ANOVA, bootstrap и тесты на соответствие распределению

Тест на равенство средних двух групп

Что мы делаем, когда дисперсии не равны?

Тест на равенство средних двух групп

Что мы делаем, когда дисперсии не равны?

Используем тест Велча, который устойчив к этой ситуации

Предположений 1-way ANOVA

1. Наши наблюдения независимы.
2. Шум в данных распределен нормально Что если нет?
3. Гомоскедастичность – дисперсии в группах одинаковы
4. Для целей нашего исследования фактор одноуровневый – внутри себя он не делится на другие группы. Иначе – n-way ANOVA, repeated measures ANOVA – разберем на следующем занятии

Robust ANOVA

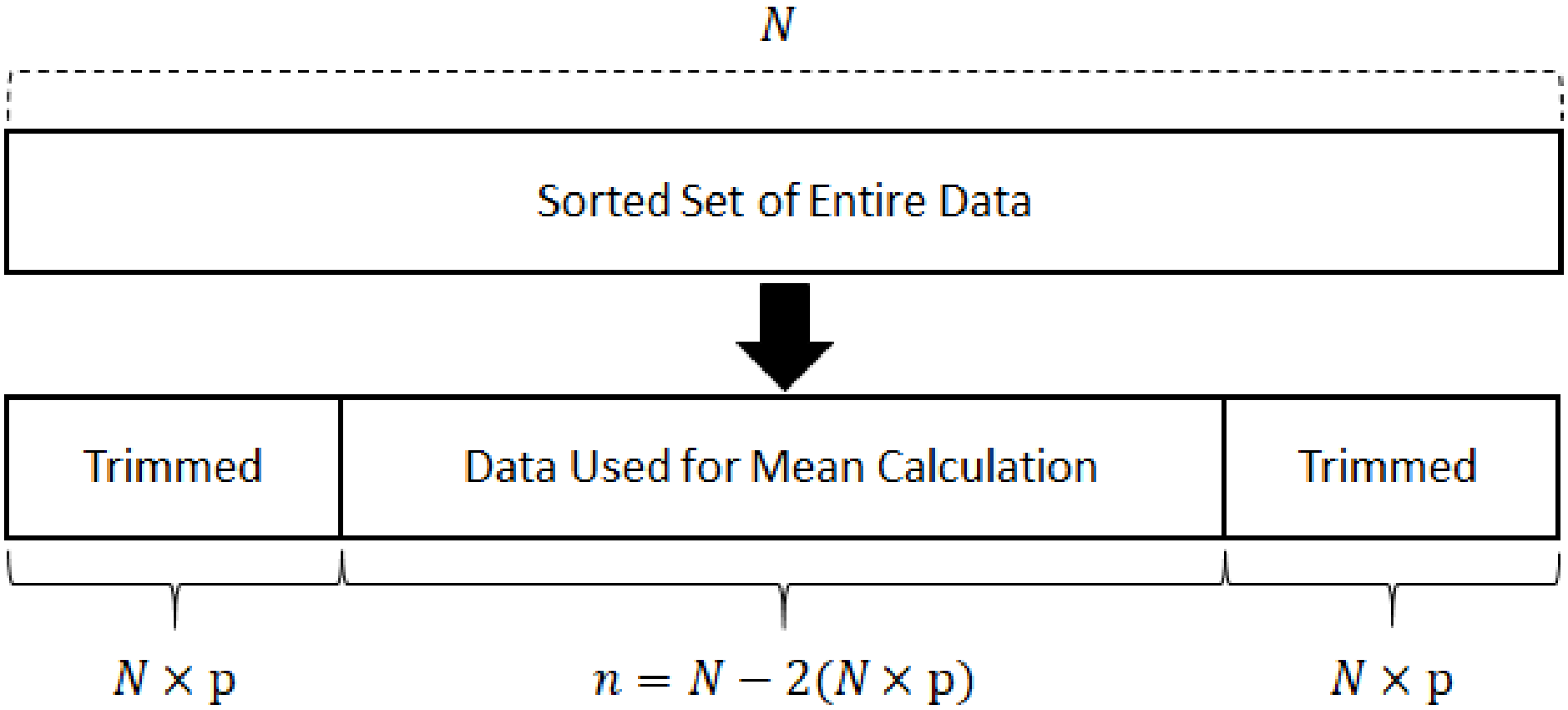
Не требует предположения о равенстве дисперсий и более устойчива к outliers

Использует несколько разных вариантов:

1) Trimmed mean
2) Median deviation
3) Boostrap

Trimmed mean

Мы знаем, что среднее неустойчиво к выбросам

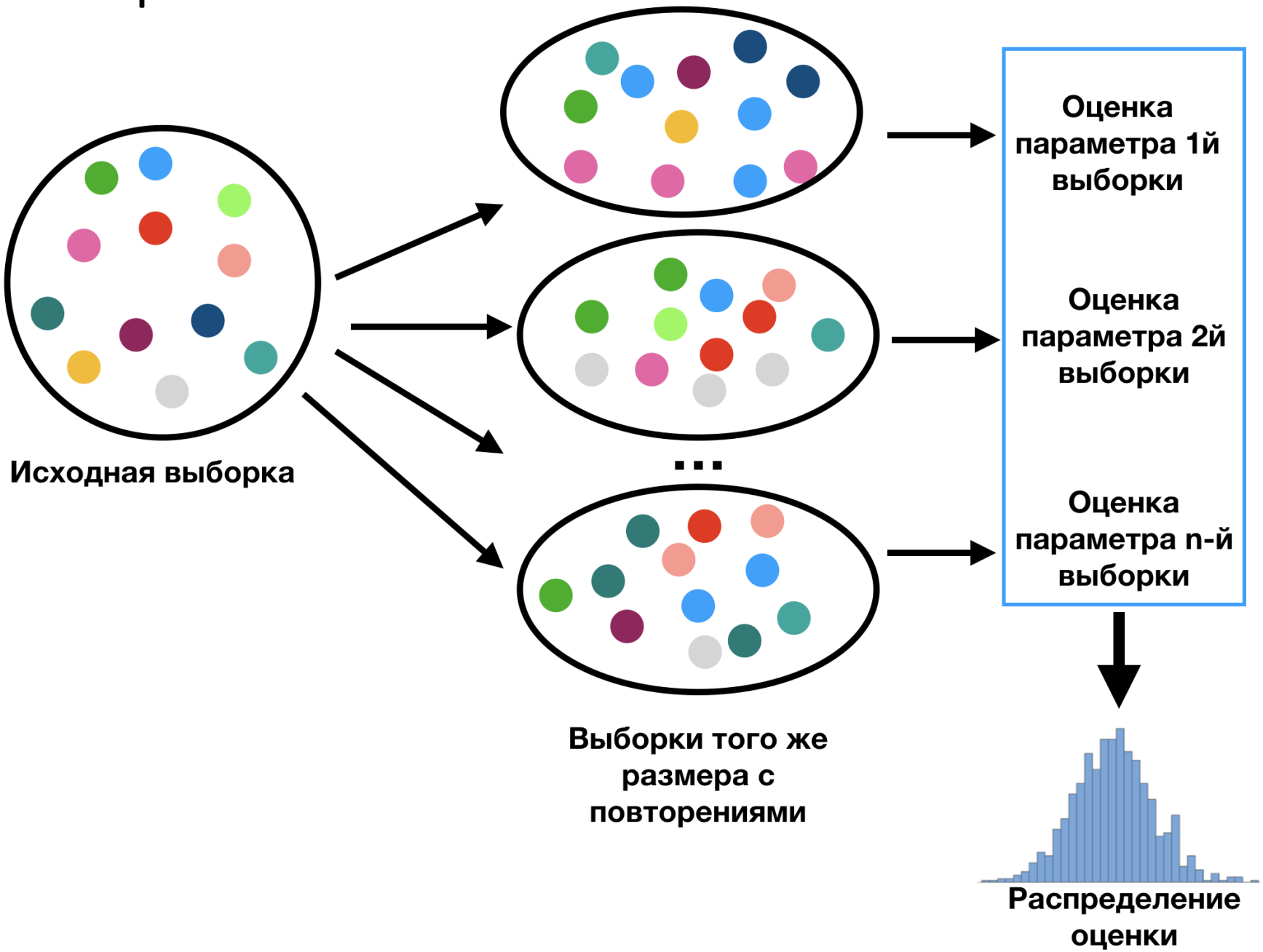


Median deviation

Вместо того, чтобы использовать среднее и дисперсию, которые неустойчивы к выбросам в данных, будем считать медиану и медиану абсолютных отклонений наблюдения от медианы

$$MAD = median(|x_i - median_x|$$

Bootstrap


Исходная выборка
...
Выборки того же размера с повторениями
Оценка параметра 1й выборки
Оценка параметра 2й выборки
Оценка параметра n-й выборки
Распределение оценки

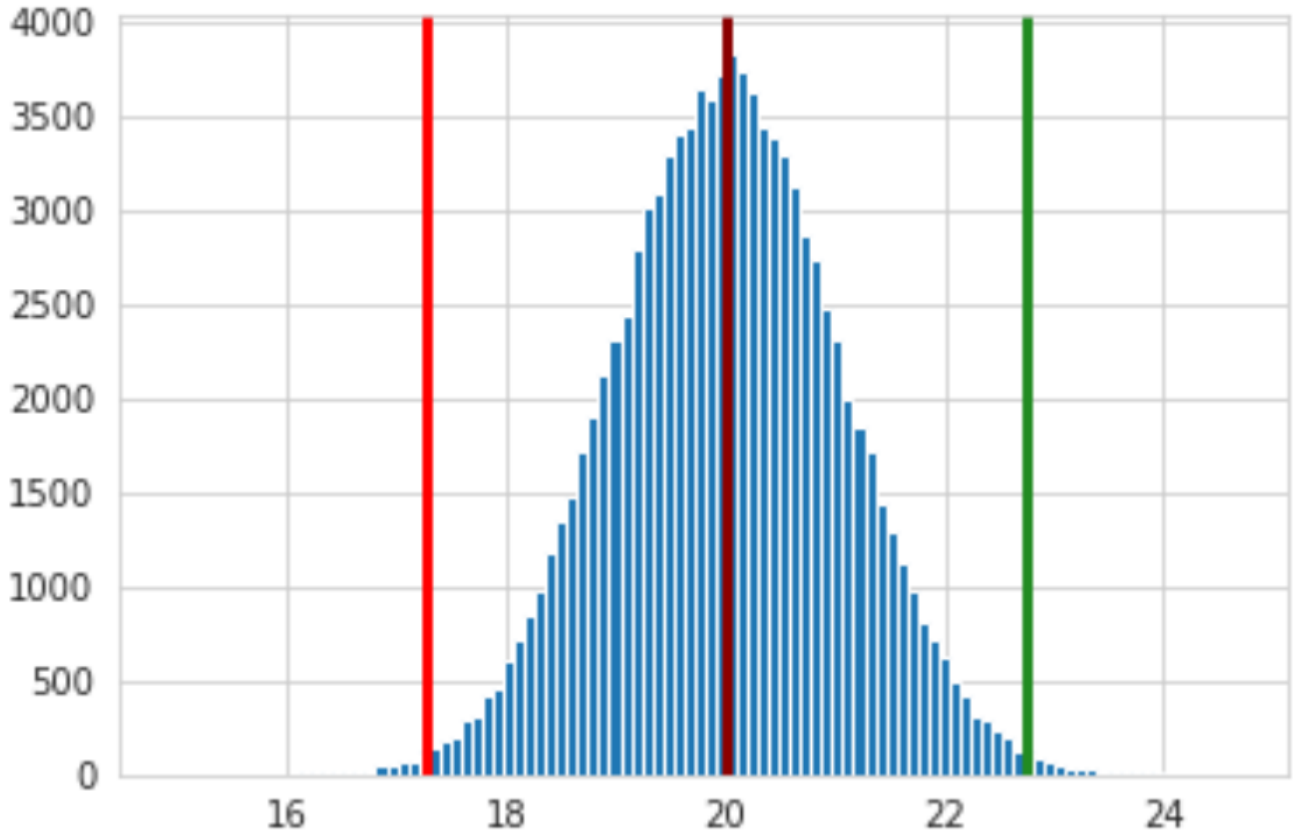
Bootstrap

Позволяет получить распределение любой оценки, которая в принципе может быть подсчитана по выборке

Как это нам может помочь в статистике?

