

Критерий Хи-квадрат

Тест на независимость

Используется как на то, есть ли значимая ассоциация между двумя факторными переменными

H0: факторы независимы

H1: факторы зависимы

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

$$df = (n - 1) \cdot (m - 1)$$

Где df - число степеней свободы, n - число разных значений первой переменной, m - число разных значений второй

Задача

Для четырех категорий людей - школьников, студентов, программистов (закончивших учебу со стажем < 5 лет и программистов (закончивших учебу) со стажем больше 5 лет имеются данные о их отношении к PHP. Отношение может быть “хороший язык”, “ну а что поделать” “ненавижу”. Проверить гипотезу о том, что категории независимы. Уровень значимости принять равным 0.01

Отношение/ Категория	Школьники	Студенты	Программис т, < 5 лет	Программис т, > 5 лет
Хороший язык	40	22	17	12
Ну а что поделать	15	12	20	35
Ненавижу	35	20	22	10

Решение

Гипотеза H0: Отношение не зависит от категории

Гипотеза H1: Отношение зависит от категории

Если отношение не зависит от категории, то $P(\text{хороший язык, категория}) = P(\text{хороший язык}) * P(\text{категория})$. То есть вероятность объекта оказаться в ячейке - произведение вероятностей в соответствующих столбце и строке

Отношение /Категория	Школьник и	Студенты	Программист, < 5 лет	Программист, > 5 лет	Сумма	Вероятность
Хороший язык	40	22	17	12	91	0.35
Ну а что поделать	15	12	20	35	82	0.32
Ненавижу	35	20	22	10	87	0.33
Сумма	90	54	59	57	260	
Вероятность	0.35	0.21	0.23	0.22		-

Решение

Тогда ожидаемые нами числа:

Отношение /Категория	Школьник и	Студенты	Программист, < 5 лет	Программист, > 5 лет	Сумма	Вероятность
Хороший язык	31,85	19,11	20,93	20,02	91	0,35
Ну а что поделать	29,12	17,472	19,136	18,304	82	0,32
Ненавижу	30,03	18,018	19,734	18,876	87	0,33
Сумма	90	54	59	57	260	
Вероятность	0,35	0,21	0,23	0,22		-

Решение

Посчитаем значение критерия хи-квадрат

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

$$\chi^2 = 35.8$$

$$df = (4 - 1) \cdot (3 - 1) = 6$$

$$P(\chi^2(6) > 33.8) = 0.0004 < 0.01$$

На уровне значимости 0.01 мы отвергаем гипотезу H_0 о независимости

Проблемы с критерием Хи-квадрат

Критерий Хи-квадрат можно применять только тогда, когда ожидаемое число наблюдений в каждой клетке больше 5.

Иначе необходимо использовать точный тест Фишера

Менеджер компании хочет узнать есть ли взаимосвязь между географическим регионом проживания и наличием компьютера Макинтош. Она опрашивает 100 человек и получает следующие данные:

	Есть Мак	Нет Мака	Всего
Северо-восточный	12	14	26
Юго-западный	26	13	39
Средний запад	17	18	35
Total	55	45	100

Для проверки гипотезы о независимости этих двух факторов, чему равны тестовая статистика и соответствующее критическое значение на 5% уровне значимости?

Студенты некоторого университета могут выбирать спецкурсы по собственному желанию. В частности, студенты могут выбрать или не выбрать спецкурс по статистике. Была получена случайная выборка из 100 студентов. В нижеследующей таблице приводятся данные о числе студентов, которые выбирали спецкурс по статистике, и о числе студентов, которые по окончании защитили диплом с отличием.

	Диплом с отличием	Обычный диплом
Изучали статистику	27	23
Не изучали статистику	18	32

Студенты некоторого университета могут выбирать спецкурсы по собственному желанию. В частности, студенты могут выбрать или не выбрать спецкурс по статистике. Была получена случайная выборка из 100 студентов. В нижеследующей таблице приводятся данные о числе студентов, которые выбирали спецкурс по статистике, и о числе студентов, которые по окончании защитили диплом с отличием.

	Диплом с отличием	Обычный диплом
Изучали статистику	27	23
Не изучали статистику	18	32

(2 pts) На 5% уровне значимости протестируйте гипотезу о том, что между изучением статистики и защитой диплома с отличием имеется взаимосвязь. Интерпретируйте Ваш результат в контексте задачи и сформулируйте необходимые условия для применения этого теста.

Студенты некоторого университета могут выбирать спецкурсы по собственному желанию. В частности, студенты могут выбрать или не выбрать спецкурс по статистике. Была получена случайная выборка из 100 студентов. В нижеследующей таблице приводятся данные о числе студентов, которые выбирали спецкурс по статистике, и о числе студентов, которые по окончании защитили диплом с отличием.

	Диплом с отличием	Обычный диплом
Изучали статистику	27	23
Не изучали статистику	18	32

(2 pts) На 5% уровне значимости протестируйте гипотезу о том, что между изучением статистики и защитой диплома с отличием имеется взаимосвязь. Интерпретируйте Ваш результат в контексте задачи и сформулируйте необходимые условия для применения этого теста.

