Статистика и анализ данных на языке R

Непараметрические тесты

Анна Валяева
21.02.2022

Занятие 3

- Нормальность
- Q-Q plot
- Непараметрические тесты

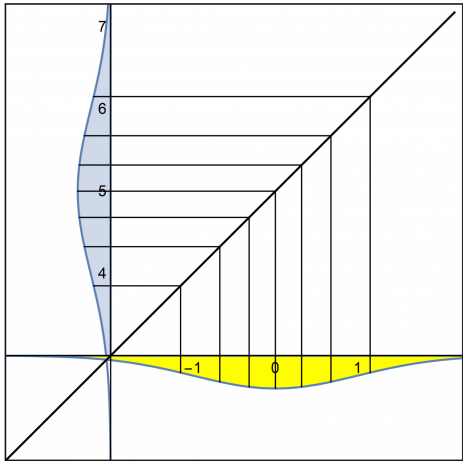
Нормальность данных

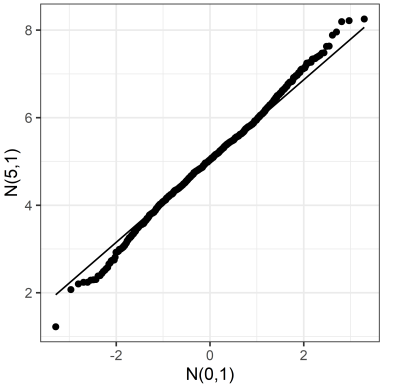
Как понять, что данные распределены нормально?

- Визуализировать распределение с помощью гистограммы / графика плотности
- Оценить визуально с помощью QQ-графика (Q-Q plot)
- Использовать статистистический тест:
  - критерий Колмогорова-Смирнова (K-S test)
  - критерий Шапиро-Уилка
  - критерий Андерсона-Дарлинга
  - критерий Крамера-фон Мизеса
  - ...

Q-Q plot

- квантили эмпирического распределения ~ квантили теоретического распределения

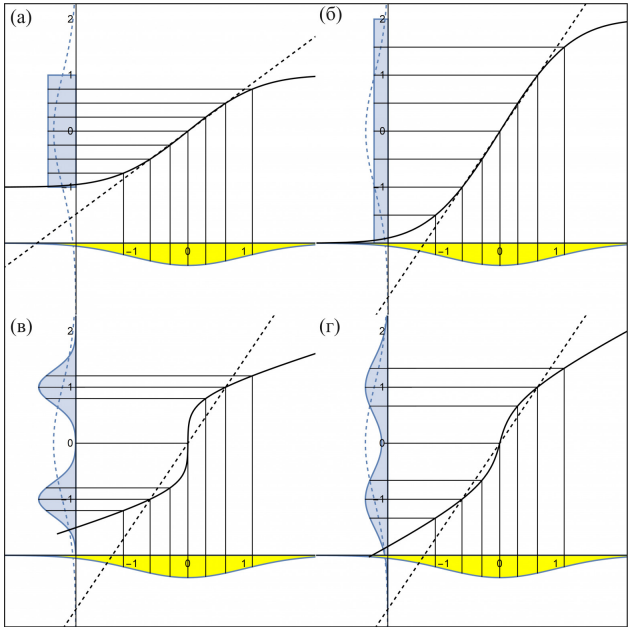




Q-Q Plots

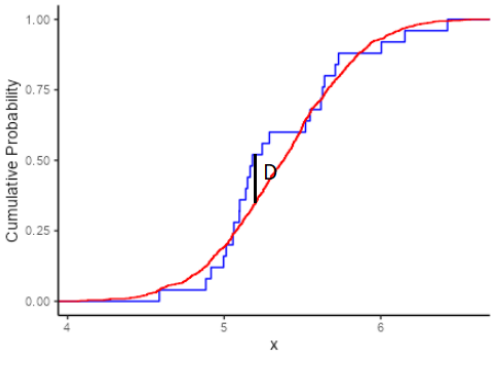
Зависимость теоретических квантилей нормального распределения N(5, 1) от теоретических квантилей стандартного нормального N(0, 1).

