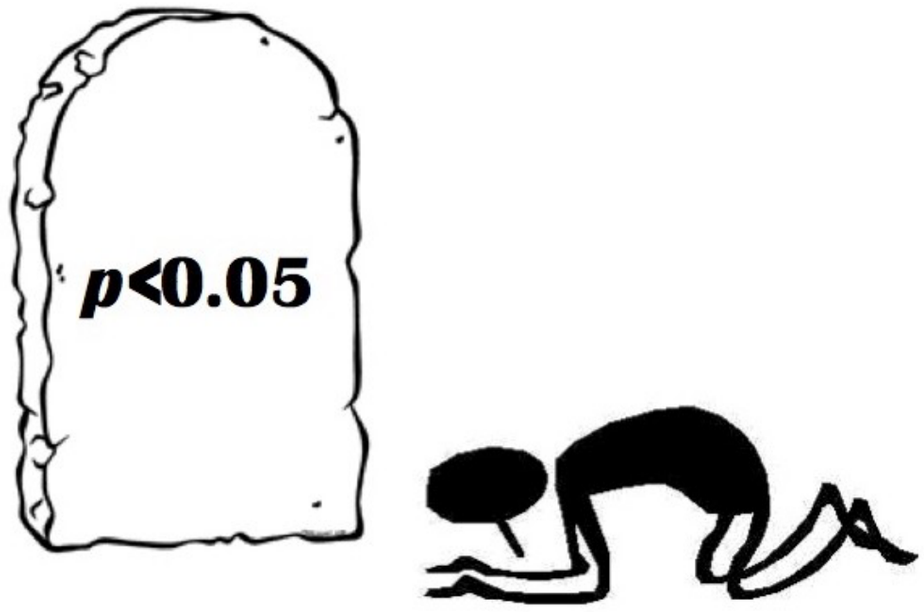
P-value - вероятность получить результат как минимум такой же критический как тот, что мы наблюдаем, считая, что нулевая гипотеза является правильной

Другими словами - если нулевая гипотеза верна, то насколько вероятно получить ту выборку, которую мы получили



За что надо бить по пальцам молотком

- Малое значение p-value (**≤ 0.05**) свидетельствует в в пользу отвердения нами нулевой гипотезы
- Большое значение p-value означает слабую значимость доказательств неправильности нулевой гипотезы, потому мы не можем отвергнуть нулевую гипотезу



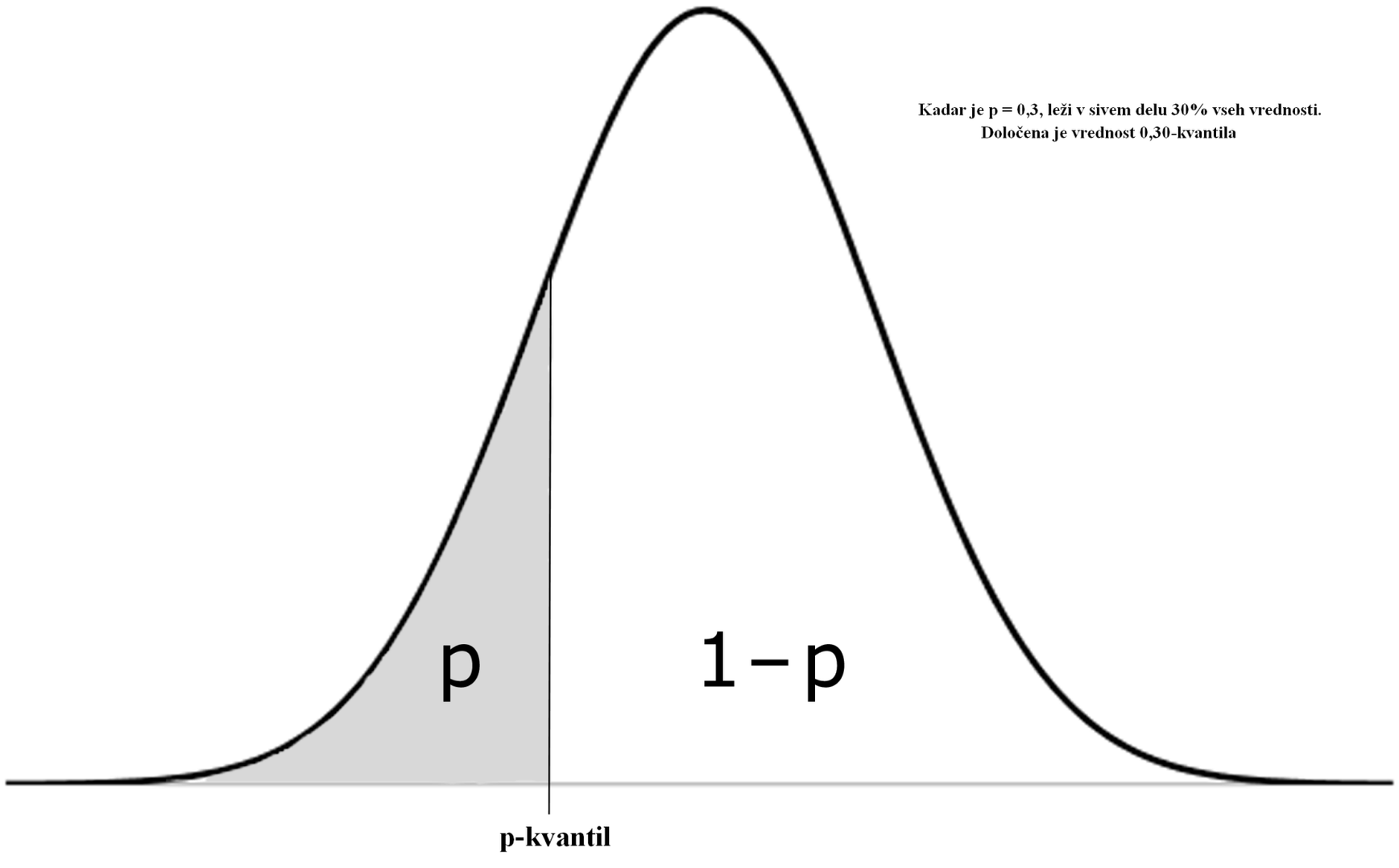
Уровень значимости

Уровень значимости статистического теста alpha - допустимая для данной задачи вероятность ошибки первого рода (то есть вероятность отклонить нулевую гипотезу, когда она на самом деле верна)

