ВЕБИНАРНЫЙ ФОРМАТ
Библиотеки Python для Data Science: продолжение

Урок 2.
Анализ данных и проверка статистических гипотез.

Светлана Медведева

План урока

- Теория вероятностей и математическая статистика
- Что такое статистическая гипотеза?
- Проверка статистических гипотез
- Критерий Шапиро-Уилка
- Критерий Стьюдента (t-test), двухвыборочный
- Критерий хи-квадрат (критерий согласия Пирсона)
- Доверительные интервалы

Теория вероятностей и математическая статистика

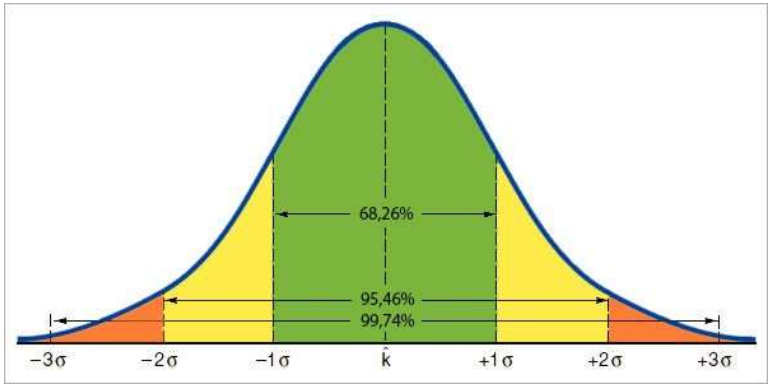
Теория вероятностей - раздел математики, изучающий случайные события, случайные величины, их свойства и операции над ними.

Математическая статистика и анализ данных пытаются по свойствам конечных выборок определить свойства случайной величины, чтобы понять, как она будет вести себя в будущем.

Что такое статистическая гипотеза?

Статистические гипотезы:

1. Нулевая
2. Альтернативная



$$x^n = (x_1, .., x_n),\ x^n \in X,\ X \sim P$$
$$H_0 : P \in \omega$$
$$H_1 : P \notin \omega$$
$$T(x^n),\ T(x^n) \sim F_0(t) \mid H_0,\ T(x^n) \nsim F_0(t) \mid H_1$$

H_0 - нулевая гипотеза

H_1 - альтернативная гипотеза

X - случайная величина

x^n - выборка размера n из случайной величины X

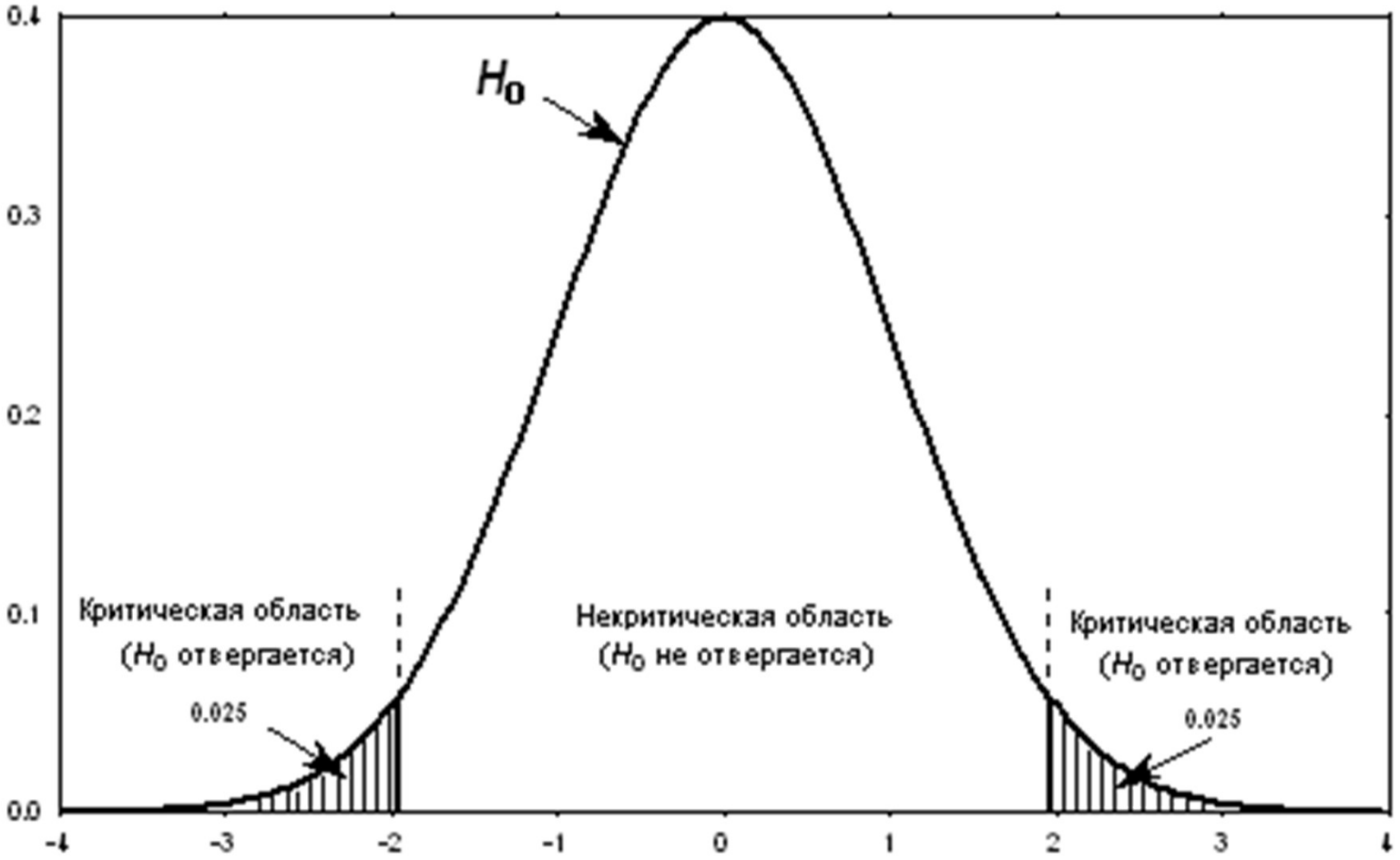
P - некоторое распределение случайной величины X