Bootstrap

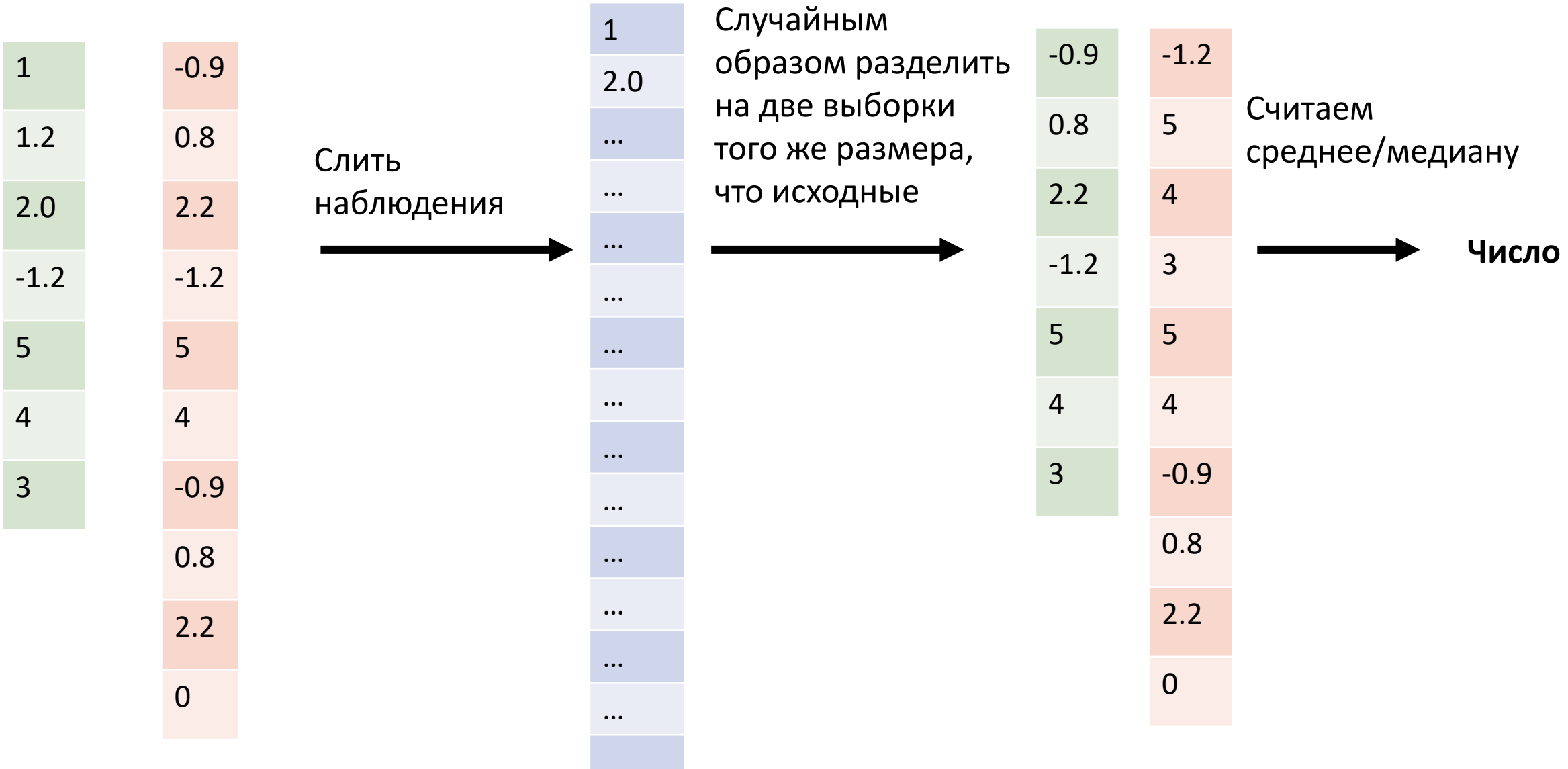
Позволяет получить распределение любой оценки, которая в принципе может быть подсчитана по выборке



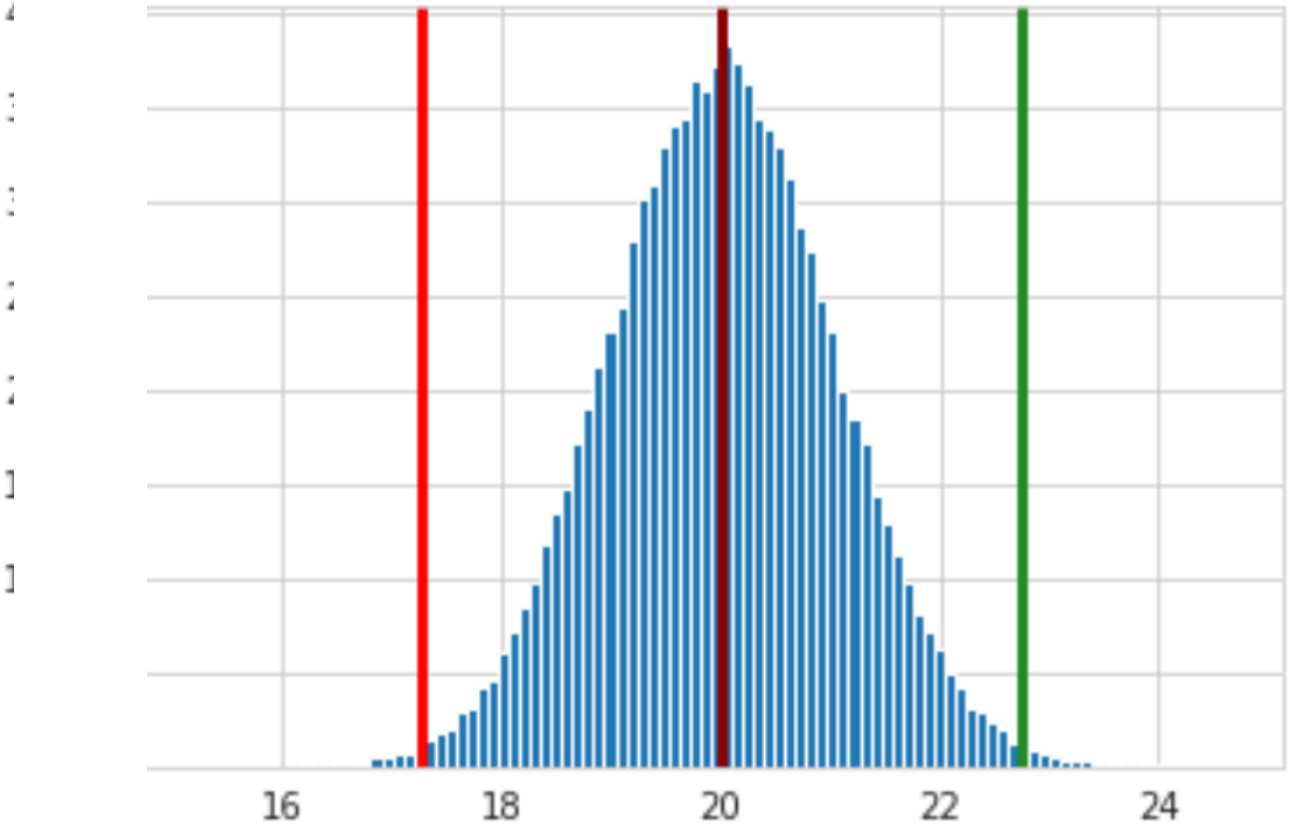
Как это нам может помочь в статистике?

Мы можем прикинуть, как выглядит распределение нашей статистики при условии H_0. Далее смотрим, сколько значений не меньше величины, посчитанной на реальном датасете. Доля таких величин – p-value

Bootstrap. Значимость различия между группами



Bootstrap. Значимость различия между группами



Смотрим, сколько раз число, посчитанное на реальных перемешанных, оказывается больше (или таким же критичным), чем то, что считаем на реальных - получаем p-value

Bootstrap

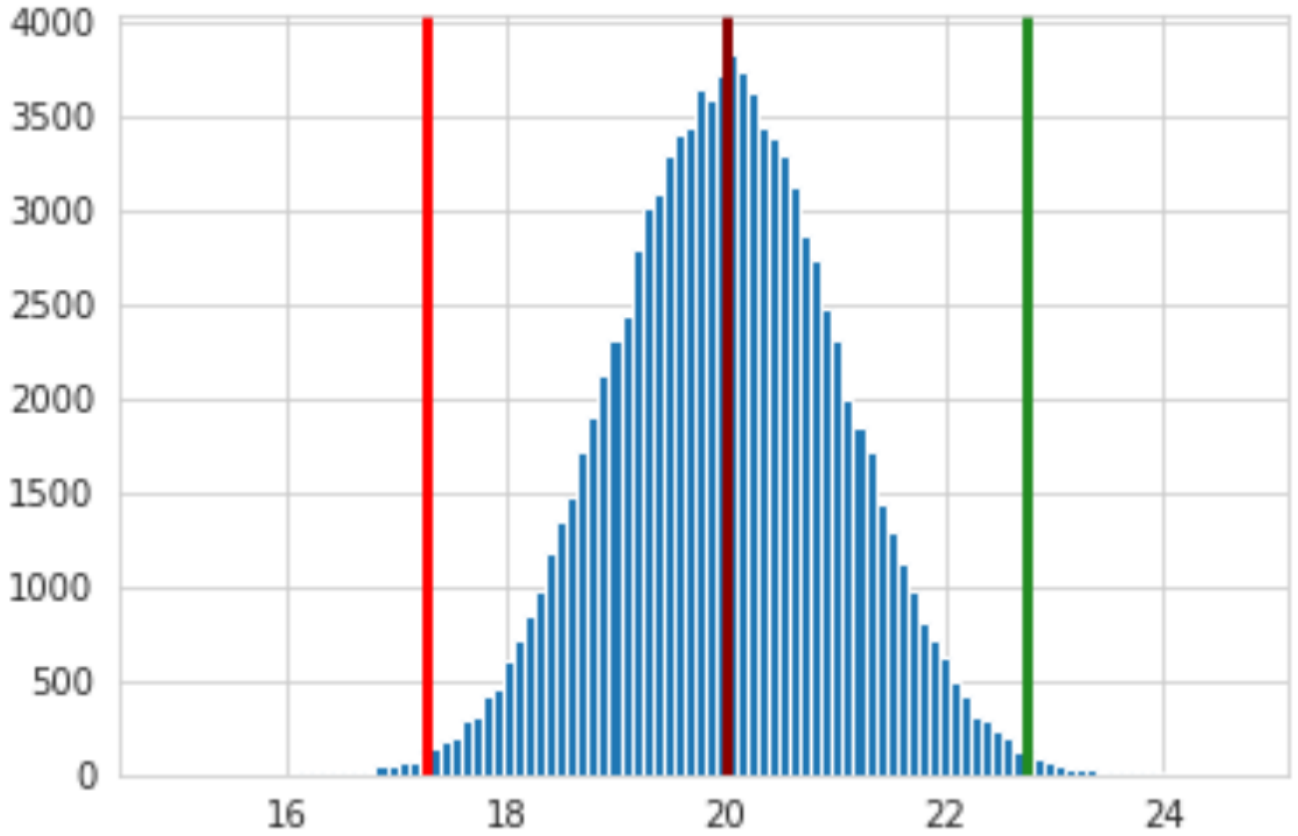
Позволяет получить распределение любой оценки, которая в принципе может быть подсчитана по выборке

Для ANOVA:

1. Если H_0 верна, то есть средние всех группа равны, то мы можем перемещать объекты между группами без нарушения H0
2. Давайте так и сделаем. 10000 раз перемешаем объекты между группами и для каждой полученной выборки посчитаем статистику ANOVA
3. Посмотрим, в скольких случаях

Bootstrap

Позволяет получить распределение любой оценки, которая в принципе может быть подсчитана по выборке



Как это **еще** нам может помочь в статистике?