(2 pts) Какой тест следует применить, если вопрос задачи ставится так: на 5% уровне значимости протестируйте гипотезу о том, что доля дипломов с отличием среди студентов, изучавших статистику, выше чем соответствующая доля среди студентов, не изучавших статистику? Сравните с результатом пункта (a).

(2 pts) На 5% уровне значимости протестируйте гипотезу о том, что между изучением статистики и защитой диплома с отличием имеется взаимосвязь. Интерпретируйте Ваш результат в контексте задачи и сформулируйте необходимые условия для применения этого теста.

(2 pts) Какой тест следует применить, если вопрос задачи ставится так: на 5% уровне значимости протестируйте гипотезу о том, что доля дипломов с отличием среди студентов, изучавших статистику, выше чем соответствующая доля среди студентов, не изучавших статистику? Сравните с результатом пункта (а).

(c) **(1 pt)** Основываясь на результатах этого исследования, научный руководитель требует, чтобы его студент записался на курс статистики. С точки зрения статистики объясните прав ли научный руководитель.

Точный тест Фишера

	Исход есть	Исхода нет	Всего
Фактор есть	A	B	A + B
Фактора нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

$$p(table) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

В чем проблема?:

Точный тест Фишера

	Исход есть	Исхода нет	Всего
Фактор есть	A	B	A + B
Фактора нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

$$p(table) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

В чем проблема?:?

Мы получили точечную оценку. Для получения p-value нам надо посчитать весь хвост (односторонний тест) или оба хвоста (двустронний тест)

Точный тест Фишера

**Левый хвост, сложить
вероятности всех
таблиц здесь**

**Правый хвост,
сложить вероятности
всех таблиц здесь**

Все хорошо

**Таблица, перекошенная, как
наша, но в другую сторону**

**Таблицы с
еще более
перекошенно
й в другую
сторону
связью**

	Исход есть	Исход а нет	Всего
Факто р есть	A	B	A + B
Факто ра нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

Наша таблица

	Исход есть	Исход а нет	Всего
Факто р есть	A	B	A + B
Факто ра нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

**Таблицы с еще
более
перекошенной
в нашу сторону
связью**

Пример

	Юноши	Девушки	Всего
На диете	1	9	10
Без диеты	11	3	14
Всего	12	12	24

Гипотеза Н0: Юноши и девушки сидят на диетах одинаково

Гипотеза Н1: Девушки сидят на диетах чае

$$p(\text{table}) = ?$$

Пример

Для вычисление p-value нам надо посчитать еще все таблицы, которые критичнее нашей, в данном случае она одна..

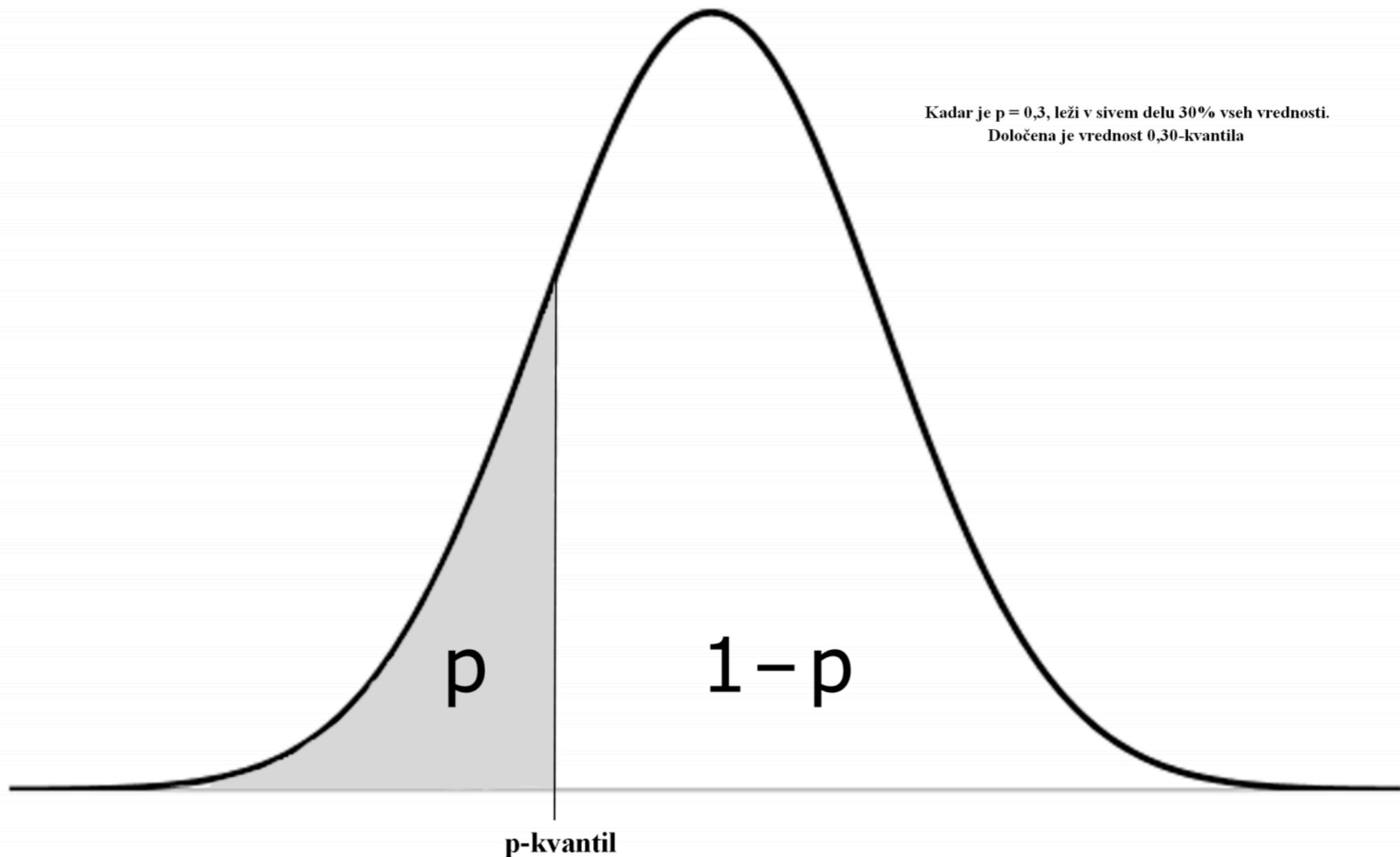
	Юноши	Девушки	Всего
На диете	0	10	10
Без диеты	12	2	14
Всего	12	12	24

