



Статистика и анализ данных на языке R

Статистический вывод, t-тест

*Анна Валяева
14.02.2022*

Занятие 2

- Идея статистического вывода
- р-значение, или p-value
- t-тест
- нормальность

Статистический вывод

[Statistical inference]

Алгоритм тестирования гипотез

1. Формулируем нулевую и альтернативную гипотезу

- H_0 - принимаемое по умолчанию скептическое предположение о том, что эффекта нет.
- H_1 или H_A - альтернативное предположение о том, что эффект есть.

2. Подбираем уровень значимости, на котором планируем отвергать нулевую гипотезу.

- Обычно $\alpha = 0.05$.
- Это значение вероятности "редкого" события, при котором это событие уже можно считать неслучайным.
- Также это допустимая вероятность ошибки первого рода (ложноположительного решения, false positive), то есть вероятность отклонить нулевую гипотезу, когда на самом деле она верна.

Алгоритм тестирования гипотез

3. Выбираем статистический тест и считаем тест-статистику.
4. Считаем p -значение.
5. Делаем вывод, сравнивая p -значение и уровень значимости.

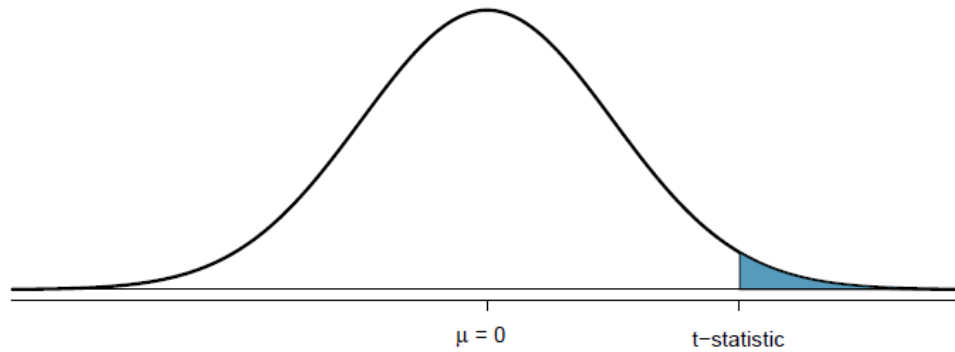
p-значение

- вероятность получить результат такой же, как сейчас видим, или еще экстремальней, **если нулевая гипотеза верна**.

Если p-значение меньше уровня значимости, то нулевую гипотезу мы отвергаем.

Если p-значение больше уровня значимости, то у нас не хватает оснований, чтобы отвергнуть нулевую гипотезу.

Распределение тест-статистики



Критерий Стьюдента

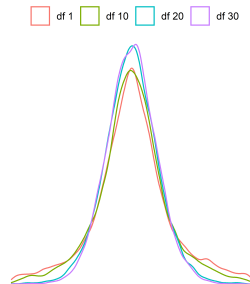
t-test

Когда используется t-тест?

- когда хотим проверить разницу между **выборочными средними**
 - если сравниваем 2 выборки, используем двухвыборочный тест
 - если сравниваем среднее 1 выборки с некоторым числом, используем одновыборочный тест
- когда выборка случайна и индивидуальные наблюдения независимы
- когда предполагаем, что генеральная совокупность распределена **нормально**
- когда предполагаем равенство дисперсий двух генеральных совокупностей (*)
- когда не знаем дисперсию генеральной совокупности

t-распределение

- непрерывное
- симметричное
- похоже на нормальное, но хвосты тяжелее
- форма зависит от числа степеней свободы (df)
- число степеней свободы зависит от размера выборки
- обычно $df = n - 1$
- чем больше степеней свободы, тем более похоже на нормальное



Одновыборочный t-критерий

- сравниваем выборочное среднее с некоторым числом
- $H_0 : \bar{X} = \mu_0$
- двусторонняя альтернативная гипотеза: $H_1 : \bar{X} \neq \mu_0$
- односторонняя альтернативная гипотеза: $H_1 : \bar{X} > \mu_0$ или $\bar{X} < \mu_0$
- тест-статистика: $t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$