Мы можем проверить устойчивость наших оценок –правильно проведенный бутстрэп влияет на оценку несильно и все отклонения – следствия истинной неуверенности оценки

Bootstrap. Доверительный интервал для коэффициента корреляции

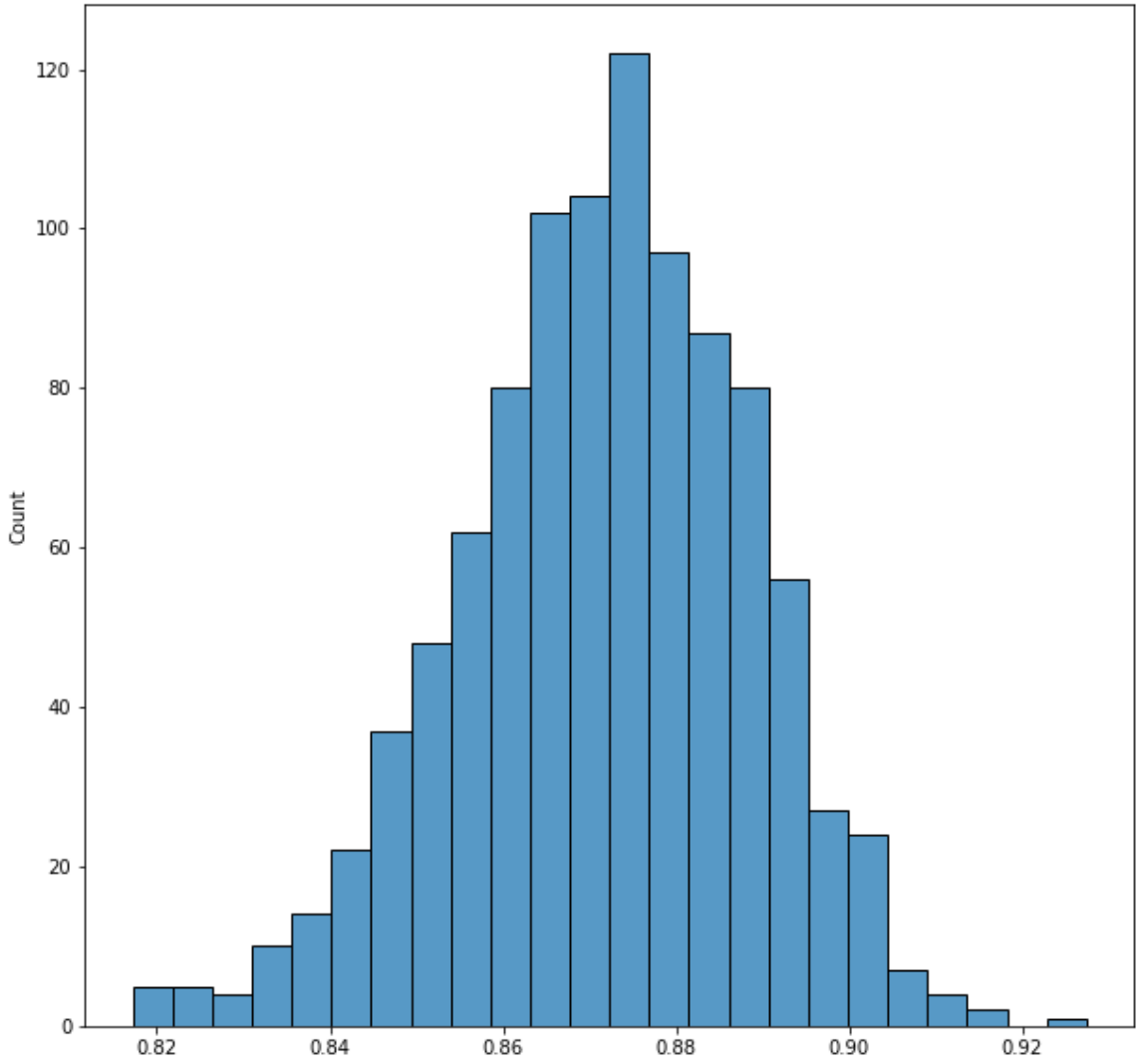
X	Y
-0.9	-2
0.8	1
2.2	2.2
-1.2	-3
5	4
4	4
-0.9	-100
0.8	20
2.2	5
0	0

Выбираем случайным образом то же самое количество пар с замещением

X	Y
-0.9	-2
0.8	1
0.8	1
2.2	2.2
-1.2	-3
2.2	2.2
-1.2	-3
5	4
4	4
2.2	5

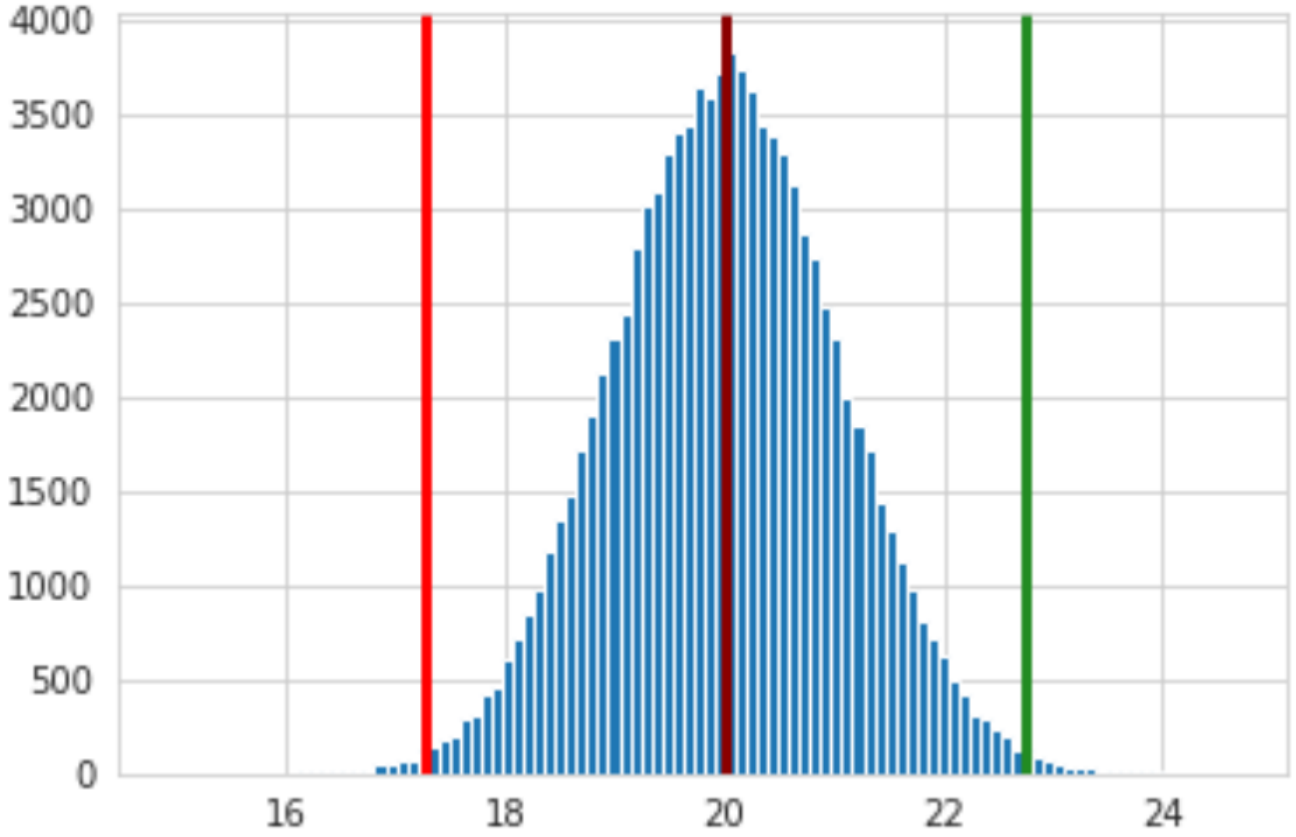
Корреляция на «случайных» данных

Bootstrap. Доверительный интервал для коэффициента корреляции



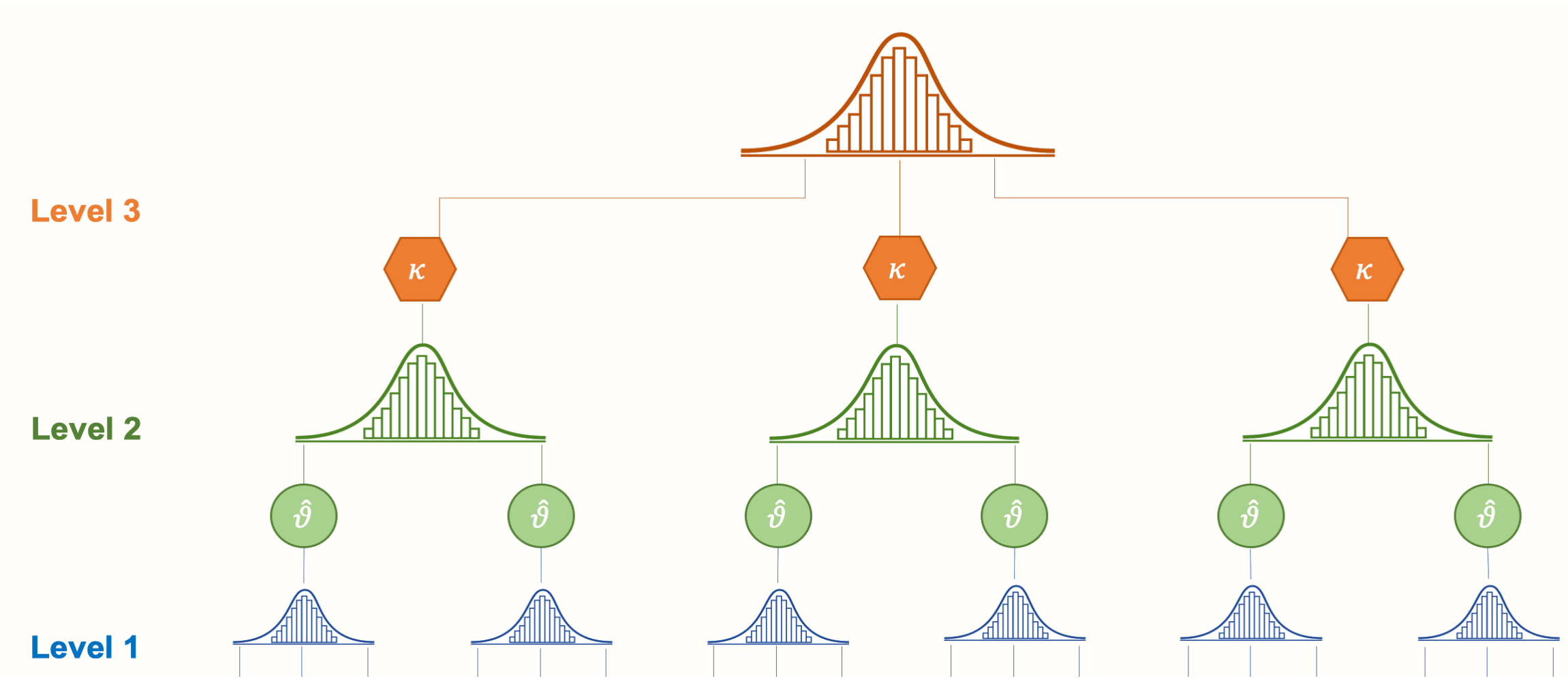
Получаем распределение. Чтобы получить из него 95% доверительный интервал – просто записываем 2.5 квантиль и 97.5 квантиль