Q-Q plot

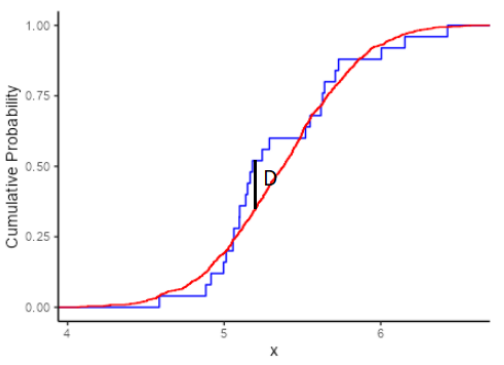


Тест Колмогорова-Смирнова [K-S test]

- критерий согласия: проверяет, что эмпирическое распределение соответствует теоретическому или что два эмпирических распределения совпадают
- H_0 : распределения совпадают
- чувствителен к отличиям в форме распределений и их сдвигу относительно друг друга
- плохо работает на маленьких выборках
- применим только для непрерывных распределений
- D-статистика - максимальная абсолютная разница между двумя кумулятивными функциями распределения (CDF)



Тест Колмогорова-Смирнова [K-S test]



- $D_n = sup|F_n(x) - F(x)|$
- $P(\sqrt{n}D_n \leq x) = H(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x}$
- $H(x)$ - распределение Колмогорова-Смирнова

Распределение Колмогорова-Смирнова

Critical values for $\sup_x |F_n(x) - F(x)|$

	Level of significance, α			
n	0.10	0.05	0.02	0.01
1	0.95000	0.97500	0.99000	0.99500
2	0.77639	0.84189	0.90000	0.92929
3	0.63604	0.70760	0.78456	0.82900
4	0.56522	0.62394	0.68887	0.73424
5	0.50945	0.56328	0.62718	0.66853
6	0.46799	0.51926	0.57741	0.61661
7	0.43607	0.48342	0.53844	0.57581
8	0.40962	0.45427	0.50654	0.54179
9	0.38746	0.43001	0.47960	0.51332
10	0.36866	0.40925	0.45662	0.48893
11	0.35242	0.39122	0.43670	0.46770
12	0.33815	0.37543	0.41918	0.44905
13	0.32549	0.36143	0.40362	0.43247
14	0.31417	0.34890	0.38970	0.41762
15	0.30397	0.33760	0.37713	0.40420
16	0.29472	0.32733	0.36571	0.39201
17	0.28627	0.31796	0.35528	0.38086
18	0.27851	0.30936	0.34569	0.37062
19	0.27136	0.30143	0.33685	0.36117
20	0.26473	0.29408	0.32866	0.35241

Тест Шапиро-Уилка

- проверяет гипотезу, что выборка пришла из нормального распределения
- H_0 : выборка является нормальной
- мощнее, чем тест Колмогорова-Смирнова (то есть с меньшей вероятностью ошибочно принимает H_0)

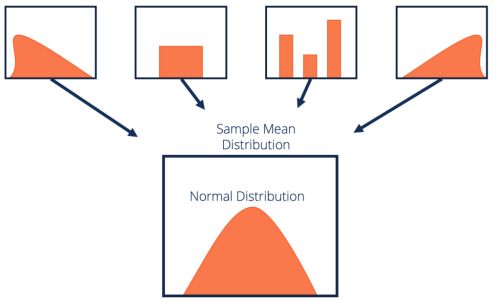
- размер выборки от 3 до 5000

- $W = \frac{\sum_{i=1}^{n} a_i X_{(i)}}{\sum_{i=1}^{n} (X_{(i)} - \overline{X})^2}$

Если n > 30

Если размер выборки достаточно большой, то Центральная Предельная Теорема (ЦПТ) позволяет нам игнорировать условие о нормальности данных.

ЦПТ: распределение значений выборочных средних близко к нормальному вне зависимости от распределения значений генеральной совокупности при условии, что выборки достаточно велики.



Что делать, если данные распределены не нормально?

- Использовать непараметрические тесты:
  - критерий Манна-Уитни для сравнения независимых выборок,
  - критерий Уилкоксона для сравнения парных выборок.