ω - некоторое семейство распределений

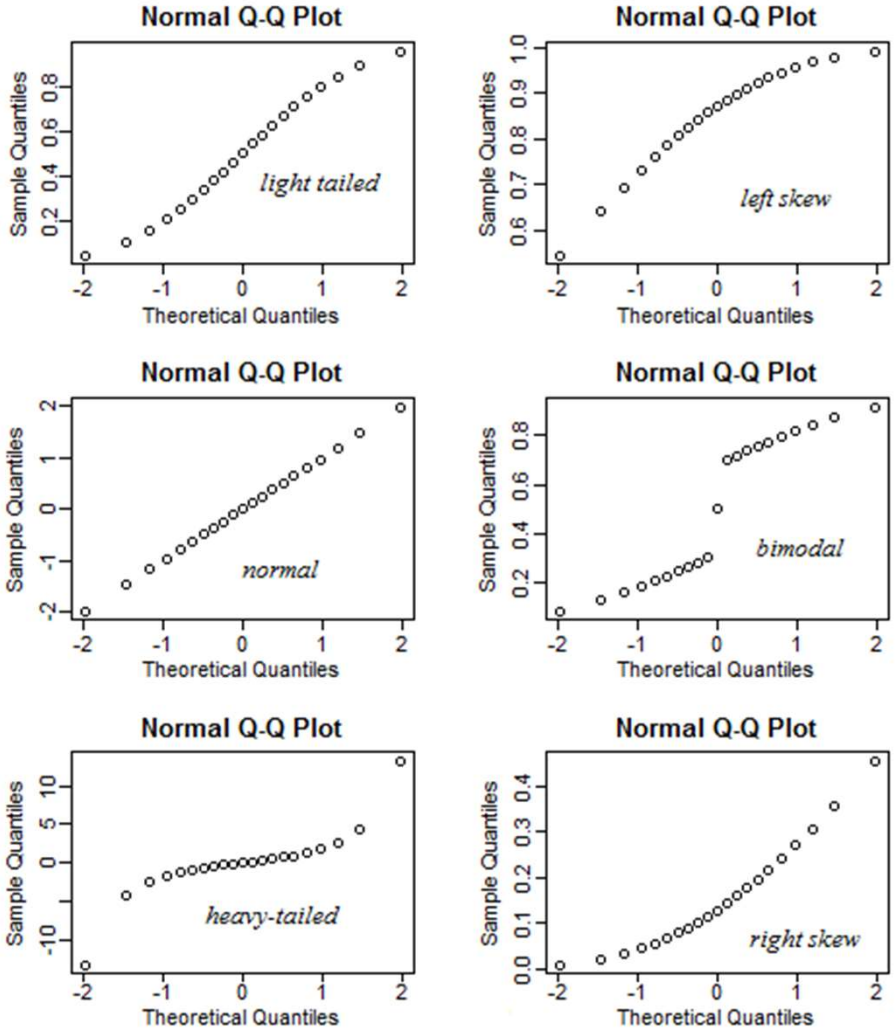
$T(x^n)$ - статистика от выборки x^n

$F_0(t)$ - нулевое распределение статистики

Критическая область



Графики Q-Q (квантиль-квантиль)