Bootstrap ≈ Monte-Carlo sampling ≈ Permutation test



Часто эти термины используются взаимозаменяемо. Monte-Carlo sampling - похожая концепция, permutation test иногда подразумевает, что мы реально делаем не случайные перестановки, а ВСЕ перестановки (как в тесте Фишера)

Two-way ANOVA

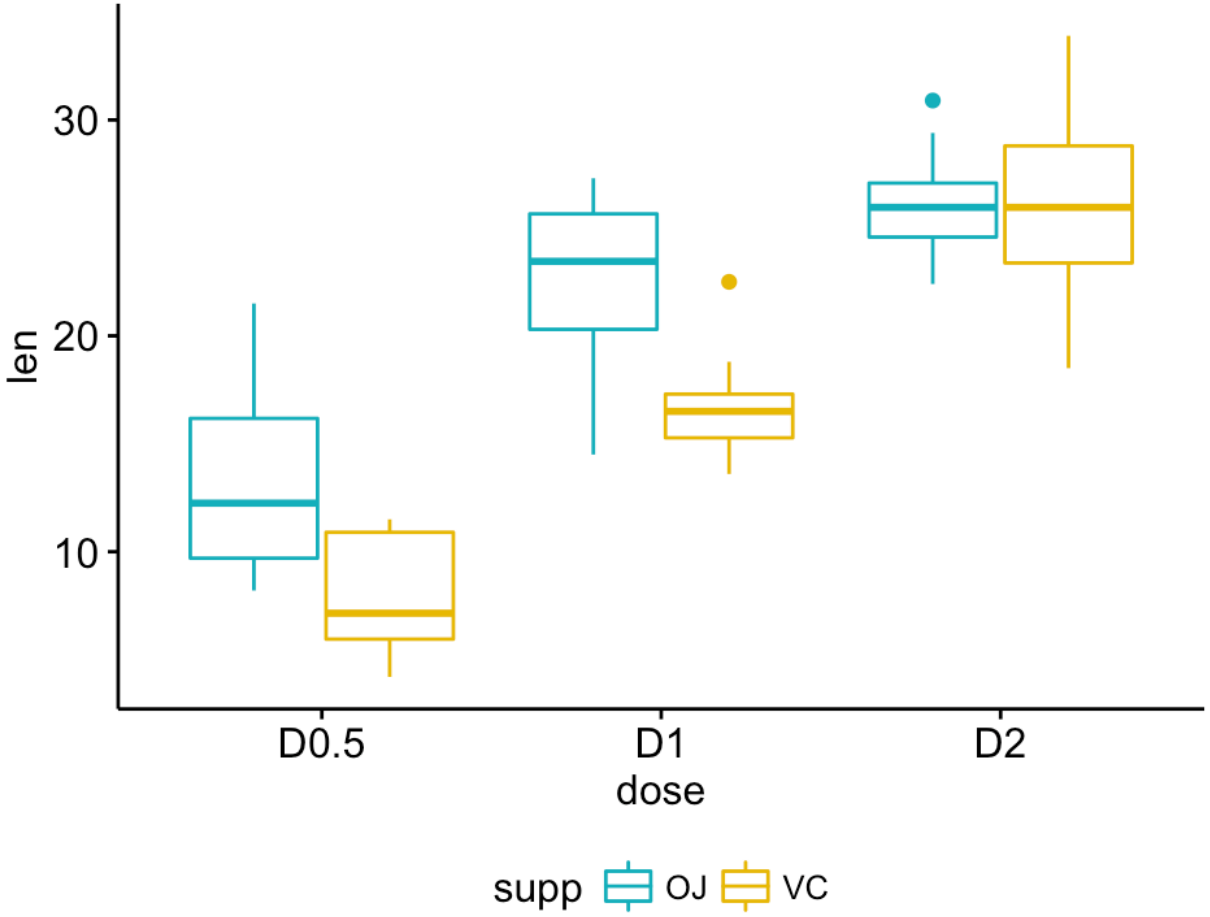


Two-way ANOVA

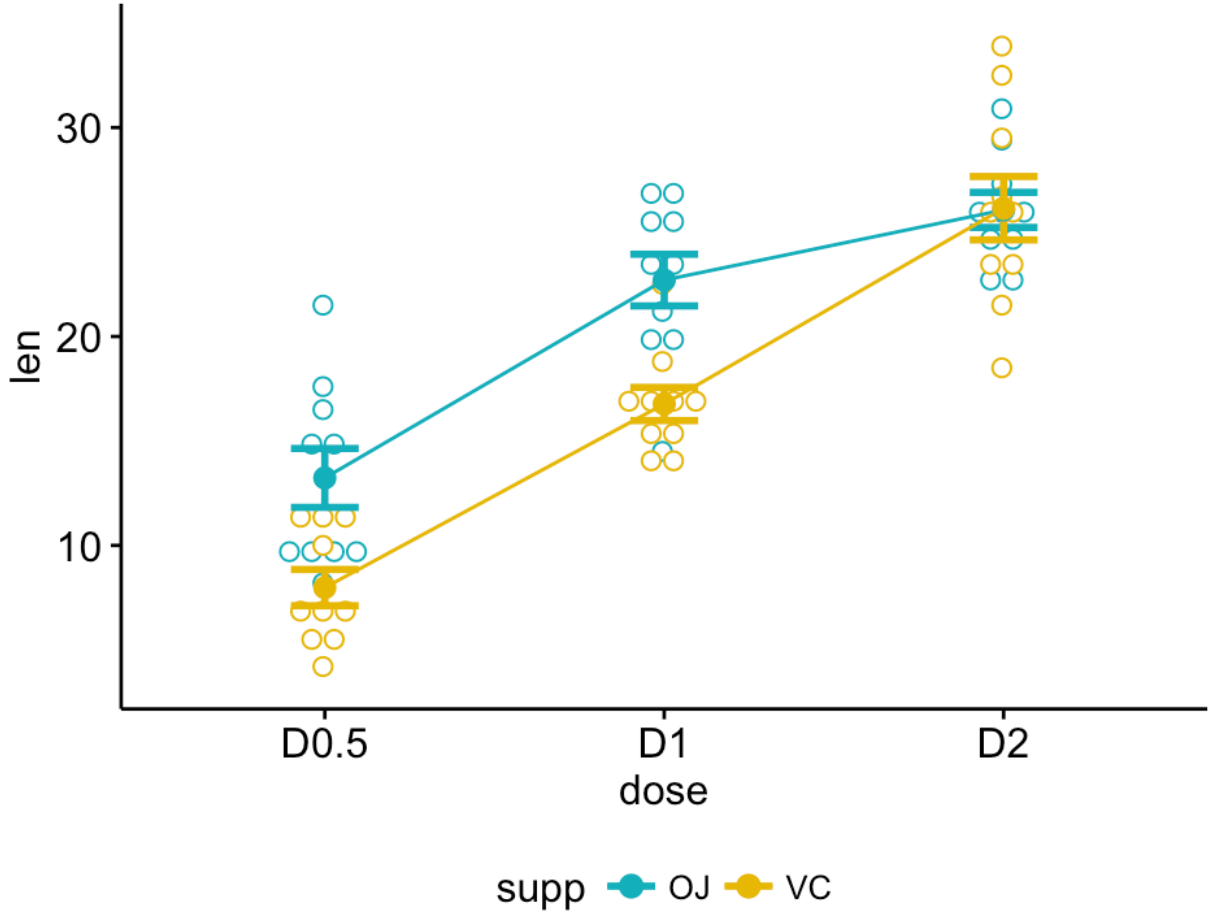
Можно интерпретировать по-разному:

1. У нас одна вещественная переменная и две категориальных. Хотим узнать взаимосвязь вещественной переменной с каждой из них и их комбинацией
2. У нас есть одна вещественная переменная и одна категориальная. Но категориальная переменная - многоуровневая

Two-way ANOVA



Не бро



Бро

Two-way ANOVA. Без взаимодействия

Предполагаем, что взаимодействия факторов нет.

Сильное предположение, в реальности обосновать достаточно сложно

Проверяем две гипотезы

H_0 - средние во всех группах по первому фактору одинаковы

H_0 - средние во всех группах по второму фактору одинаковы

Two-way ANOVA. Без взаимодействия

```
ToothGrowth$dose <- factor(ToothGrowth$dose,
                  levels = c(0.5, 1, 2),
                  labels = c("D0.5", "D1", "D2"))
res.aov2 <- aov(len ~ supp + dose, data = ToothGrowth)
summary(res.aov2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## supp          1  205.4   205.4   14.02 0.000429 ***
## dose          2 2426.4  1213.2   82.81  < 2e-16 ***
## Residuals    56  820.4    14.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-way ANOVA. С взаимодействием

Проверяем три гипотезы

H_0 - средние во всех группах по первому фактору одинаковы

H_0 - средние во всех группах по второму фактору одинаковы

H_0 - влияние первого фактора зависит от влияния второго

Two-way ANOVA. С взаимодействием