$$p(table_1) = ?$$

$$Pvalue = ?$$

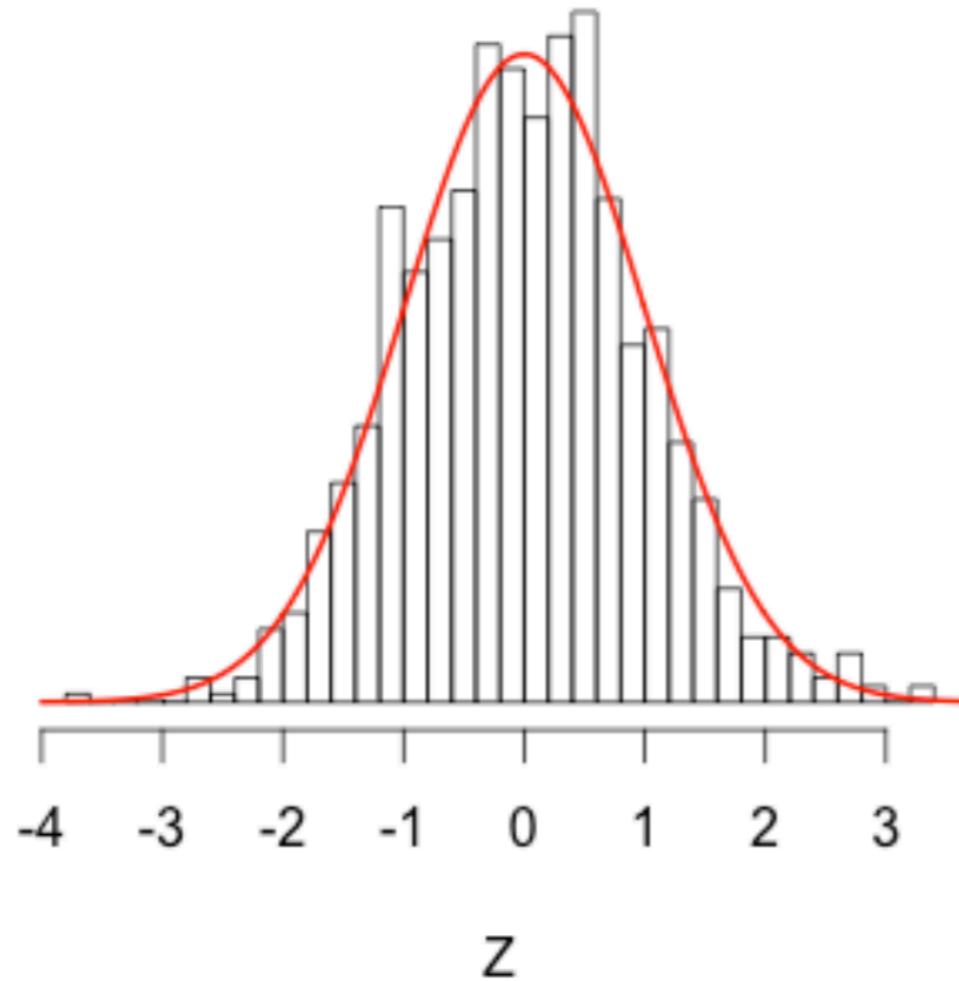
Квантили

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

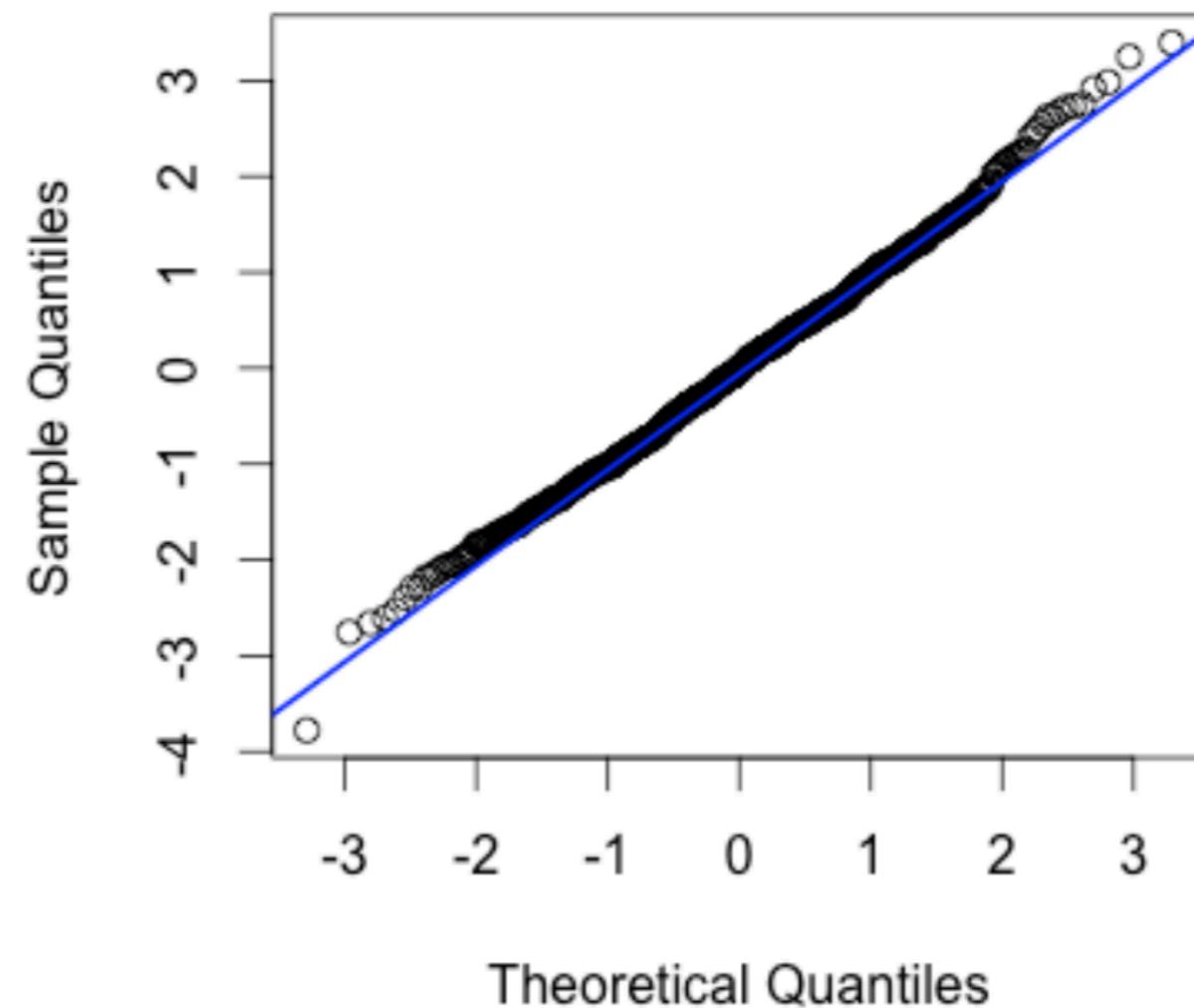


Проверка выборки на нормальность

Gaussian Distribution

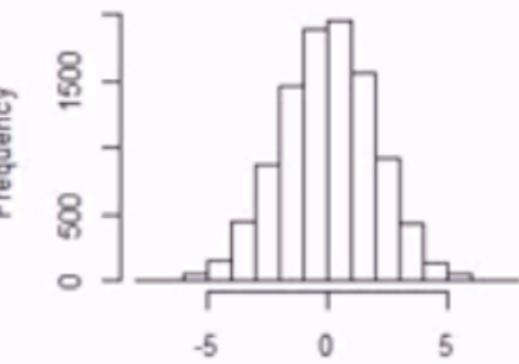


Normal Q-Q Plot