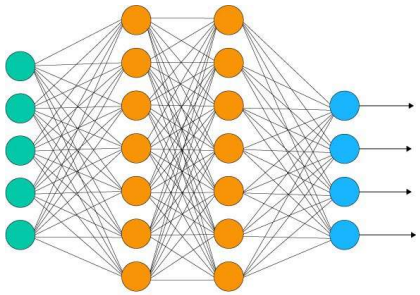


Проверка статистических гипотез

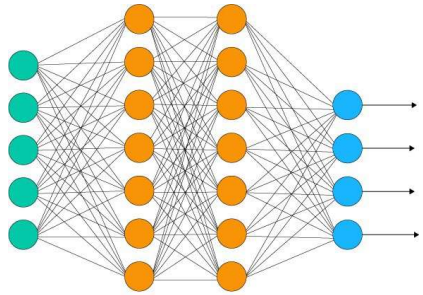
1. Сформулировать гипотезы $H0$ и $H1$
2. Выбрать подходящий статистический критерий, исходя из сформулированных гипотез, размера выборки(ок) и т.д.
3. Зафиксировать уровень значимости α
4. На множестве значений выбранной статистики T определить критическую область Ω_α наименее вероятных значений, таких, что $P(T \in \Omega_\alpha | H0) = \alpha$, как правило, рассматривается двусторонняя критическая область: $(-\infty; x_{\alpha/2}) \cup (x_{1-\alpha/2}; +\infty)$
5. Рассчитать значение статистики T и достигаемые уровень значимости $p-value* = P(T \geq t | H0)$
6. Если $p-value < \alpha$, $H0$ отвергается в пользу $H1$, т.к вероятность получить такие данные (выборку), при верности $H0$, крайне мала.

Ошибки первого и второго рода

H_0	верная	ложная
принимается	H_0 верно принята	H_0 неверно принята (ошибка второго рода)
отклоняется	H_0 неверно отвергнута (ошибка первого рода)	H_0 верно отвергнута



собака



кошка

Ошибка 1-го рода

Ошибка 2-го рода

Виды статистических тестов

1. Тесты нормальности

- Критерий Шапиро-Уилка

2. Корреляционные тесты

- Коэффициент корреляции Пирсона
- Тест хи-квадрат

3. Параметрические статистические проверки гипотез

- Студенческий т-тест
- Парный студенческий т-тест
- Анализ дисперсионного теста (ANOVA)
- Повторные измерения ANOVA Test

4. Непараметрические статистические проверки гипотез

- U-тест Манна-Уитни
- Тест Уилкоксона со знаком
- Kruskal-Wallis H Test
- Тест Фридмана

Виды статистических тестов

1. Тесты нормальности

- Критерий Шапиро-Уилка (*scipy.stats.shapiro*)
- D'Agostino's K-squared test (*scipy.stats.normaltest*)
- Тест Андерсона-Дарлинга (*scipy.stats.anderson*)

2. Корреляционные тесты

- Коэффициент корреляции Пирсона (*scipy.stats.pearsonr*)
- Тест хи-квадрат (*scipy.stats.chi2_contingency*)

3. Параметрические статистические проверки гипотез

- Student T-test (*scipy.stats.ttest_ind*)
- Парный Student T-test (*scipy.stats.ttest_rel*)
- Анализ дисперсионного теста (ANOVA) (*scipy.stats.f_oneway*)

4. Непараметрические статистические проверки гипотез

- U-тест Манна-Уитни (*scipy.stats.mannwhitneyu*)
- Kruskal-Wallis H Test (*scipy.stats.kruskal*)
- Тест Фридмана (*scipy.stats.friedmanchisquare*)