```
res.aov3 <- aov(len ~ supp * dose, data = ToothGrowth)
summary(res.aov3)
```

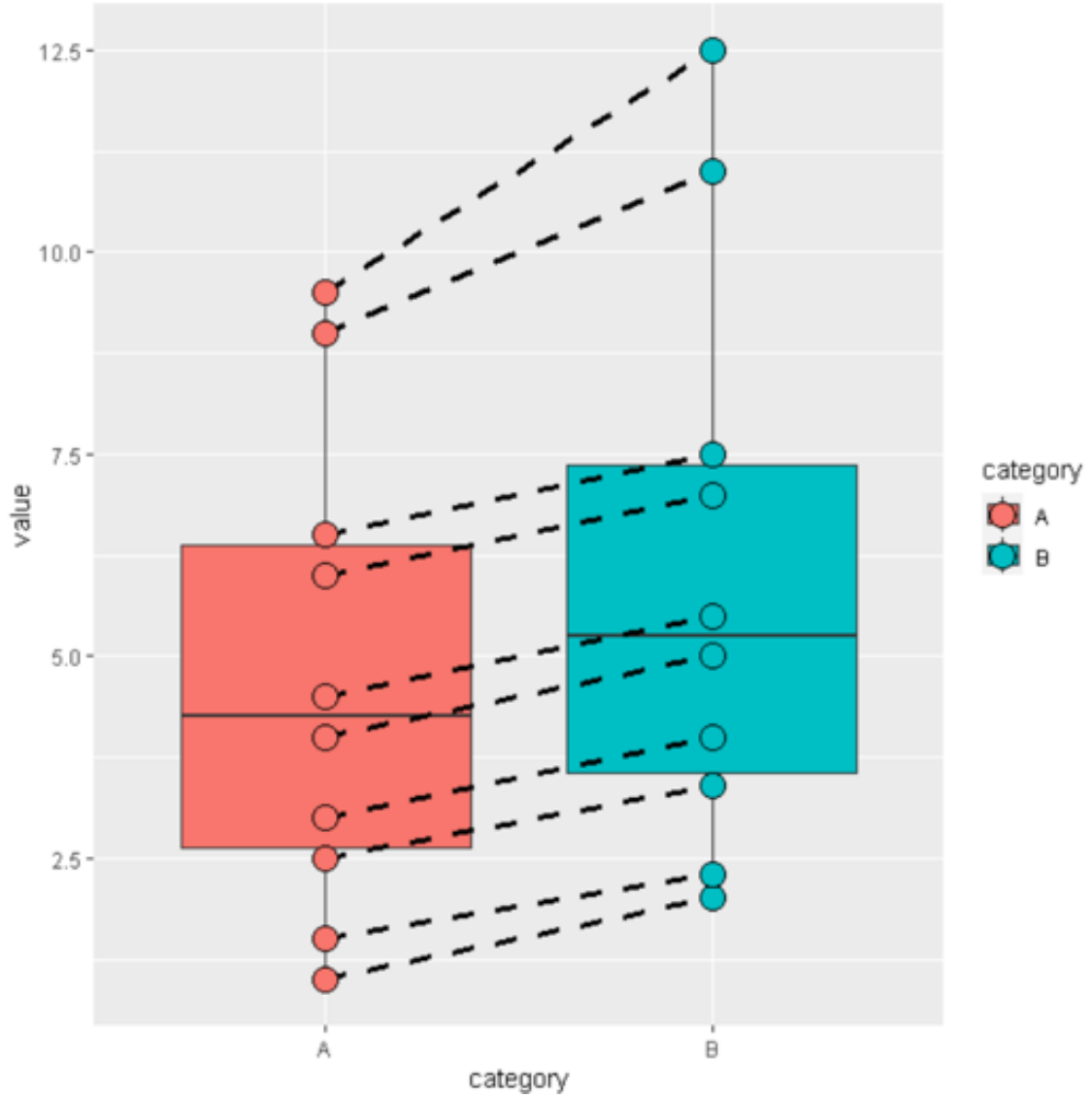
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## supp          1  205.4   205.4  15.572 0.000231 ***
## dose          2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose     2  108.3    54.2   4.107 0.021860 *
## Residuals    54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiway ANOVA

То же самое, но для бОльшего количества факторов

Парные t-test

Наблюдения связаны – мы меряли одно и то же для разных людей



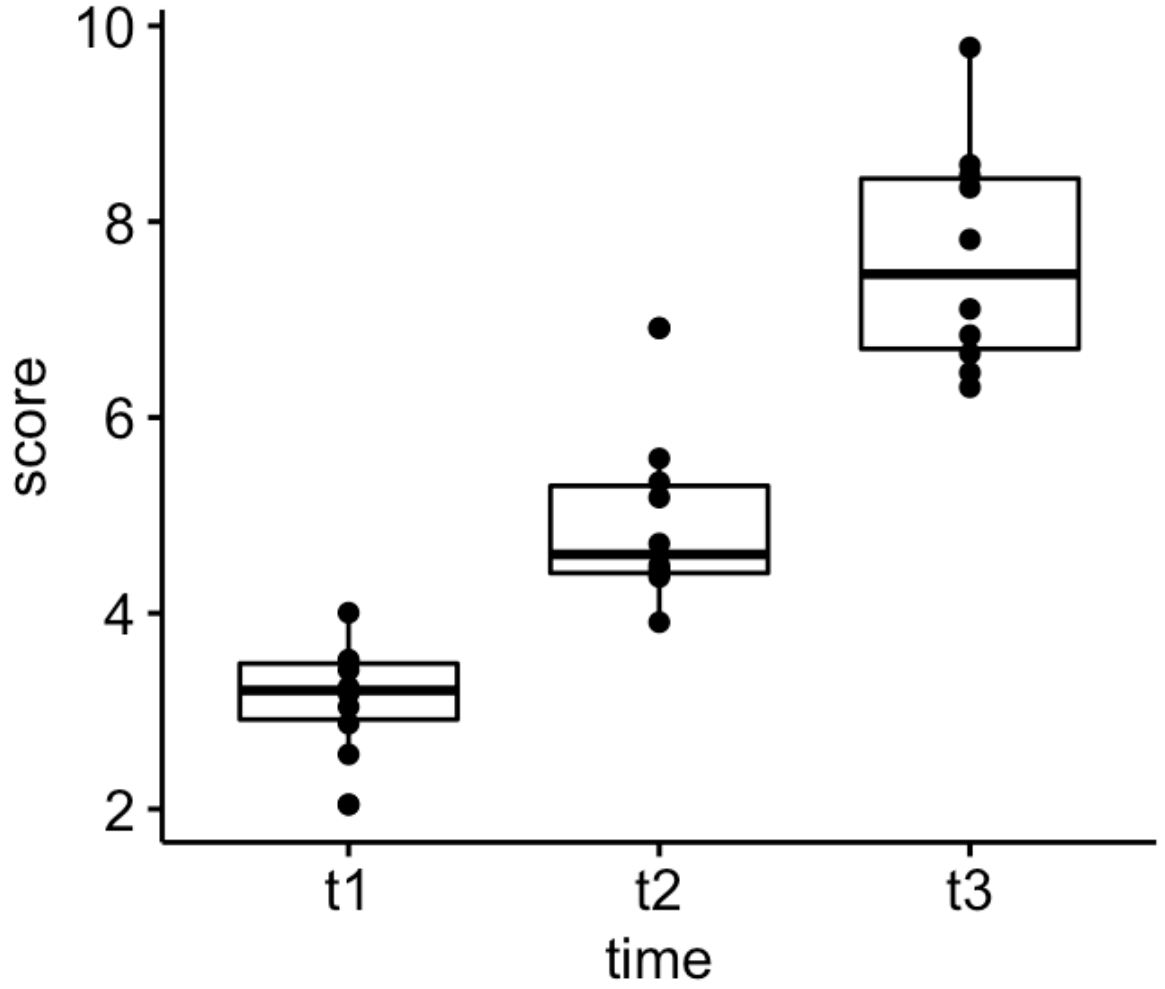
Repeated measures 1-way ANOVA

Нам не интересно влияние второго фактора. Но он есть и мы должны его учесть.

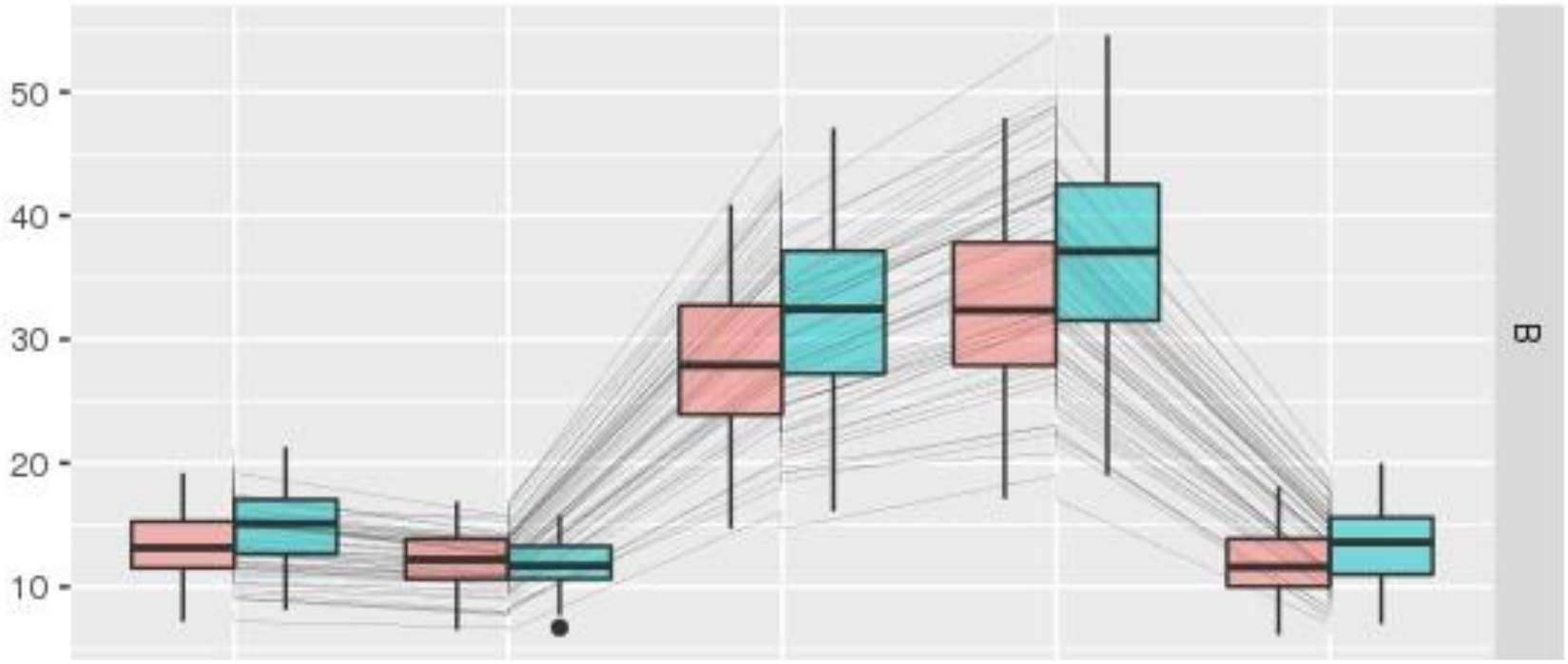
Например – мы делаем измерения в разные периоды времени (до терапии, сразу после терапии, 2 месяца после и тд) для одних и тех же людей. Нам интересно влияние первого фактора.

Второй же фактор (конкретный человек) «просто» повторностью – мы делаем лекарство не для Джона, а для людей в целом. Конкретный человек – просто случайный образец из генеральной совокупности

Repeated measures 1-way ANOVA



Repeated measures 2-way ANOVA

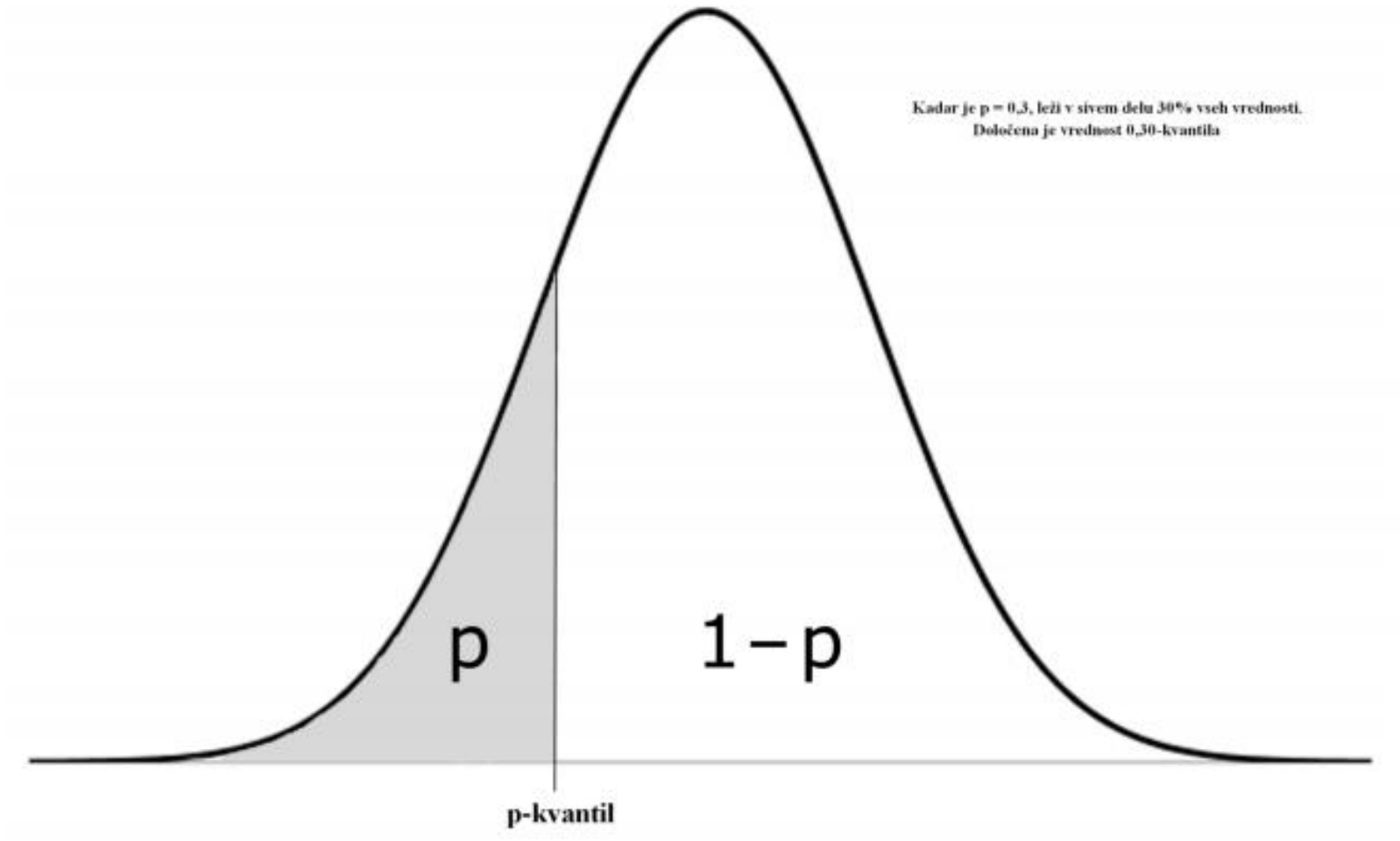


Тесты на соответствие распределению

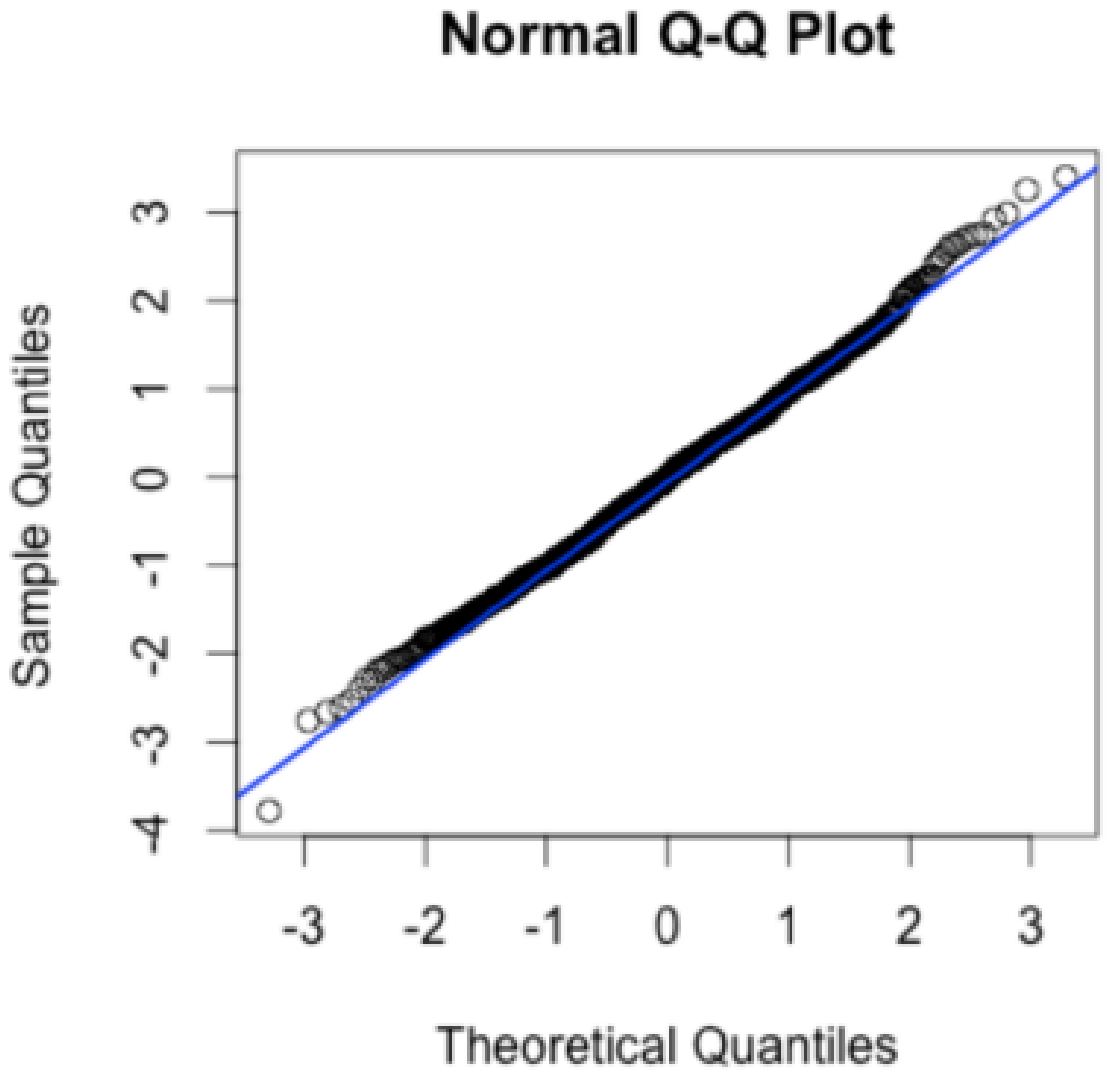
Часто хотим проверить, а соответствую ли наши данные какому-то распределению

Квантили

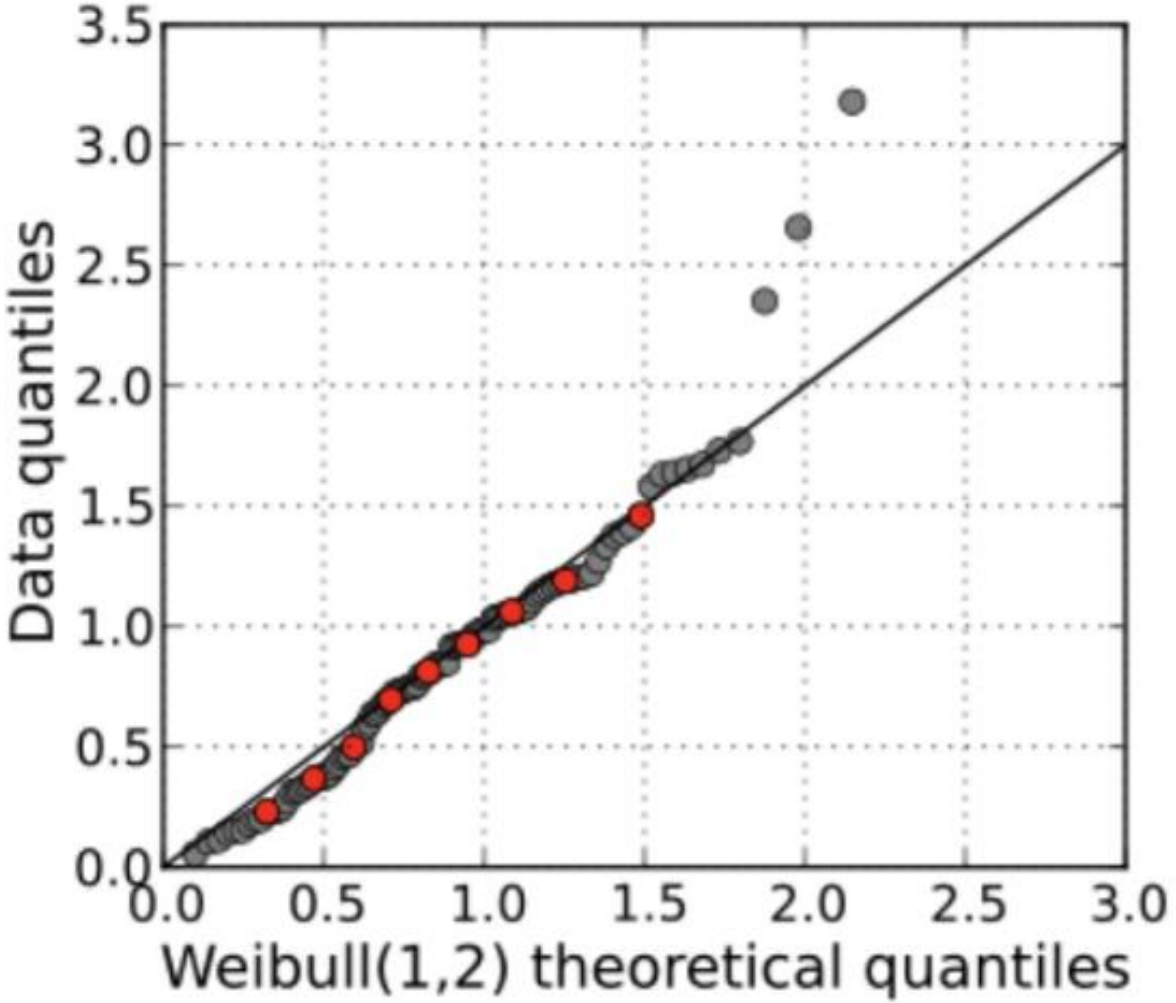
Квантиль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.



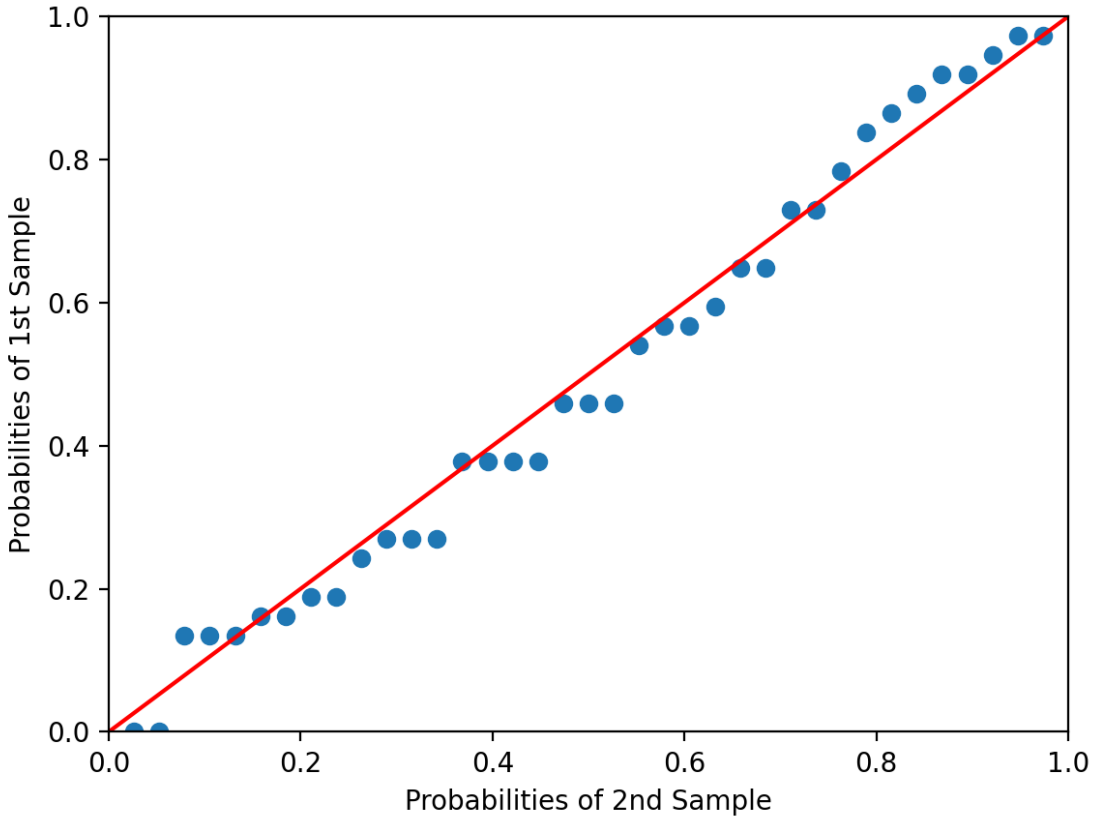