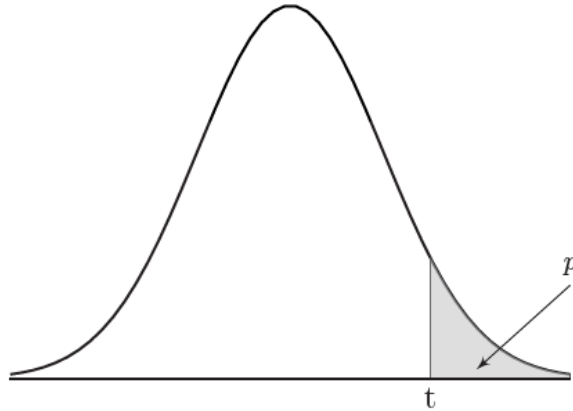
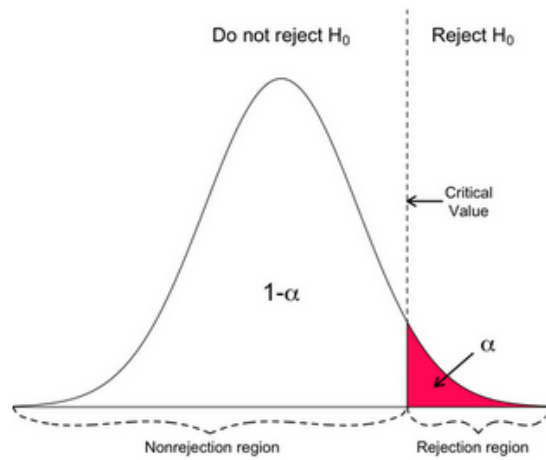


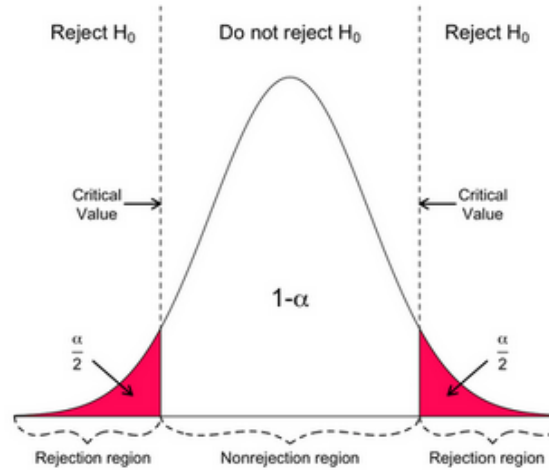
Таблица верхних критических значений t-распределения

<i>df</i>	Tail Probability <i>p</i>											
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.00	1.38	1.96	3.08	6.31	12.71	15.89	31.82	63.66	127.32	318.31	636.62
2	0.82	1.06	1.39	1.89	2.92	4.30	4.85	6.96	9.93	14.09	22.33	31.60
3	0.77	0.98	1.25	1.64	2.35	3.18	3.48	4.54	5.84	7.45	10.21	12.92
4	0.74	0.94	1.19	1.53	2.13	2.78	3.00	3.75	4.60	5.60	7.17	8.61
5	0.73	0.92	1.16	1.48	2.02	2.57	2.76	3.37	4.03	4.77	5.89	6.87
6	0.72	0.91	1.13	1.44	1.94	2.45	2.61	3.14	3.71	4.32	5.21	5.96
7	0.71	0.90	1.12	1.42	1.90	2.37	2.52	3.00	3.50	4.03	4.79	5.41
8	0.71	0.89	1.11	1.40	1.86	2.31	2.45	2.90	3.35	3.83	4.50	5.04
9	0.70	0.88	1.10	1.38	1.83	2.26	2.40	2.82	3.25	3.69	4.30	4.78
10	0.70	0.88	1.09	1.37	1.81	2.23	2.36	2.76	3.17	3.58	4.14	4.59
11	0.70	0.88	1.09	1.36	1.80	2.20	2.33	2.72	3.11	3.50	4.03	4.44
12	0.69	0.87	1.08	1.36	1.78	2.18	2.30	2.68	3.06	3.43	3.93	4.32
13	0.69	0.87	1.08	1.35	1.77	2.16	2.28	2.65	3.01	3.37	3.85	4.22
14	0.69	0.87	1.08	1.34	1.76	2.15	2.26	2.62	2.98	3.33	3.79	4.14
15	0.69	0.87	1.07	1.34	1.75	2.13	2.25	2.60	2.95	3.29	3.73	4.07
16	0.69	0.86	1.07	1.34	1.75	2.12	2.23	2.58	2.92	3.25	3.69	4.01
17	0.69	0.86	1.07	1.33	1.74	2.11	2.22	2.57	2.90	3.22	3.65	3.96
18	0.69	0.86	1.07	1.33	1.73	2.10	2.21	2.55	2.88	3.20	3.61	3.92