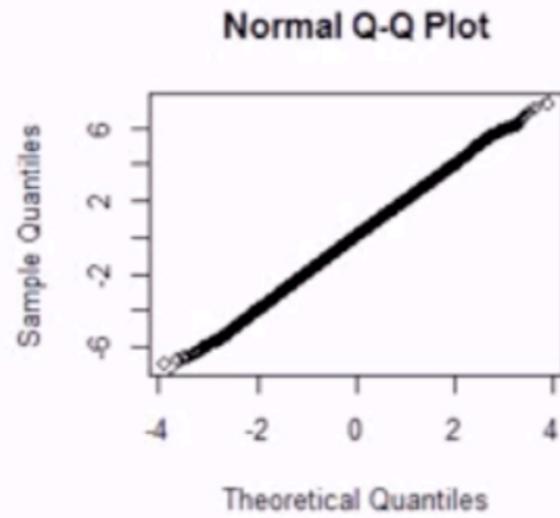


Q-Q Plot

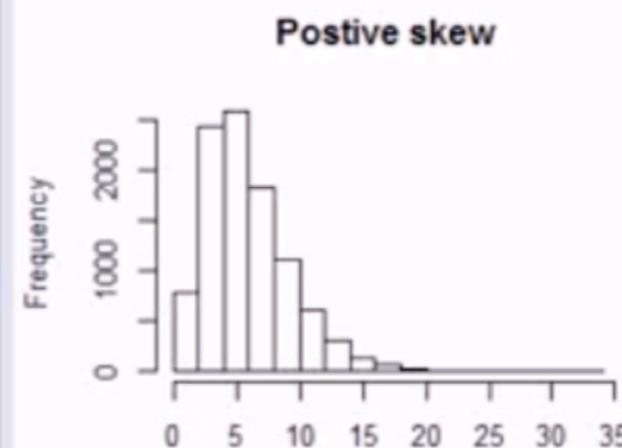
Symmetric distribution



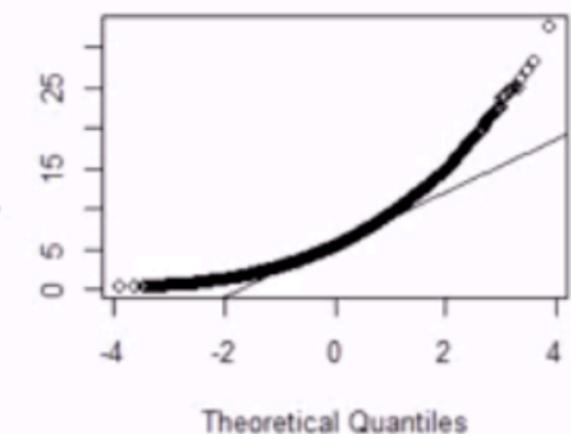
Normal Q-Q Plot



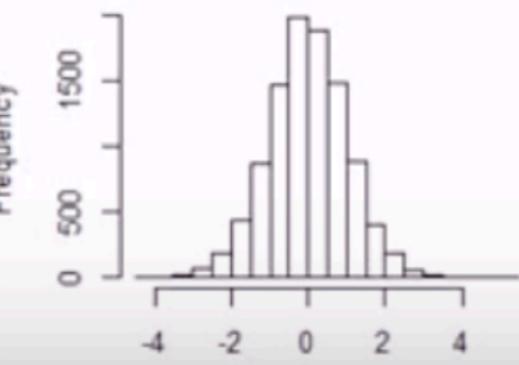
Positive skew



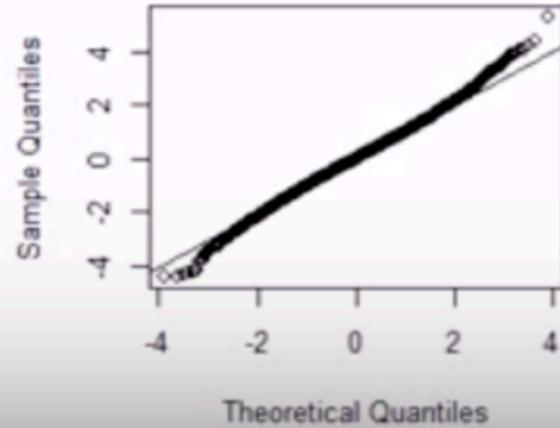
Normal Q-Q Plot



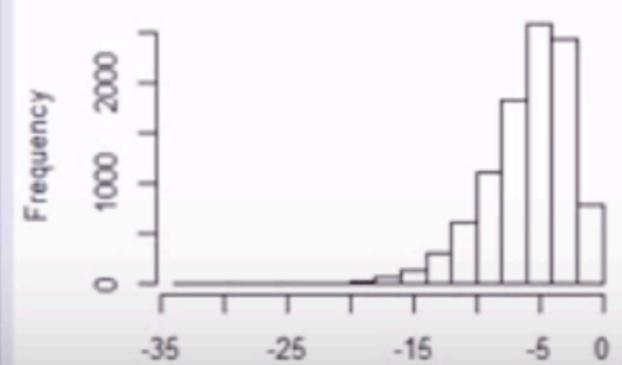