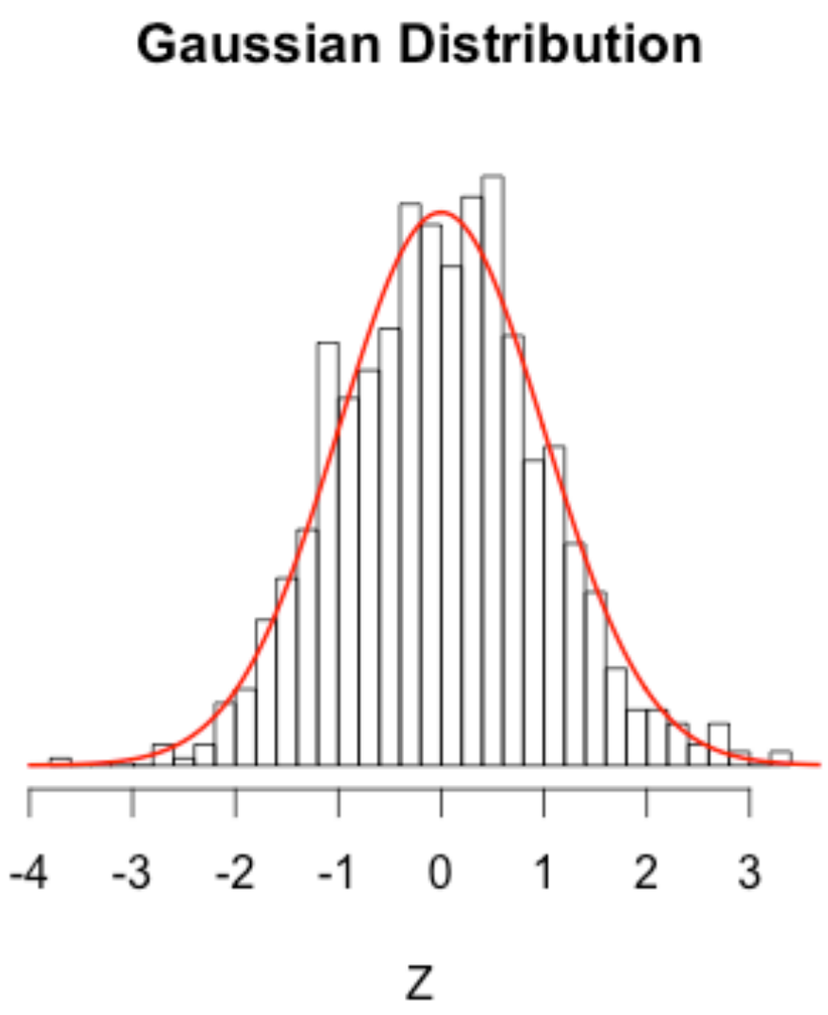
- Малое значение p-value (≤ **alpha**) свидетельствует в пользу отвержения нами нулевой гипотезы
- Большое значение p-value означает слабую значимость доказательств неправильности нулевой гипотезы, потому мы не можем отвергнуть нулевую гипотезу