Односторонняя гипотеза



Двусторонняя гипотеза



Пример 1

У 16 курильщиков был измерен объем легких. Среднее значение оказалось равным 5.0 литров, стандартное отклонение - 0.7 литров. Значительно ли снижен объем легких курильщика по сравнению со средним по популяции не-курильщиков, равным 5.4 литра?

$$\bar{X} = 5.0, s = 0.7, n = 16$$

$$\mu_0 = 5.4$$

$$H_0 : \mu = 5.4, H_1 : \mu < 5.4$$

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{5.0 - 5.4}{\frac{0.7}{\sqrt{16}}} = -2.29$$

$$t(15) = 1.75 \text{ при } \alpha = 0.05$$

$$|-2.29| > t(15), \text{ отвергаем } H_0$$

$$p\text{-value} = 0.0185$$

Пример 2

У 16 курильщиков был измерен объем легких. Среднее значение оказалось равным 5.0 литров, стандартное отклонение - 0.7 литров. Отличается ли объем легких курильщика от среднего по популяции некурильщиков, равного 5.4 литра?

$$\bar{X} = 5.0, s = 0.7, n = 16$$

$$\mu_0 = 5.4$$

$$H_0 : \mu = 5.4, H_1 : \mu \neq 5.4$$

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{5.0 - 5.4}{\frac{0.7}{\sqrt{16}}} = -2.29$$

$$t(15) = 2.13 \text{ при } \alpha = 0.025$$

$$|-2.29| > t(15), \text{ отвергаем } H_0$$

$$p\text{-value} = 0.037$$

Двухвыбросный t-тест для независимых выборок

- проверяем, равны ли средние генеральных совокупностей, из которых были взяты выборки
- $H_0 : \mu_1 - \mu_2 = 0, H_1 : \mu_1 - \mu_2 \neq 0$
- наблюдения из двух выборок не связаны и выбраны из своих совокупностей независимо
- генеральные совокупности имеют приблизительно нормальное распределение
- дисперсии двух совокупностей приблизительно равны

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

где s_p^2 - объединенная дисперсия

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Пример 3

Две группы спортсменов по 10 человек в каждой - лыжники и футболисты - выполняли различные физические упражнения, чтобы определить в какой физической форме они находятся. Результаты выполнения всех упражнений были суммированы в виде единого показателя, который принимает значения от 0 до 100. Опыт использования подобного метода оценки с этим способом подсчета результатов показывает, что эти показатели распределены в генеральной совокупности приблизительно нормально.

Для лыжников средний показатель оказался равен 87.95, дисперсия - 32.38. Для футболистов среднее равно 85.19, дисперсия - 31.18.

Кто находится в лучшей спортивной форме, лыжники или футболисты?

Пример 3

Для лыжников средний показатель оказался равен 87.95, дисперсия - 32.38. Для футболистов среднее равно 85.19, дисперсия - 31.18.

Кто находится в лучшей спортивной форме, лыжники или футболисты?

$$s_p^2 = 31.78$$

$$t = 1.01$$

$$t(18) = 2.10 \text{ при } \alpha = 0.025$$

$$|1.01| < t(18), \text{ нет оснований отвергнуть } H_0$$

$$p\text{-value} = 0.328$$

Двухвыбросочный t-тест для парных выборок

- наблюдения в двух выборках каким-то образом связаны
- проверяем, есть ли сдвиг между средними генеральных совокупностей, из которых были взяты выборки
- $H_0 : \mu_1 - \mu_2 = 0, H_1 : \mu_1 - \mu_2 \neq 0$

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}}$$

$\bar{d}$ - средняя разница между парными наблюдениями, образует "новую" выборку

s_d - стандартное отклонение разниц

$df = n - 1$, где n - число пар

Однородность дисперсий

t-критерий для выборок с неравной дисперсией

Функция `t.test()` в R по умолчанию применяет t-тест Велча (Welch's t-test).

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

и более сложная формула для расчета количества степеней свободы.