QQplot – визуальное сравнение нашего распределения с другим

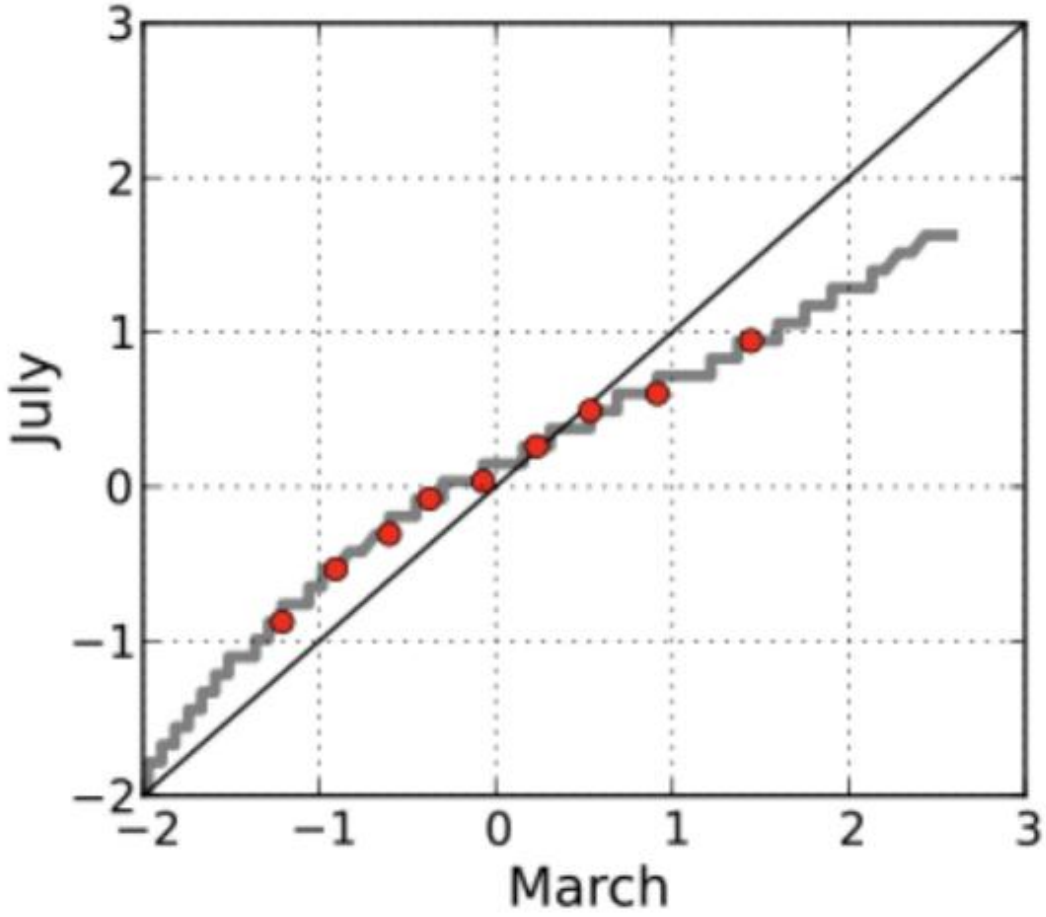


Не обязательно сравнивать с нормальным распределением

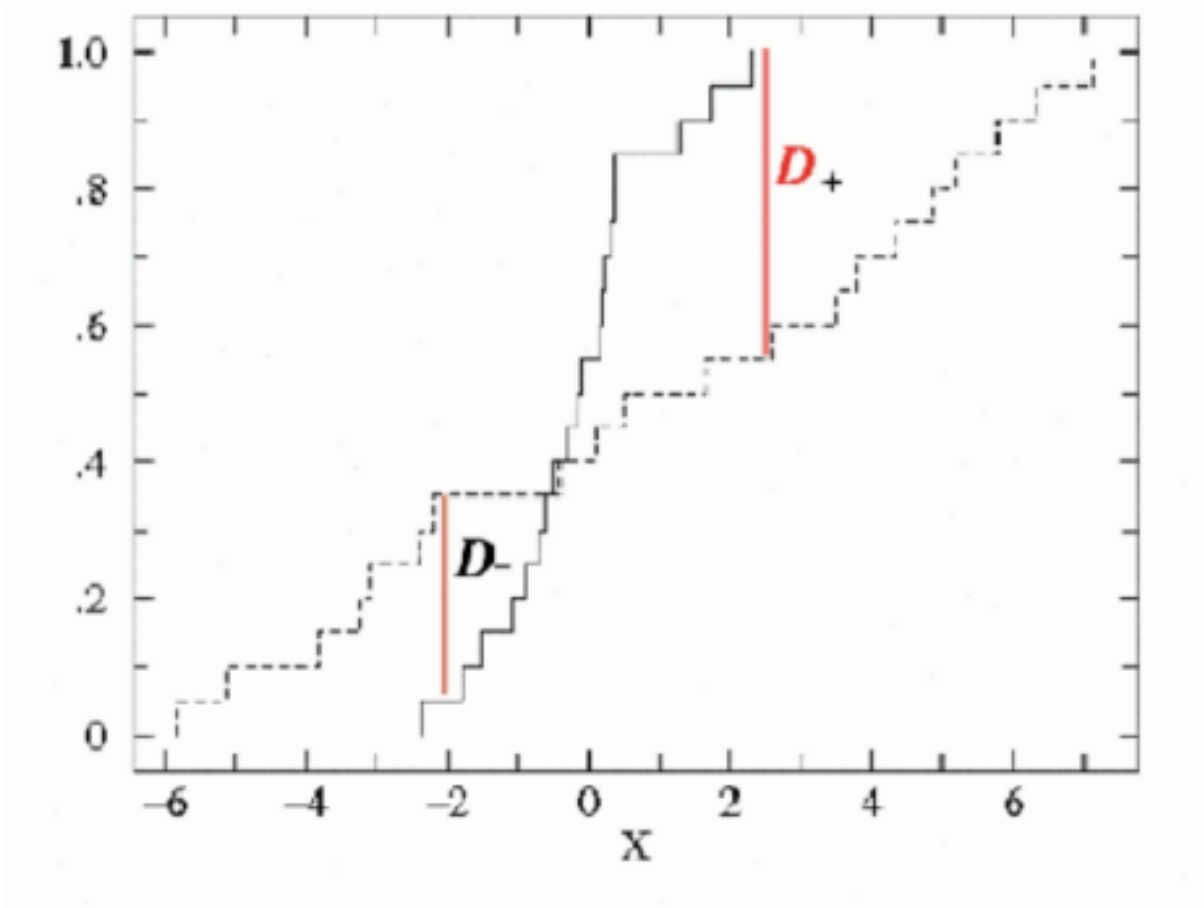


QQplot – можно сравнивать две выборки





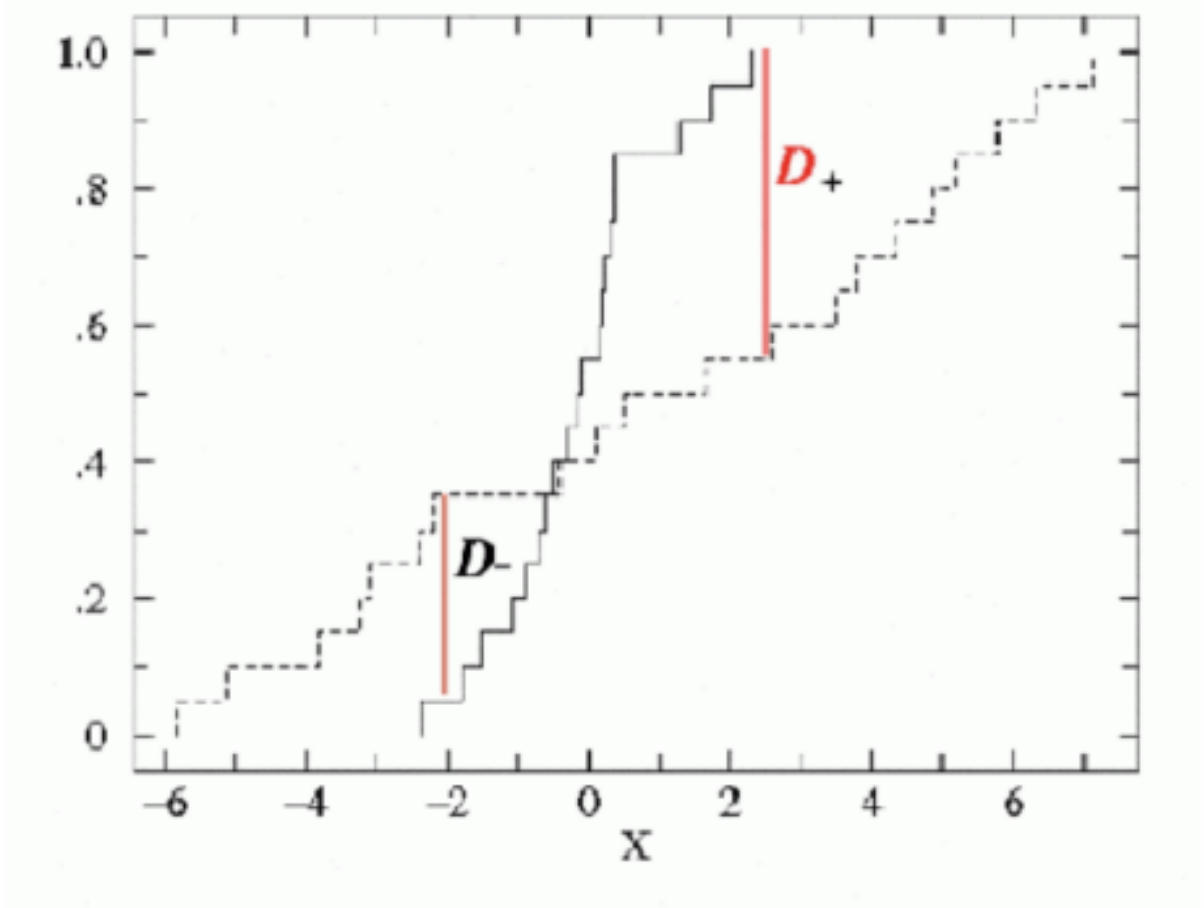
Тест Колмогорова-Смирнова



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

Можно сравнивать только с известным, полностью заданным распределением

Тест Колмогорова-Смирнова

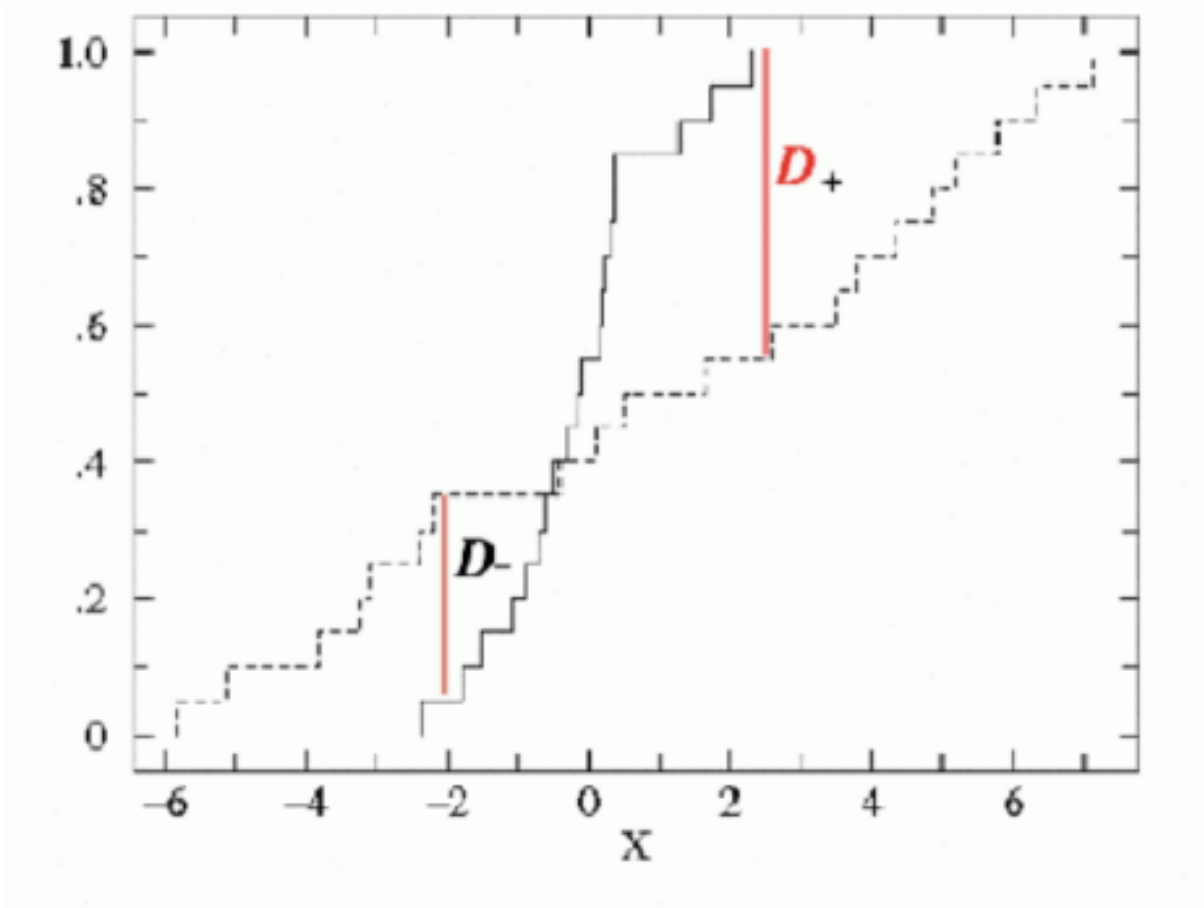


$$D_n = \sup_x |F_n(x) - F(x)|.$$

Можно сравнивать только с известным, полностью заданным распределением

Нельзя проверить, соответствуют ли наши данные нормальному распределению вообще

Тест Колмогорова-Смирнова



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

Можно сравнивать только с известным, полностью заданным распределением

Нельзя проверить, соответствуют ли наши данные нормальному распределению вообще

Можно, если использовать монте-карло сэмплирование. Но в R по-умолчанию не он

Тест Шапиро-Уилка

Проверяем нормальность выборки

Очень строгий тест, на больших выборках крайне склоненнен отвергать нулевую гипотезу

Хи-квадрат тест на соответствие распределению

Пусть у нас есть мультиномиальное распределение следующего вида

Значение	val_0	val_1	val_2	val_3	val_4	…
Вероятность	p0	p1	p2	p3	p4	…

Хи-квадрат тест на соответствие распределению

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$df = n - 1$$

n - число ячеек

Пример