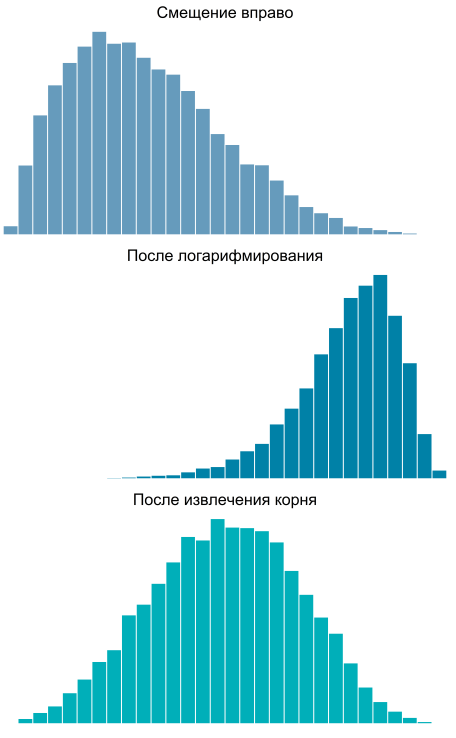
- Преобразовать данные таким образом, чтобы их распределение стало приблизительно нормальным.

Преобразование данных

Если данные смещены **вправо** (есть тяжелый правый хвост), поможет:

- логарифмирование или
- извлечение квадратного корня

Если после логарифмирования распределение данных становятся нормальным, значит, исходные данные были распределены лог-нормально.

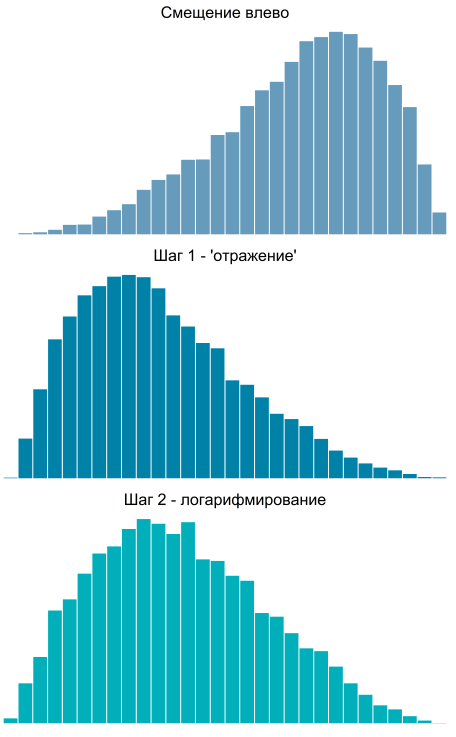


Преобразование данных

Если данные смещены **влево** (есть тяжелый левый хвост), поможет:

- "зеркальное отражение" данных и
- все то же, что и со смещенными вправо данными.

$$Y_i = (X_{max} + 1) - X_i$$



Непараметрические тесты

Непараметрические тесты

U-критерий Манна—Уитни

То же, что и

- Wilcoxon rank-sum test,
- unpaired two-samples Wilcoxon test,
- Mann–Whitney–Wilcoxon (MWW).

T-критерий Вилкоксона

То же, что и

- Wilcoxon signed-rank test,
- paired samples Wilcoxon test.

Непараметрические тесты

U-критерий Манна—Уитни

- проверяет, что значения в одной из выборок в среднем больше, чем в другой
- используется на 2-х независимых выборках
- H_0 : выборки распределены одинаково
- H_A : выборки распределены неодинаково
- использует U-статистику

T-критерий Вилкоксона

- проверяет, что сдвиги (разница) между 2 связными выборками в ту или иную сторону происходят случайно
- H_0 : сдвиги распределены симметрично около нуля
- H_A : сдвиги не распределены симметрично около нуля
- использует W-статистику

U-критерий Манна—Уитни

Алгоритм:

- $X_1, \ldots, X_n и Y_1, \ldots, Y_m$ - 2 независимые выборки
- приписываем ранги всем наблюдениям из объединенных выборок
- R_1 - сумма рангов наблюдений из выборки 1
- R_2 - сумма рангов наблюдений из выборки 2
- $U_1 = R_1 - \frac{n(n+1)}{2}$, $U_2 = R_2 - \frac{m(m+1)}{2}$
- $U = \min(U_1, U_2)$
- $U \sim N(\mu, \sigma)$, $\mu = \frac{nm}{2}$, $\sigma = \sqrt{\frac{nm(n+m+1)}{12}}$, если $n, m \geq 10$

Если размеры выборок меньше 10, лучше не использовать нормальную аппроксимацию - не считать z-значение, а использовать табличные критические значения U-статистики.