Symmetric with fat tails



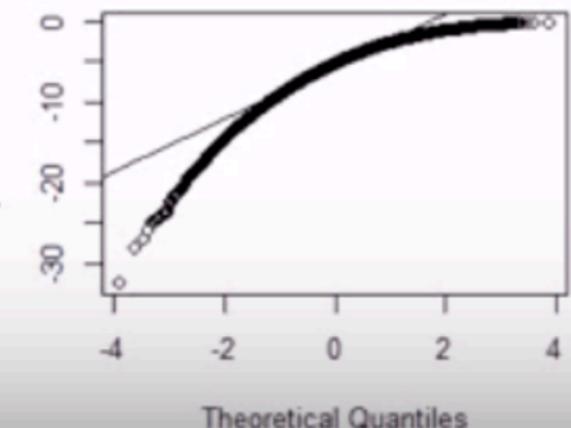
Normal Q-Q Plot



Negative skew



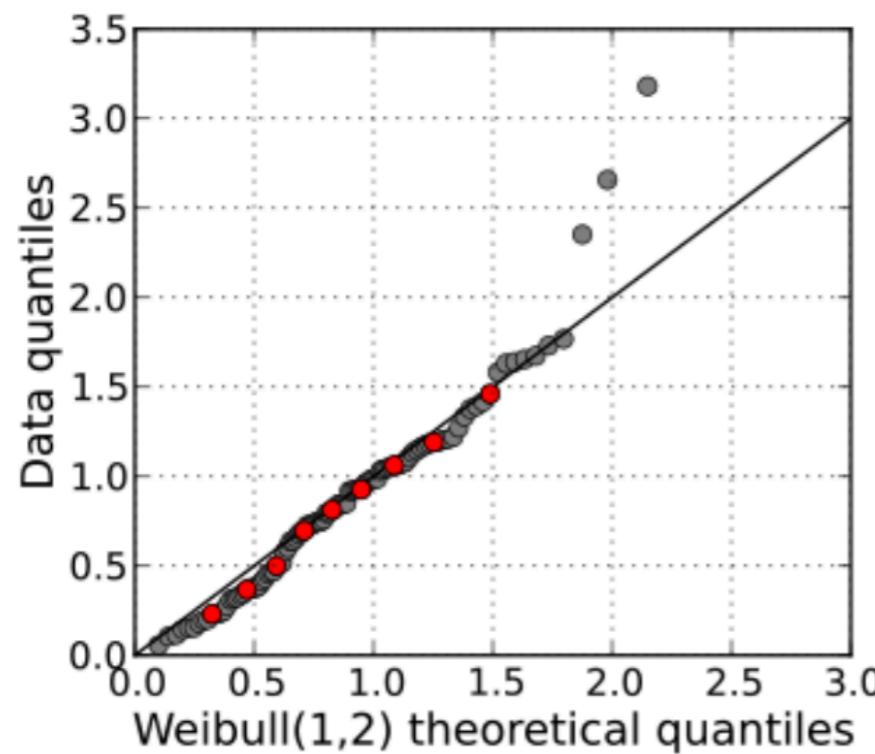
Normal Q-Q Plot



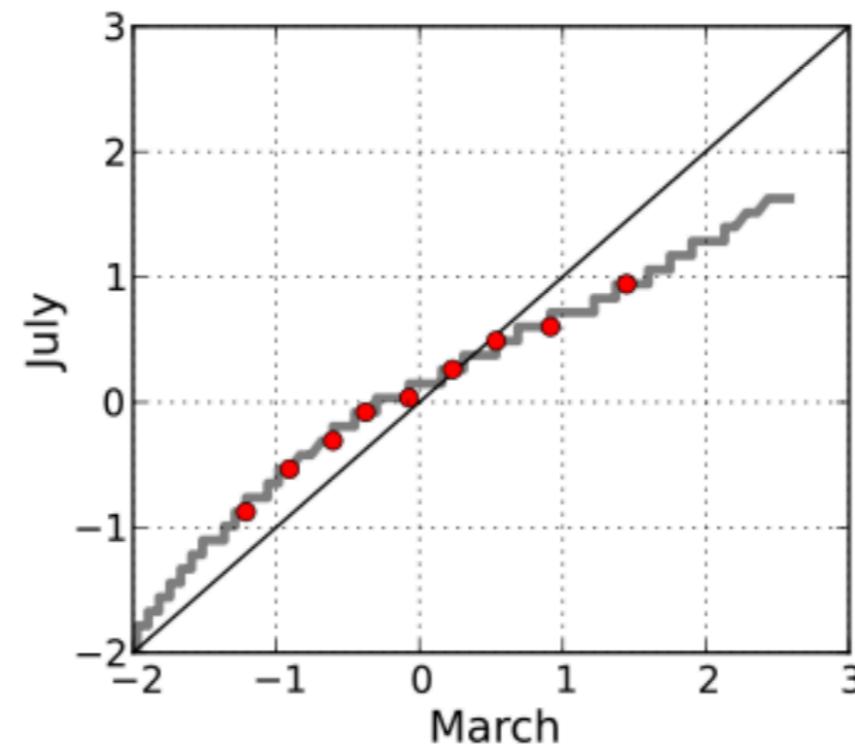
Q-Q Plot

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

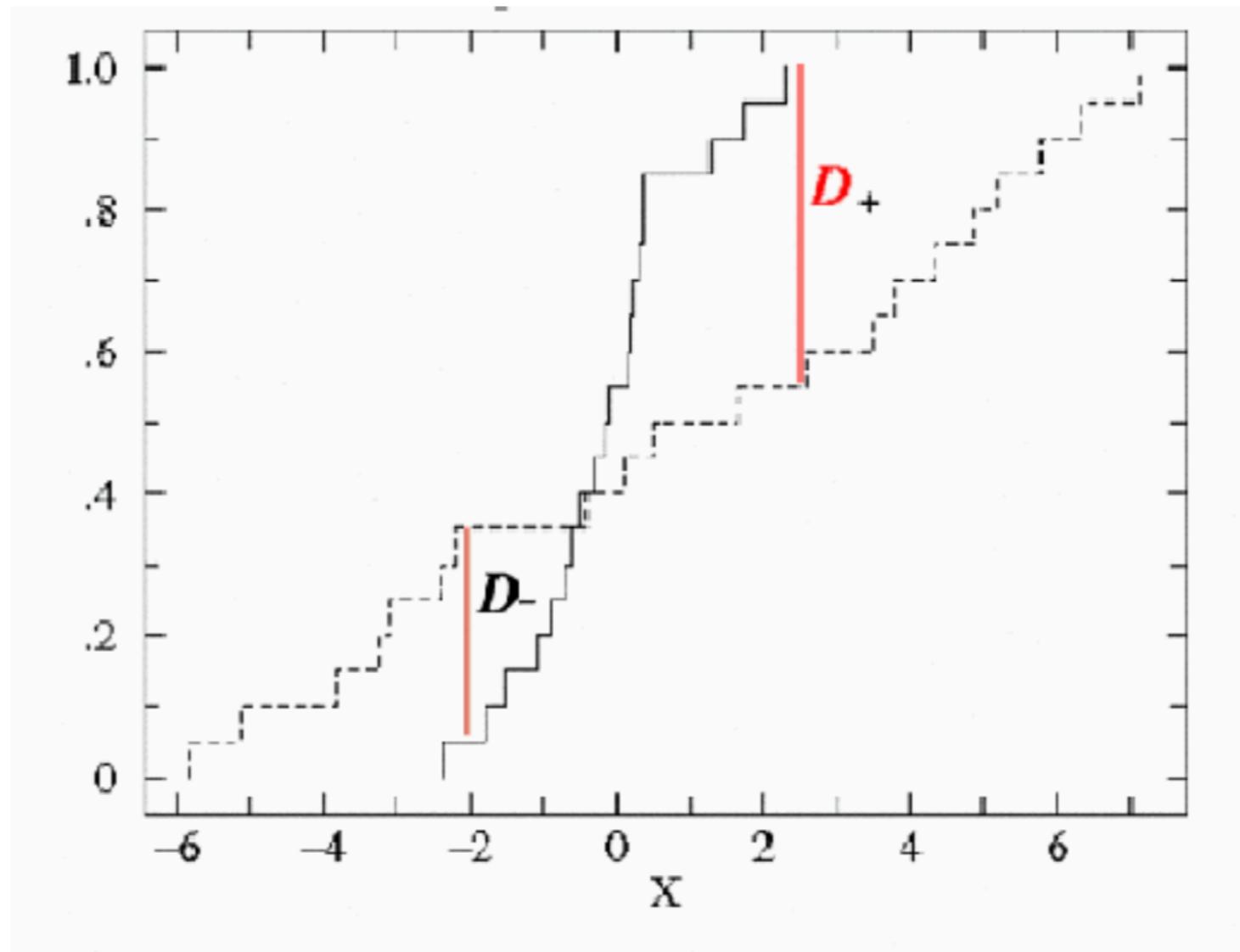


Можно взять другое теоретическое распределение



Можно взять другое теоретическое распределение

Колмогоров-Смирнов



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

При проверке гипотезы о том, что выборка 0.6, 0.27, 0.76, 0.13, 0.73, 0.55, 0.15, 0.11, 0.21, 0.25 была взята из непрерывного равномерного распределения на отрезке [0, 1], чему равно значение тестовой статистики?

x	0.11	0.13	0.15	0.21	0.25	0.27	0.55	0.60	0.73	0.76
eCDF(x)	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
uni-CDF	0.11	0.13	0.15	0.21	0.25	0.27	0.55	0.60	0.73	0.76
d	0.01	0.07	0.15	0.19	0.25	0.33	0.15	0.20	0.17	0.24