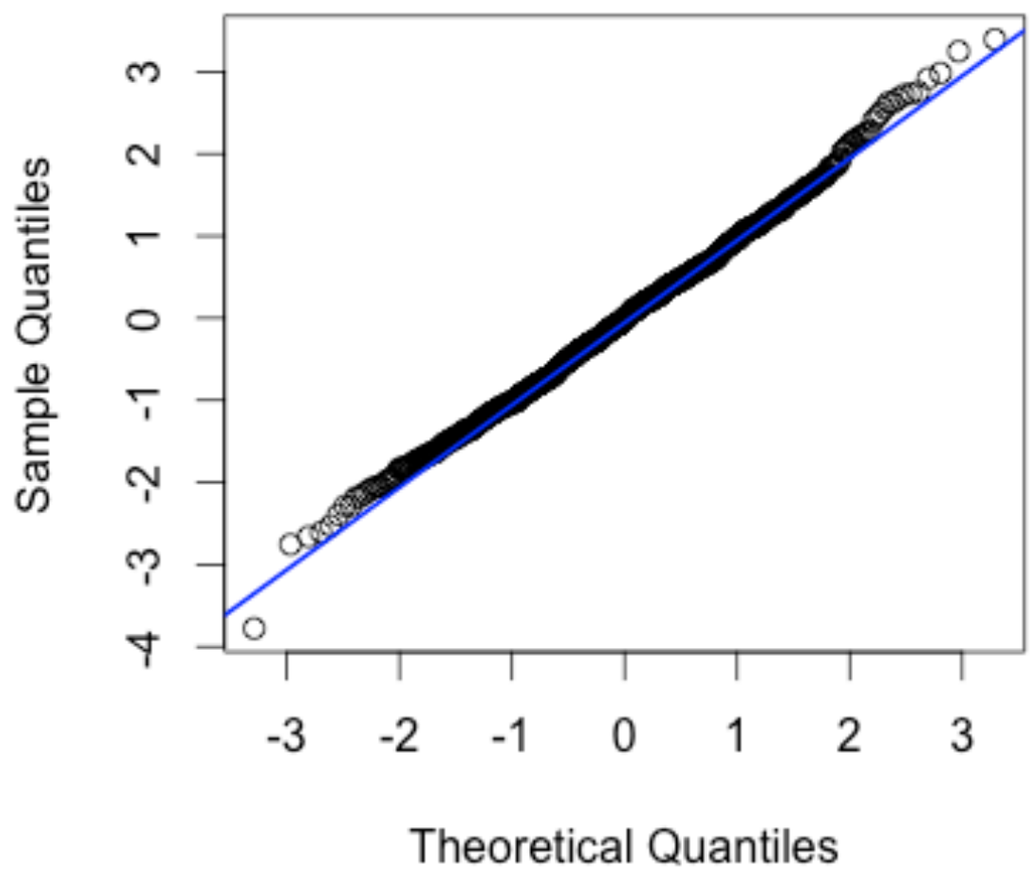
Квантили

Квантиль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

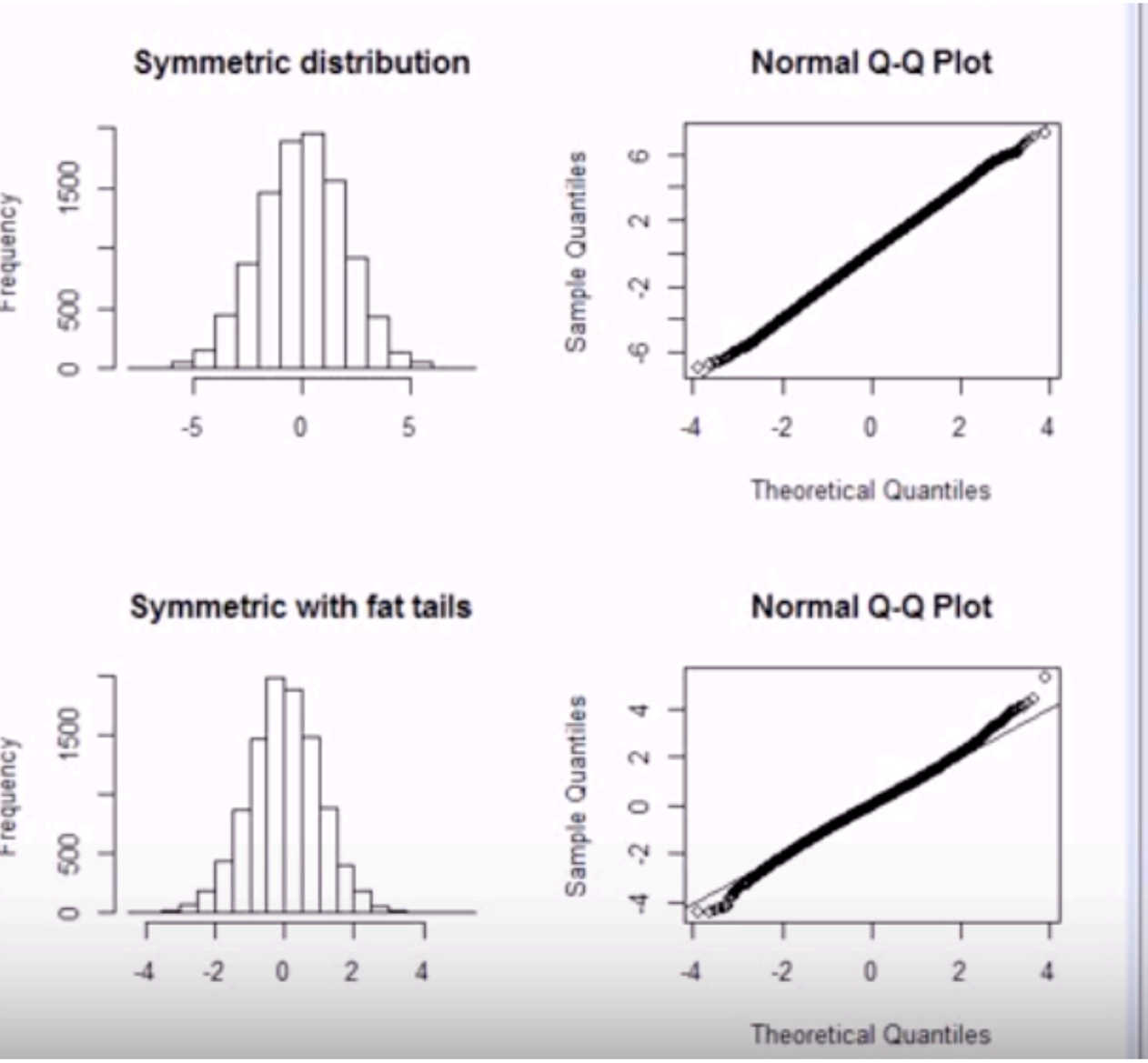


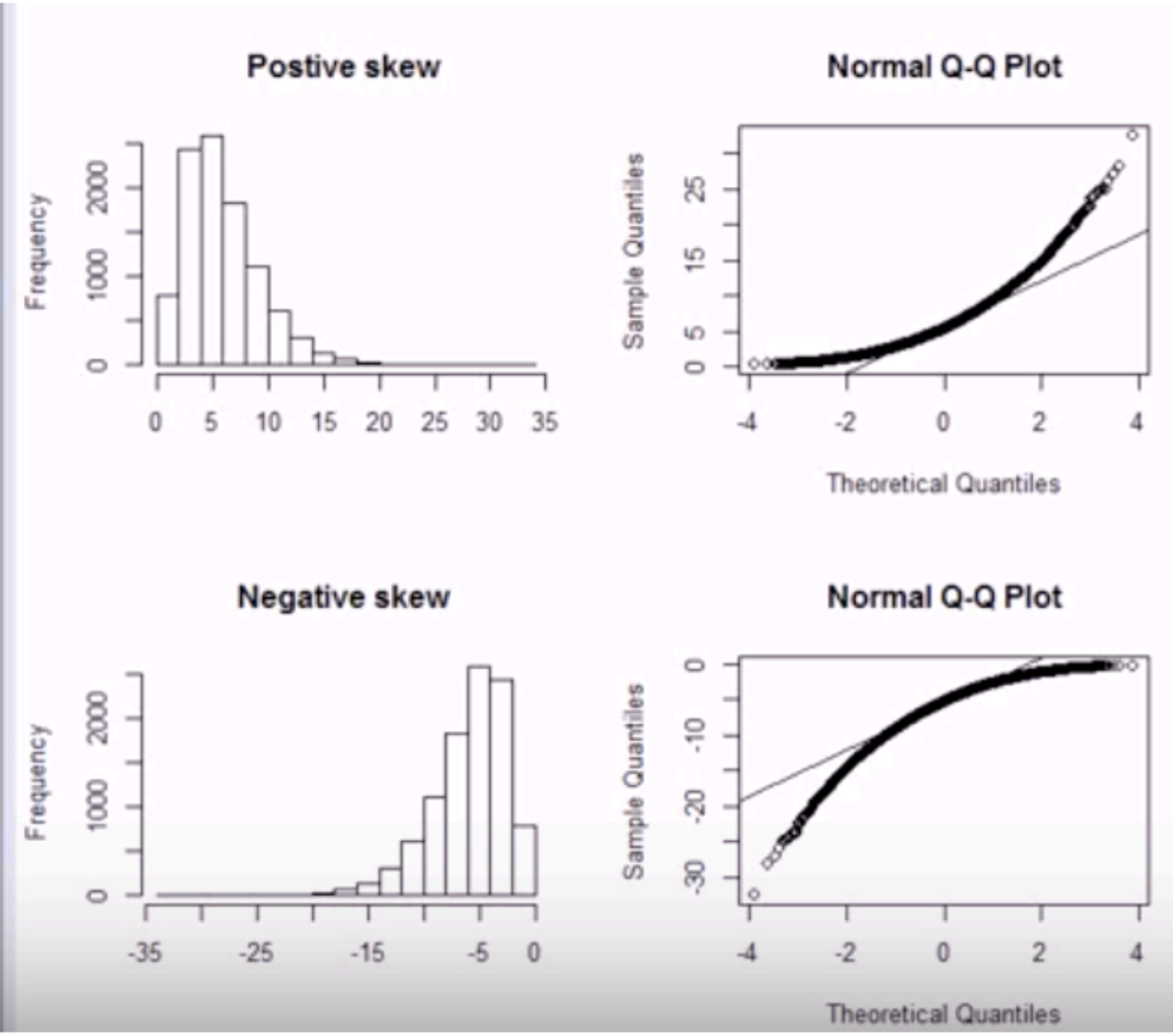
Проверка выборки на нормальность



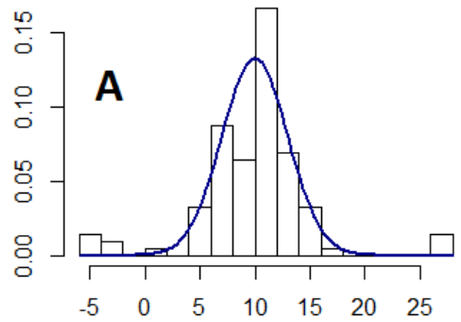


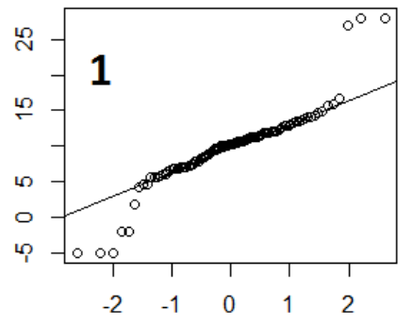
Q-Q Plot

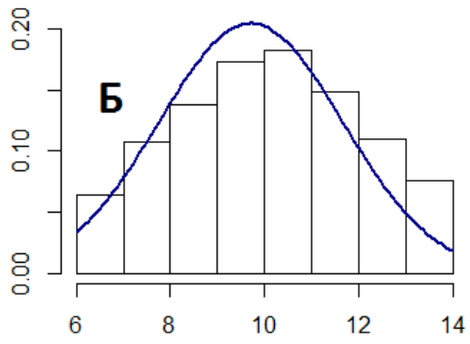

Symmetric distribution
Frequency
0
500
1500
-5
0
5
Normal Q-Q Plot
Sample Quantiles
-6
-2
2
6
-4
-2
0
2
4
Theoretical Quantiles
Symmetric with fat tails
Frequency
0
500
1500
-4
-2
0
2
4
Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
-4
-2
0
2
4
Theoretical Quantiles

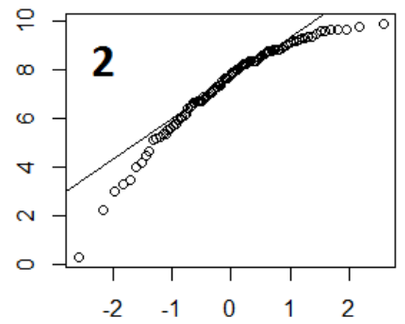

Postive skew
Frequency
0
1000
2000
0
5
10
15
20
25
30
35
Normal Q-Q Plot
Sample Quantiles
0
5
15
25
-4
-2
0
2
4
Theoretical Quantiles
Negative skew
Frequency
0
1000
2000
-35
-25
-15
-5
0
Normal Q-Q Plot
Sample Quantiles
-30
-20
-10
0
-4
-2
0
2
4
Theoretical Quantiles