Программист Петя считает, что количество лайков, которые соберут посты с шутками на тему неприятных особенностей языка, одинаковы. Для теста были выбраны языки C++, Python, Javascript, Java и R. Количество лайков для постов про эти языки составило соответственно:

17, 23, 72, 44, 65

Прав ли Петя? Уровень значимости 0.001, так он не хочет никого в случае чего обидеть незаслуженно.

Пример

	C++	Python	Javascript	Java	R
Число лайков	17	23	72	44	65
Вероятность при условии H0	0.20	0.20	0.20	0.20	0.20

Пример

Гипотеза H0: Все языки получили равное число лайков, распределение лайков равномерное

Гипотеза H1: Языки получили значимо разное число лайков

Если распределение лайков равномерное, то ожидаемое число лайков для каждого языка:

$$E = 221/5 = 44.2$$

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} = 56.2$$

$$df = n - 1 = 4$$

$$P(\chi^2(4) > 56.2) = 1e - 11 < 0.001$$

Число степеней свободы

Не всегда n – 1

Оно зависит от того, сколько параметров, оцененных по выборке, мы использовали для подсчета вероятности попасть в ячейку

Пример

Наблюдается число студентов, опаздывающих на 0 минут, минуту, две, три, четыре и 5 минут и более

Значение	0	1	2	3	4	5 и более
Студентов	14	30	33	14	6	3

Проверьте гипотезу о том, что число студентов распределено по Пуассону

Пример

Наблюдается число студентов, опаздывающих на 0 минут, минуту, две, три, четыре и 5 минут (больше не опаздывают)

Значение	0	1	2	3	4	5
Студентов	14	30	33	14	6	3

Если число студентов распределено по Пуассону, то lambda = (14 * 0 + 30 * 1 + 2 * 33 + 3 * 14 + 4 * 6 + 5 * 3) / 100 = 1.77

Можно подсчитать (по формуле или с использованием функции dpoiss R вероятность значения попасть в каждую из ячеек)