Тест Шапиро-Уилка

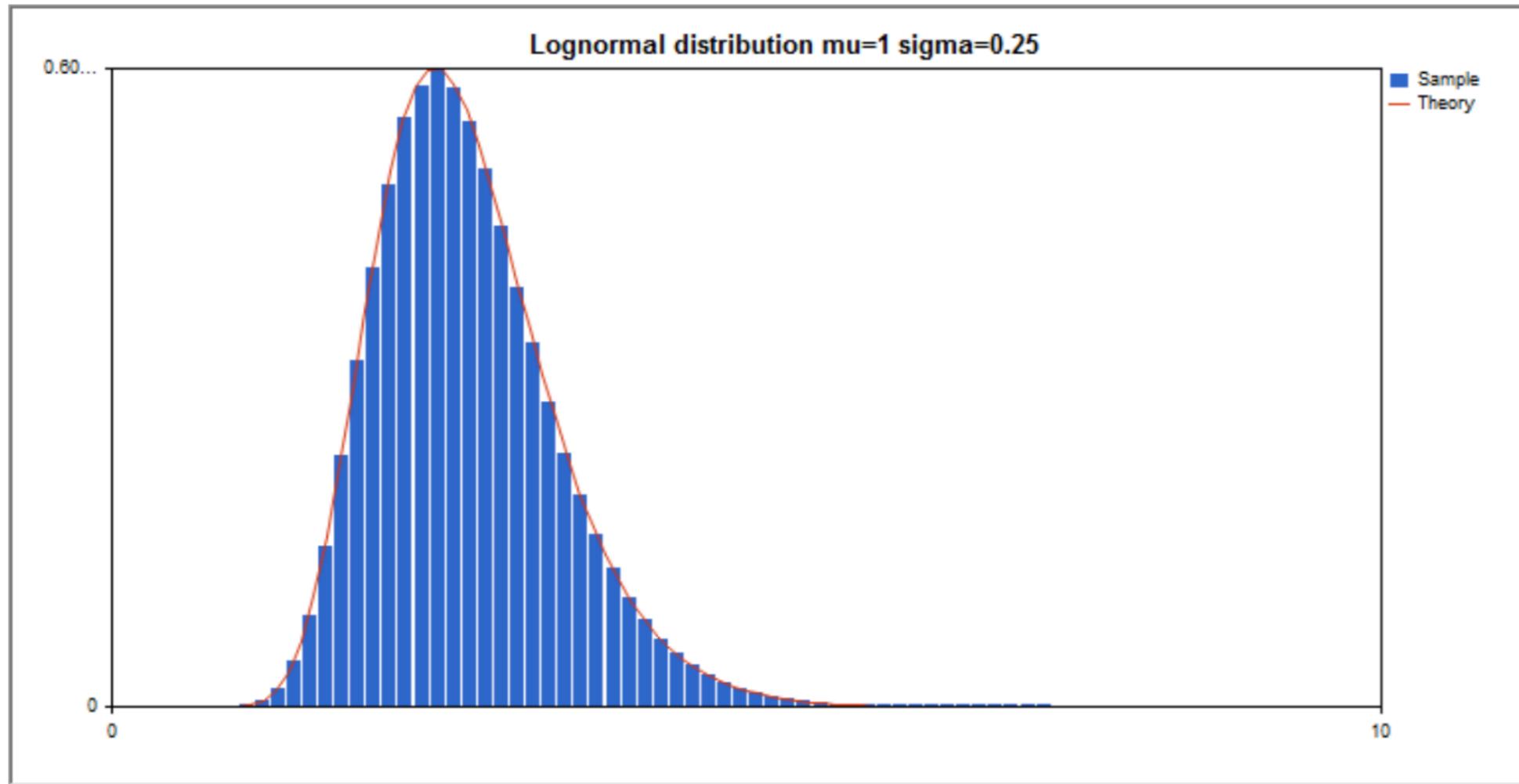
Специально для проверки нормальности распределения малых, численностью от трех до пятидесяти элементов

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 $(X_i, Y_i), i = 1 \dots n$

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

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

Тест 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
 - $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_n\}$ 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 $n \geq 10$

Профессор в бизнес-школе хочет сравнить цены на новые учебники в университетском магазине и соответствующие цены в городе. Профессор случайно выбирает 12 названий учебников для курсов бизнес-школы и сравнивает их цены (в долларах) в университетском магазине и в случайно выбранном магазине в городе. Результаты приведены в таблице

Книга	Цена на кампусе	Цена в городе
1	55.00	50.95
2	47.50	45.75
3	50.50	50.95
4	38.95	38.50
5	58.70	56.25
6	49.90	45.95
7	39.95	40.25
8	41.50	39.95
9	42.25	43.00
10	44.95	42.25
11	45.95	44.00
12	56.95	55.60

Тест Манна-Уитни

Mann-Whitney U -test

Wilcoxon-Mann-Whitney test

- X and Y are two populations
- $H_0 : P(X > Y) = P(Y > X)$
- $H_a : P(X > Y) \neq P(Y > X)$
- U -statistic
 - $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$ are two samples
 - Assign ranks to all the observations $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$
 - 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

Студентов обучают по двум методикам, A и B , причем занятия проходят в отдельных классах. Преподаватель интересуется различиями между двумя методиками и сравнивает оценки студентов. Есть достаточно оснований считать, что распределение оценок далеко от нормального.

Методика А	48	39	37	36	59	69	60	45	38
Методика Б	63	64	79	30	31	42	56	74	67
	78	66	25						

Чему равно значение тестовой статистики?