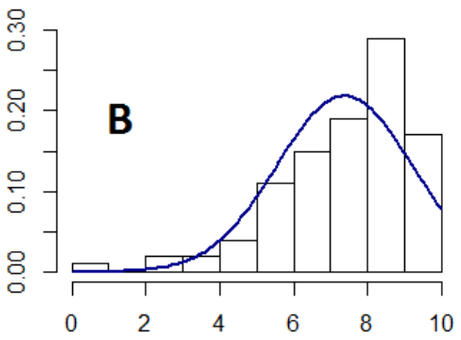
Тренировка

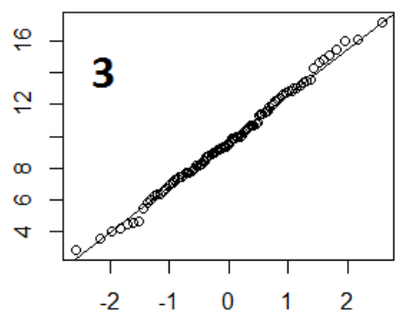

А
0.00
0.05
0.10
0.15
-5
0
5
10
15
20
25

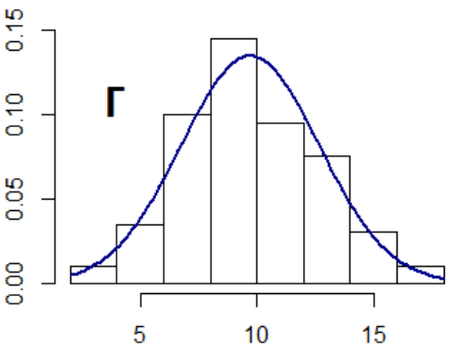

1
-5
0
5
15
25
-2
-1
0
1
2

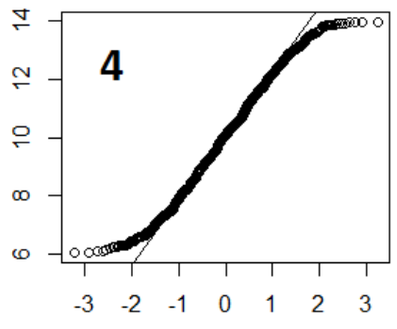

Б
0.00
0.10
0.20
6
8
10
12
14


2
0
2
4
6
8
10
-2
-1
0
1
2


В
0.00
0.10
0.20
0.30
0
2
4
6
8
10


3
4
6
8
12
16
-2
-1
0
1
2

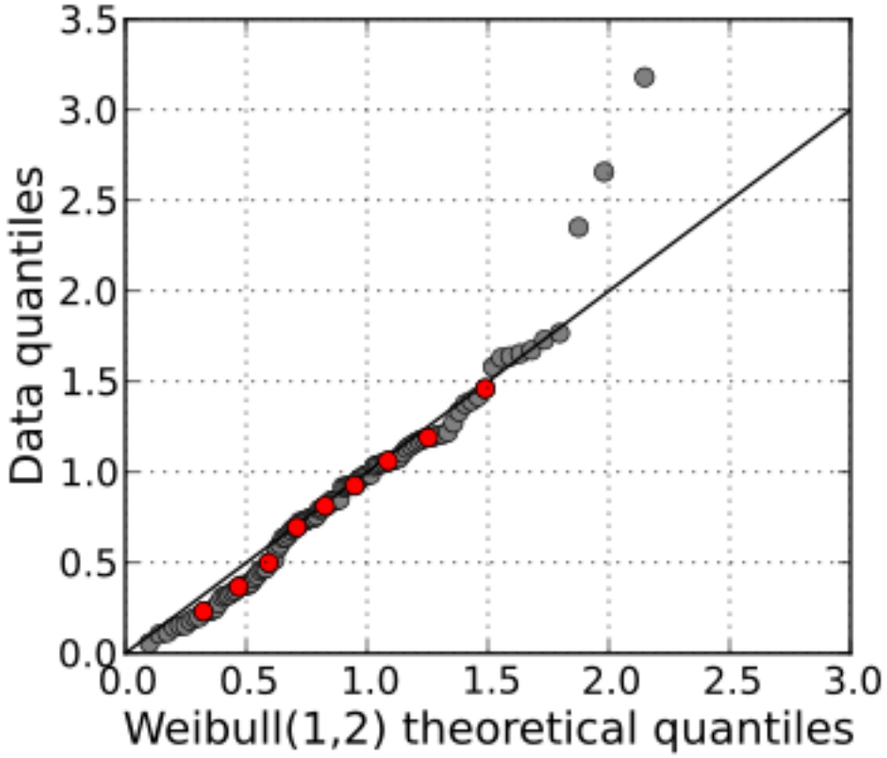

Г
0.00
0.05
0.10
0.15
5
10
15


4
6
8
10
12
14
-3
-2
-1
0
1
2
3

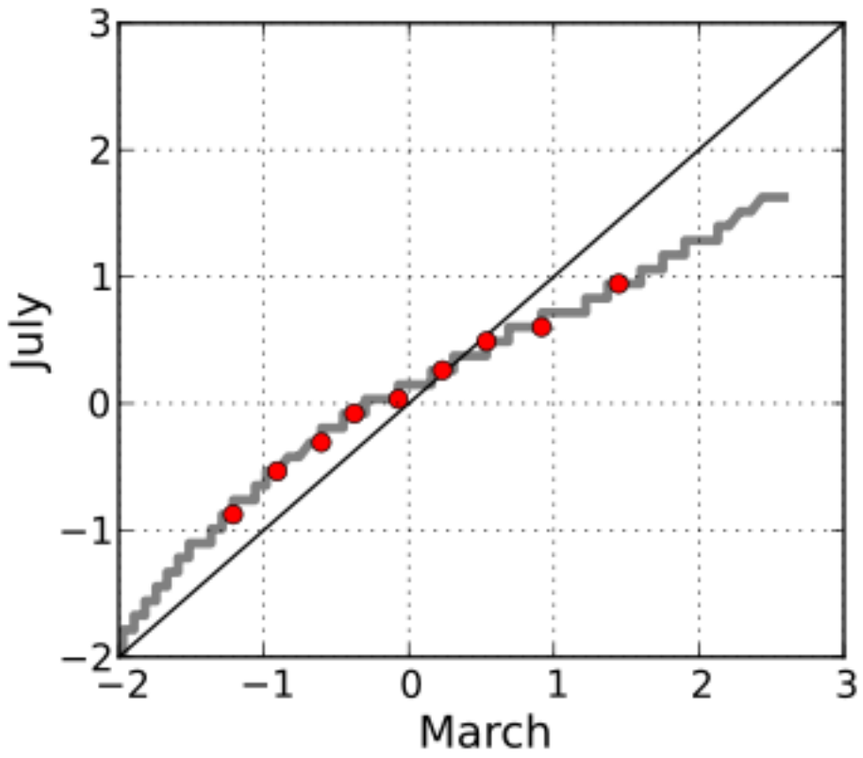
Q-Q Plot

Строго говоря, Q-Q Plot позволяет сравнить два любых распределения, хоть чаще всего и используют для проверки для нормальность.