Пример 1

У пяти легкоатлетов был замерен пульс во время отдыха: 60, 58, 67, 61 и 59. И так же во время отдыха был замерен пульс у семи людей, не занимающихся спортом: 83, 60, 75, 91, 82, 71 и 84. Отличается ли пульс у легкоатлетов и людей, не занимающихся спортом?

Пример 1

- 58, 59, 60, 60, 61, 67, 71, 75, 82, 83, 84, 91
- 1, 2, 3.5, 3.5, 5, 6, 7, 8, 9, 10, 11, 12
- $R_1 = 1 + 2 + 3.5 + 5 + 6 = 17.5$
- $R_2 = 3.5 + 7 + 8 + 9 + 10 + 11 + 12 = 60.5$
- $U_1 = R_1 - \frac{n(n+1)}{2} = 17.5 - \frac{5(5+1)}{2} = 2.5$
- $U_2 = R_2 - \frac{m(m+1)}{2} = 60.5 - \frac{7(7+1)}{2} = 32.5$
- $\min(U_1, U_2) = 2.5$

Пример 1

- $\mu = \frac{nm}{2} = \frac{5 \times 7}{2} = 17.5$
- $\sigma = \sqrt{\frac{nm(n+m+1)}{12}} = \sqrt{\frac{5 \times 7(5+7+1)}{12}} = 6.16$
- $z = \frac{U - \mu}{\sigma} = \frac{2.5 - 17.5}{6.16} = -2.43$
- $P(z < -2.43) = 0.0075$

Если размеры выборок меньше 10, лучше не использовать нормальную аппроксимацию - не считать z-значение, а использовать табличные критические значения U-статистики.

T-критерий Вилкоксона

Алгоритм:

- $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$ - 2 зависимые выборки
- считаем сдвиги $d_i = |X_i - Y_i|$, выбрасываем наблюдения с $d_i = 0$
- сортируем d_i по возрастанию
- $W = \sum sgn(X_i - Y_i) * R_i$, где R_i - ранг d_i
- $W \sim N(\mu, \sigma)$, $\mu = 0$, $\sigma = \sqrt{\frac{n(n+1)(2n+1)}{6}}$, если $n \geq 10$

Если размеры выборок меньше 10, лучше не использовать нормальную аппроксимацию - не считать z-значение, а использовать табличные критические значения W-статистики.

Пример 2

Некоторое вещество, подавляющее аппетит, было протестировано на 10 крысах. Вес животных был измерен до начала добавления исследуемого вещества в корм и после двух недель эксперимента. Используя собранные данные, можно ли утверждать, что исследуемое вещество влияет на аппетит и набор веса?

Крыса	ДО	ПОСЛЕ	Крыса	ДО	ПОСЛЕ
1	400	394	6	393	380
2	405	411	7	361	352
3	388	354	8	376	364
4	377	389	9	360	357
5	372	372	10	420	403

Пример 2

Крыса	ДО	ПОСЛЕ	d	|d|	sign	rank	sign * rank
1	400	394	-6	6	-1	2,5	-2,5
2	405	411	6	6	1	2,5	2,5
3	388	354	-34	34	-1	9	-9
4	377	389	12	12	1	5,5	5,5
5	372	372	0	0	0	0	0
6	393	380	-13	13	-1	7	-7
7	361	352	-9	9	-1	4	-4
8	376	364	-12	12	-1	5,5	-5,5
9	360	357	-3	3	-1	1	-1
10	420	403	-17	17	-1	8	-8

$W = -29,\ \sigma = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{10(10+1)(20+1)}{6}} = 19.62$

$z = \frac{W-\mu}{\sigma} = \frac{-29-0}{19.62} = -1.48$

$P(z < -1.48) = 0.0694$

Пояснения про двустороннюю гипотезу

Обратите внимание, что в Примерах 1 и 2 вопросы были поставлены так, что альтернативную гипотезу нужно было формулировать как двустороннюю. Но когда мы искали p-значение в таблице, то обращали внимание только на один - левый - хвост распределения. В то время как при двусторонней альтернативной гипотезе нам интересны оба хвоста - и левый, и симметричный правый - поэтому для получения p-значения нам нужно полученное из таблицы значение умножить еще на 2.