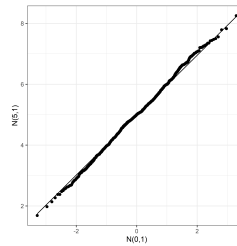
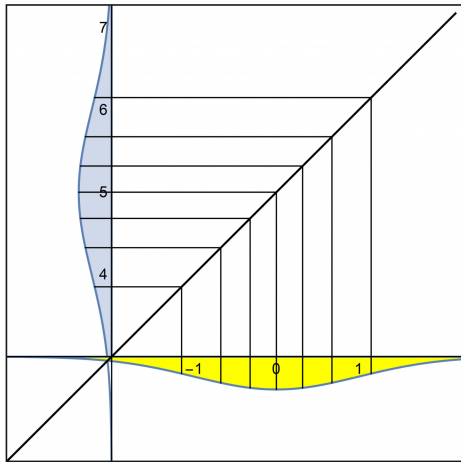
Нормальность данных

Как понять, что данные распределены нормально?

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

Q-Q plot

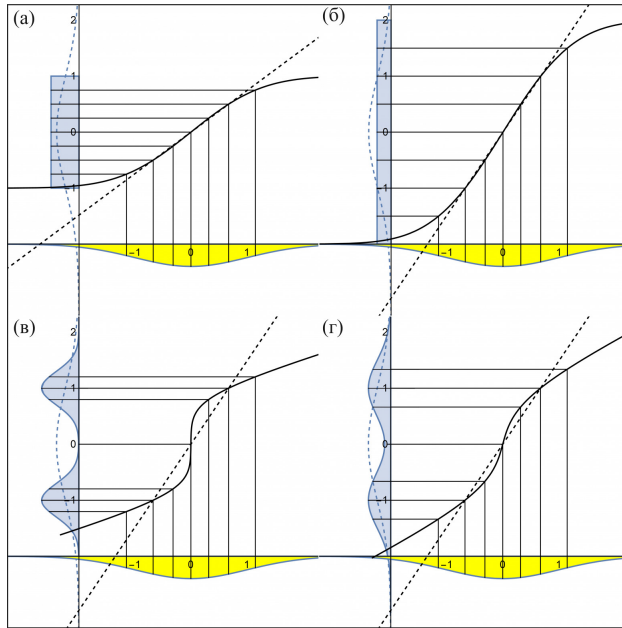
- квантили эмпирического распределения ~ квантили теоретического распределения



Q-Q Plots

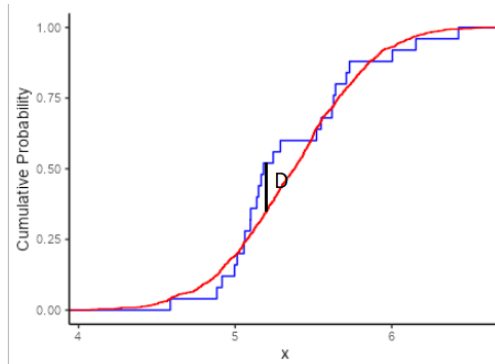
Зависимость теоретических квантилей нормального распределения $N(5, 1)$ от теоретических квантилей стандартного нормального $N(0, 1)$.

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 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

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?

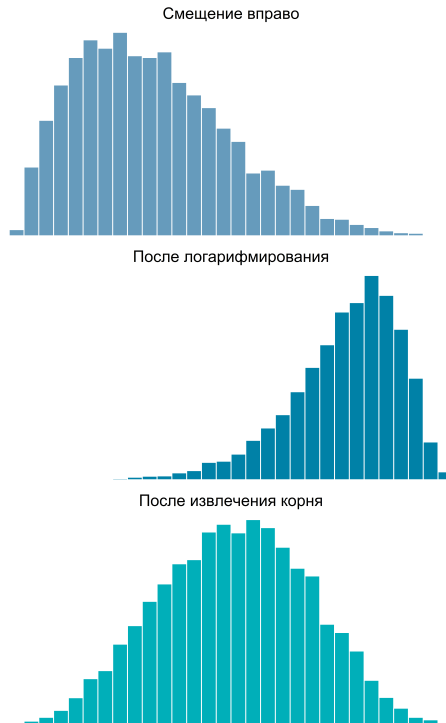
- Использовать непараметрические тесты:
 - критерий Манна-Уитни для сравнения независимых выборок,
 - критерий Уилкоксона для сравнения парных выборок.
- Преобразовать данные таким образом, чтобы их распределение стало приблизительно нормальным.

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.

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