Фактически, это первый непараметрический тест, который мы с вами узнали

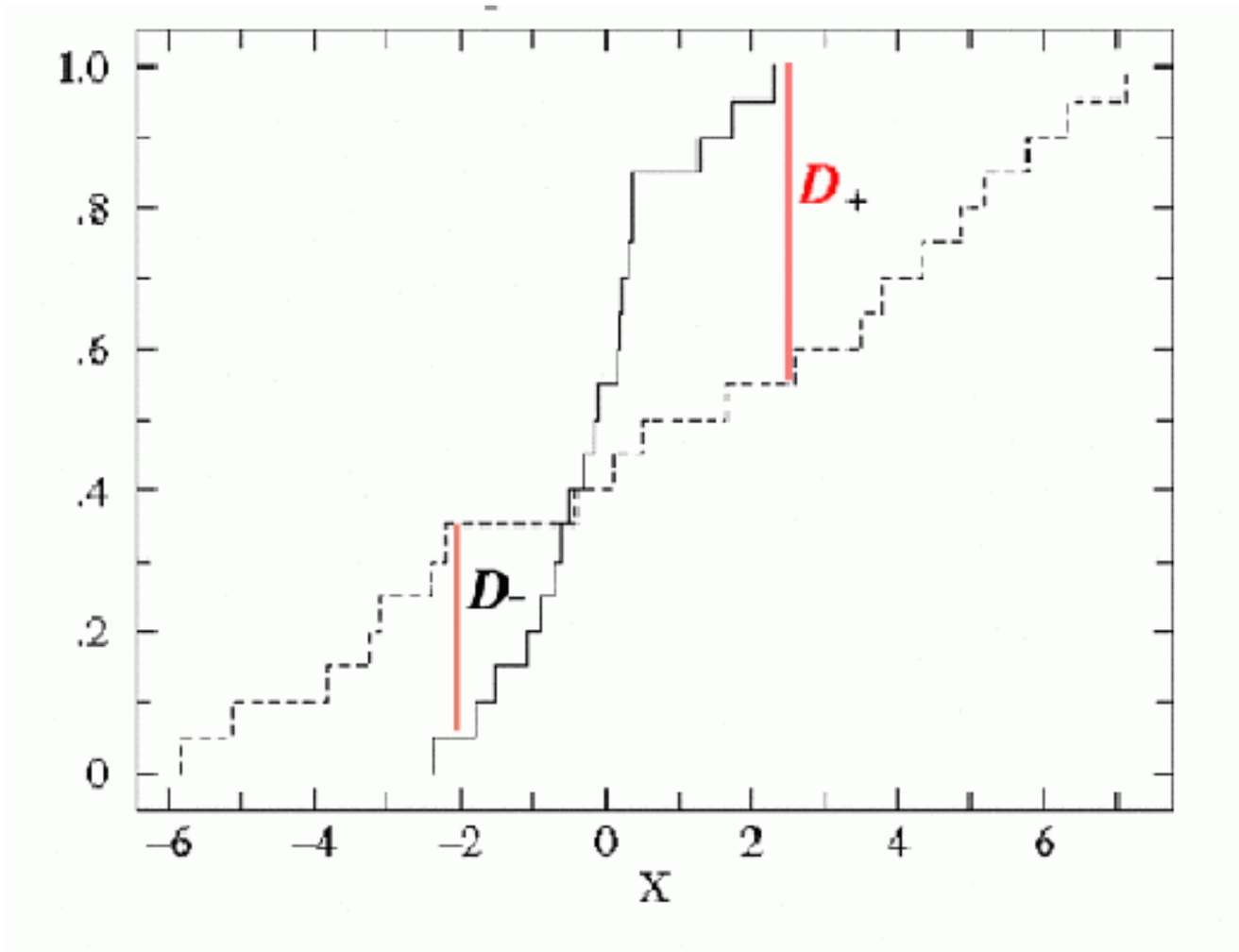


Можно взять другое теоретическое распределение



Можно взять другое теоретическое распределение

Колмогоров-Смирнов



$$D_n = \sup_x |F_n(x) - F(x)|.$$

Тест Шапиро-Уилка

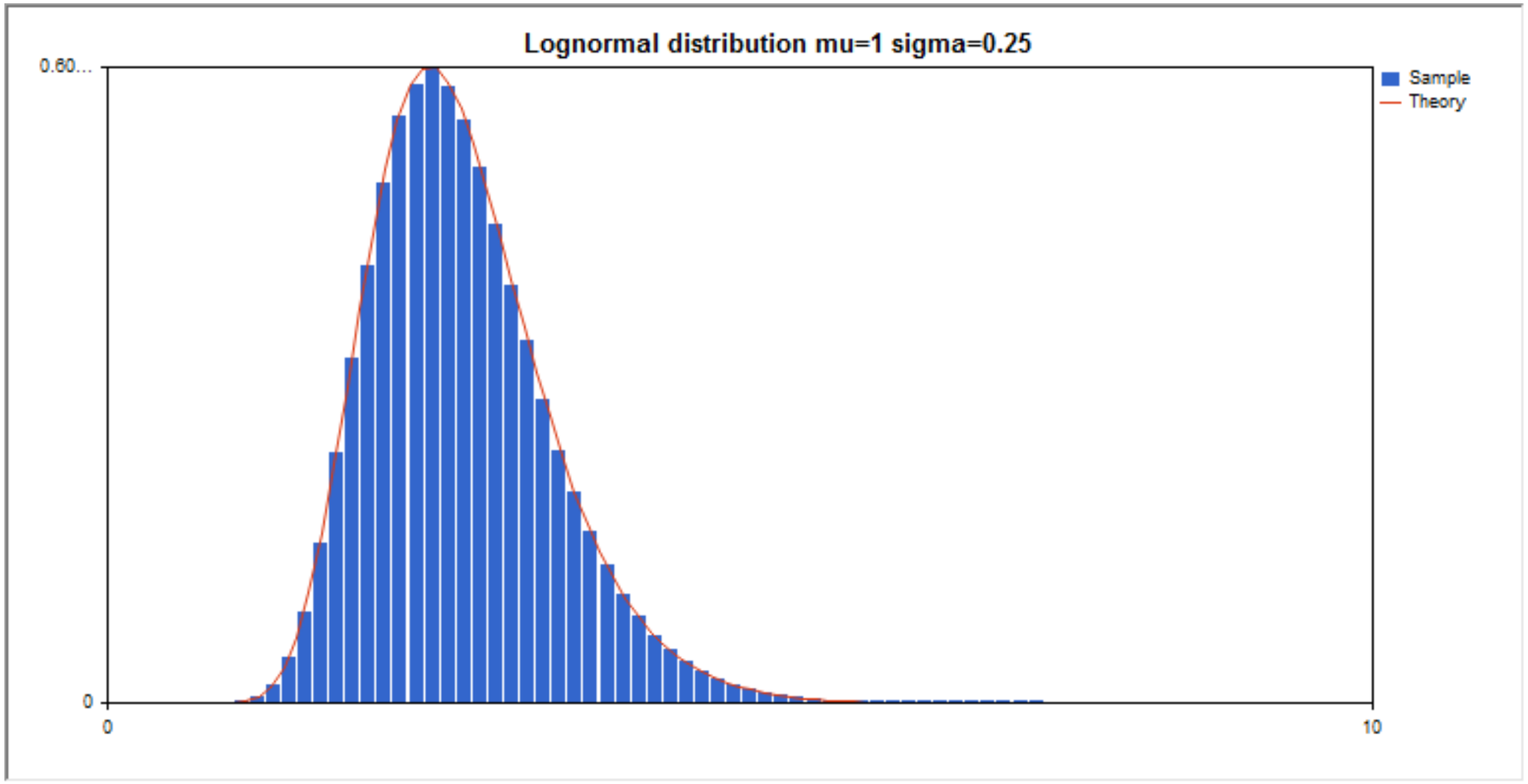
Специально для проверки нормальности распределения малых, численностью от трех до пятидесяти элементов