Пример

Значение	0	1	2	3	4	5
Студентов	14	30	33	14	6	3
p	0.17	0.30	0.27	0.16	0.07	0.02

Остается подсчитать ожидаемое число студентов

Значение	0	1	2	3	4	5
Студентов	14	30	33	14	6	3
p	0.17	0.30	0.27	0.16	0.07	0.02
Ожидаемое	17	30	27	16	7	2

Пример

Теперь можно подсчитать значение статистики, оно равно 2.76

В этом случае у нас было условие на то, что наблюдений суммарно N = 100 и на то, что lambda нашего Пуассоновского распределения равна 1.77 (мы ее считали из наших наблюдений)

Потому суммарно получаем число степеней свободы равным n - 2 = 6 - 2 = 4.

Получаем, что p-value близко к 1, то есть у нас нет оснований отвергать гипотезу о том, что наблюдения распределены по Пуассону.

Пример

Сколько степеней свободы надо в случае:

- Проверить гипотезу о том, что наши наблюдения распределены по нормальному закону?
- Проверить гипотезу о том, что наши наблюдения распределены по равномерному закону?
- Проверить гипотезу о том, что наши наблюдения распределены по Пуассону с параметром 2?

Пример

Сколько степеней свободы надо в случае:

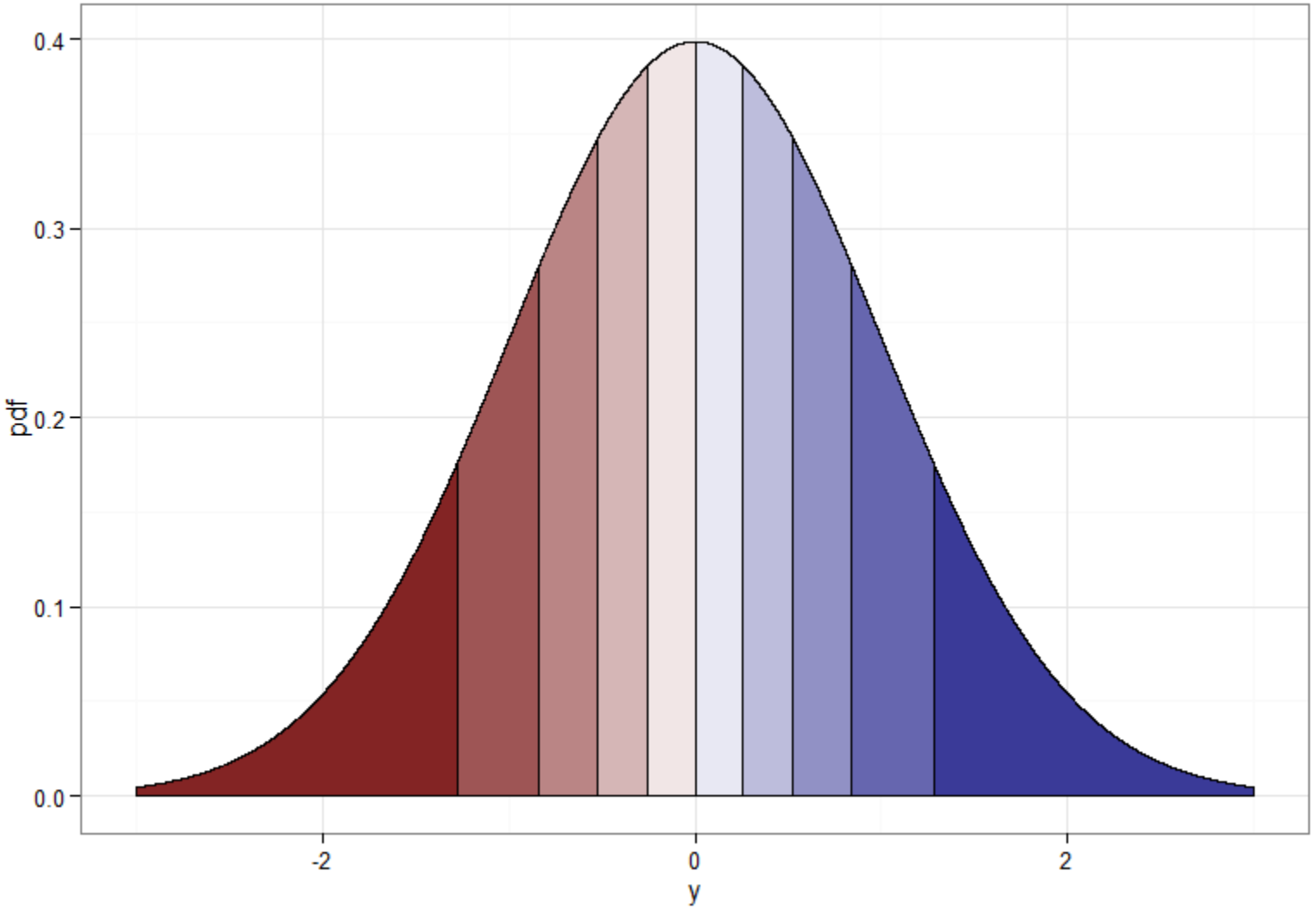
- Проверить гипотезу о том, что наши наблюдения распределены по нормальному закону?

  n - 1 - 2 = n - 3, считаем среднее и дисперсию
- Проверить гипотезу о том, что наши наблюдения распределены по равномерному закону?

  n - 1, мы это делаем по-умолчанию
- Проверить гипотезу о том, что наши наблюдения распределены по Пуассону с параметром 2?

  n - 1, мы взяли параметр не из наблюдений

Как сделать тест на нормальность?



Бинировать нормальное распределение