Очень строгий критерий, крайне склонен отвергать нормальность выборки, причем даже для данных, взятых из реального нормального распределения

Что значит, что наблюдаемый вами признак имеет нормальное распределение?

Указание: Само распределение возникает как результат сложения многих независимых случайных воздействий

Лог-нормальное распределение



Случайная величина имеет логарифмически нормальное распределение, если логарифм этой величины имеет нормальное распределение и она определена на положительной полуоси.
Логнормальное распределение часто используется в моделировании таких переменных, как персональные доходы, возраст новобрачных (точнее, первый раз вступающих в брак) или допустимое отклонение от стандарта вредных веществ в продуктах питания.

Непараметрические тесты

Исследователь ничего не знает о параметрах исследуемых совокупностей и виде их распределения: близки ли они к нормальному типу или какому-либо другому.

Соответственно, эти тесты не требуют этих ограничений от исследуемых признаков

Для их вычисления не требуется большого объема данных

Они являются более робастными (применимыми в широком диапазоне условий), чем их параметрические аналоги

Недостатки:

1) низкая статистическая мощность (менее чувствительные);

2) меньшая гибкость

Знаковый критерий

Sign test

The sign test is a method to find consistent *ordinal* differences between pairs of observations. It determines if one member in the pair of observations tends to be greater than the other member. Unlike t-test, there is no assumption of normality for small samples, neither any other assumption about the nature of the random variable.

- $H_0 : \text{median}_1 = \text{median}_2$
- $H_a : \text{median}_1 > \text{median}_2$

Sample (X_i, Y_i), $i = 1 \ldots n$

$\hat{p} =$ sample proportion of $X_i > Y_i$

Ties are split randomly between $X_i > Y_i$ and $X_i < Y_i$

Тест Манна-Уитни

Mann-Whitney U-test

Wilcoxon-Mann-Whitney test

- X and Y are two populations
- $H_0 : \mathrm{P}(X > Y) = \mathrm{P}(Y > X)$
- $H_a : \mathrm{P}(X > Y) \neq \mathrm{P}(Y > X)$
- U-statistic
  - $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ are two samples
  - Assign ranks to all the observations $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$
  - $R_1 =$ the sum of ranks for the observations which came from sample 1
  - $R_2 =$ the sum of ranks for the observations which came from sample 2
  - $U_1 = R_1 - \frac{n_1(n_1+1)}{2} \quad U_2 = R_2 - \frac{n_2(n_2+1)}{2}$
  - $U = \max\{U_1, U_2\}$
  - In case of ties there is a small correction to this procedure

Обобщение - Критерий Краскела — Уоллиса

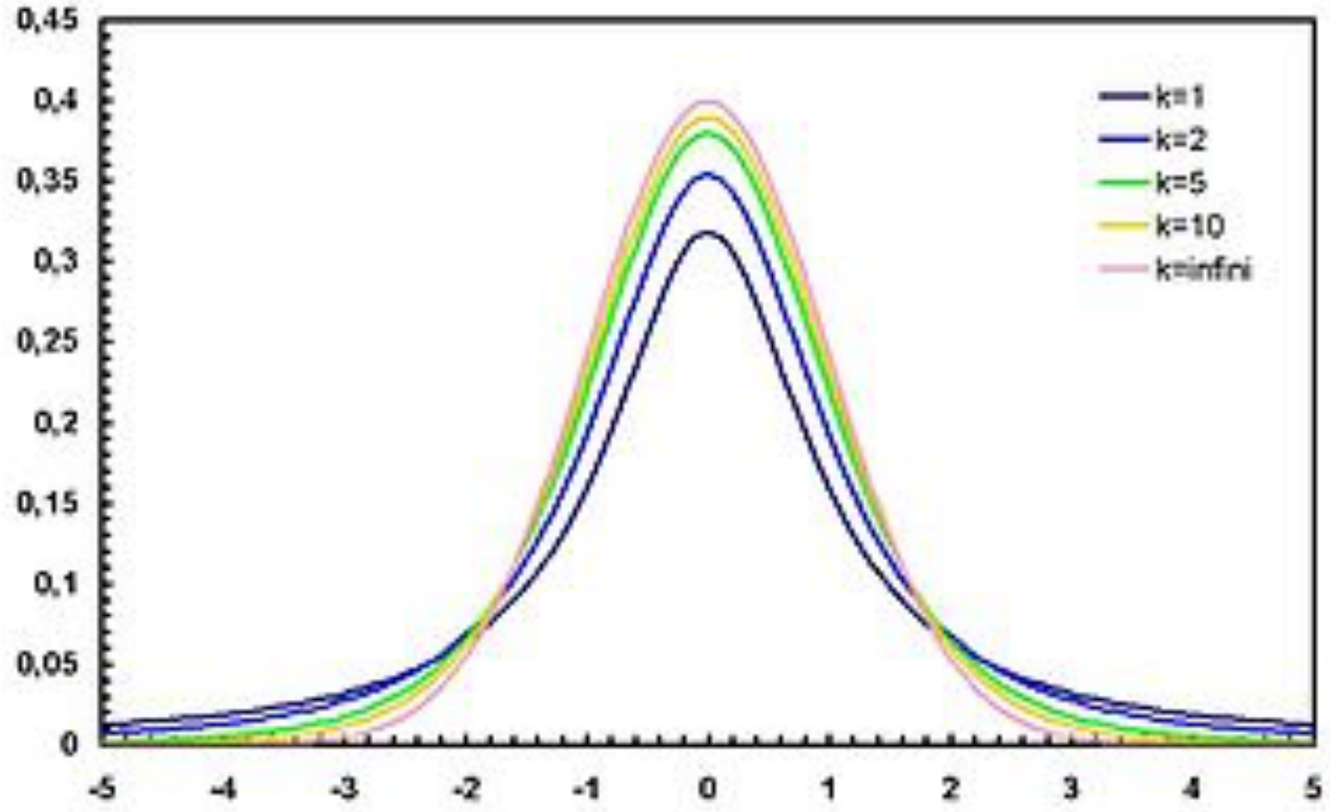
Тест Вилкоксона

Wilcoxon signed-rank test

The Wilcoxon signed-rank test is used to assess whether the differences are symmetric and centered around zero

- H_0 : differences follow a symmetric distribution around zero
- H_1 : differences don't follow a symmetric distribution around zero
- W-statistic
  - $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$ are *paired* samples
  - Compute $d_i = |X_i - Y_i| = 0$ and exclude elements with $d_i = 0$
  - Sort d_i ascending
  - $W = \sum sgn(X_i - Y_i) * R_i$, where R_i is the rank of d_i
  - $W \sim N\left(\mu = 0, \sigma = \sqrt{\frac{n(n+1)(2n+1)}{6}}\right)$ for $n \geq 10$

Распределение Стьюдента



Параметрические тесты

1) **Нормальный тест**
2) **Тест Стьюдента**
3) Критерий Хи-квадрат

Применение

- Сравнение выборочного среднего с заданным значением

- Сравнение двух выборочных средних при известных дисперсиях (дисперсии известны заранее, а не посчитаны из этих данных)

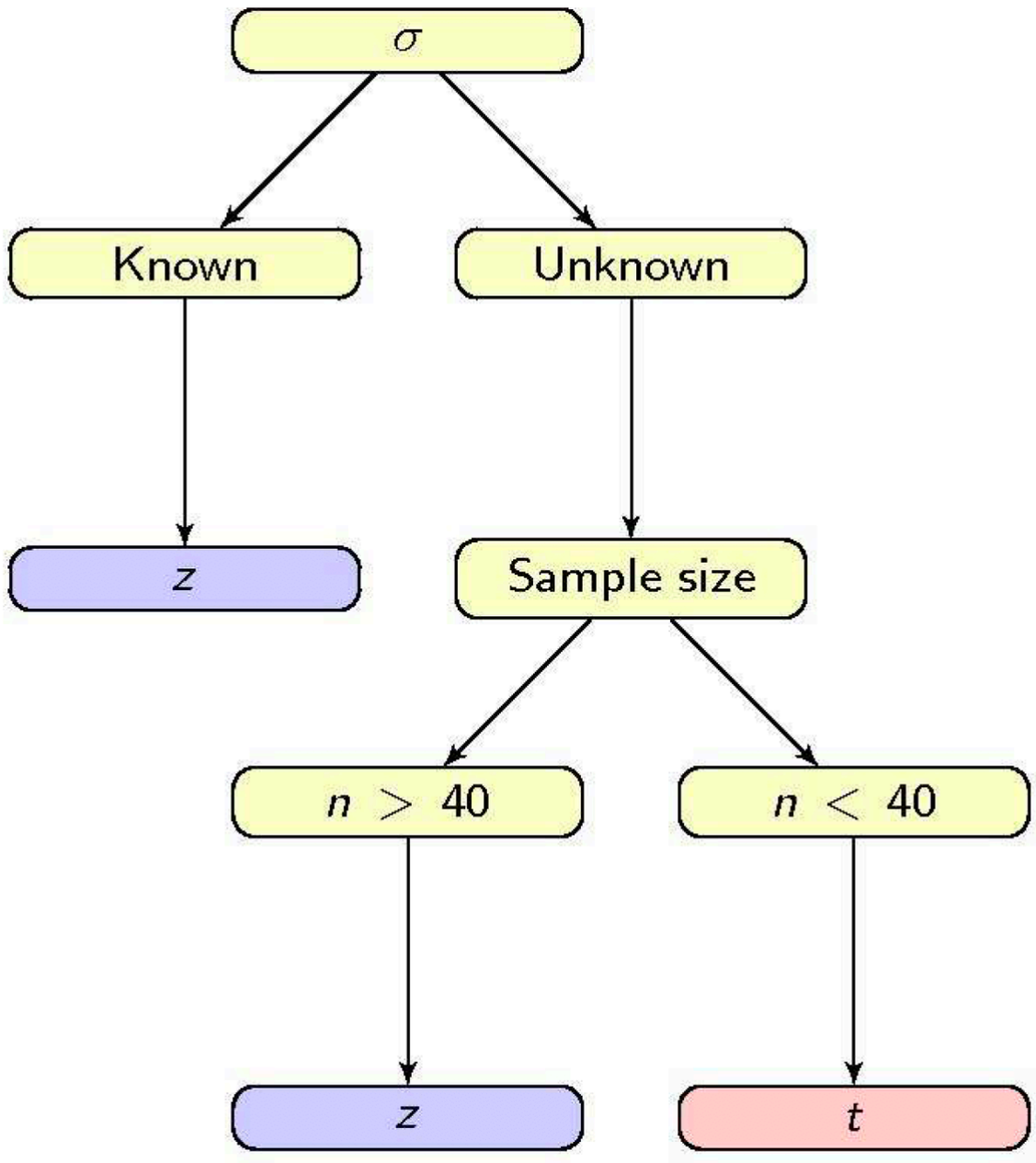
- Сравнение двух выборочных средних при неизвестных равных дисперсиях

- Сравнение двух выборочных средних при неизвестных неравных дисперсиях

Применение

- Сравнение двух выборочных средних в связанных выборках

Когда что применять



Проблема множественного тестирования