Критерий Шапиро-Уилка

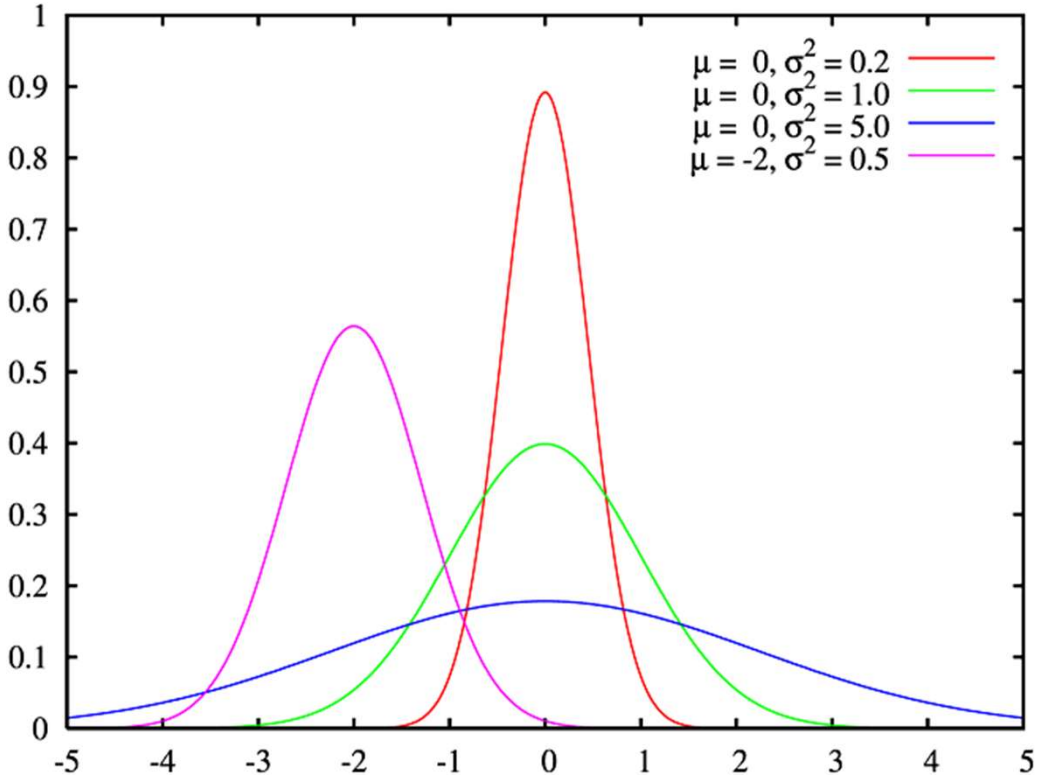
Проверка распределения случайной величины на нормальность.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$x^n = (x_1, .., x_n),\ x^n \in X$$

$$H_0 :\ X \sim N(\mu, \sigma^2)$$
$$H_1 :\ H_0 \text{ неверна}$$

$$W(x^n) = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Критерий Стьюдента (t-test), двухвыборочный

Проверка равенства средних значений (мат. ожиданий) в двух выборках.

$$x_1^{n_1} = (x_{11}, .., x_{1n_1}),\ x_1^{n_1} \in X_1,\ X_1 \sim N(\mu_1, \sigma_1^2), \sigma_1 \text{неизвестна}$$
$$x_2^{n_2} = (x_{21}, .., x_{2n_1}),\ x_2^{n_2} \in X_2,\ X_2 \sim N(\mu_2, \sigma_2^2), \sigma_2 \text{неизвестна}$$

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 <\neq> \mu_2$$

$$T(x_1^{n_1}, x_2^{n_2}) = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{S_1^2}{n_1} - \frac{S_2^2}{n_2}}}$$
$$T(x_1^{n_1}, x_2^{n_2}) \sim St$$

Нет выбросов
N > 30

Критерий хи-квадрат (критерий согласия Пирсона)

Оценка статистической значимости различий двух или нескольких относительных показателей (частот, долей).

$x^n = (x_1, .., x_n),\ x^n \in X$

H_0 : Эмпирические (наблюдаемые) и теоретические (ожидаемые) частоты согласованы

H_1 : H_0 неверна

$\chi^2(x^n) = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$

O (Observed) - наблюжаемые частоты

E (Expected) - ожидаемые частоты

K - количество оцениваемых частот

$\chi^2(x^n) \sim \chi^2$

Выборки независимые

Доверительные интервалы

Вид интервальной оценки, которая задаёт числовые границы, в которых, с определённой вероятностью, находится истинное значение оцениваемого параметра.

Порядок расчета доверительного интервала (для мат. ожидания):

1. Задать уровень достоверности (confidence level), α=95%=0.95
2. Найдите по таблице Z-оценок или рассчитать коэффициент достоверности (confidence coefficient) - $Z_{\alpha/2}$, для α = 0.95,$Z_{\alpha/2}$ = 1.96
3. Рассчитать доверительный интервал (confidence interval):

$$CI = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

где $\bar{x}$ - выборочное среднее, σ - стандартное отклонение, n - размер выборки

Практическая часть

Дополнительные материалы

1. http://datalearning.ru/study/Courses/mathstat/lections/lection04.pdf
2. https://machinelearningmastery.com/statistical-hypothesis-tests/
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/pdf/ijem-10-486.pdf
4. Сара Бослаф. «Статистика для всех».
5. Hypothesis Testing: A Visual Introduction To Statistical Significance, *Scott Hartshorn.*
6. Using and Interpreting Statistics in the Social, Behavioral, and Health Sciences, *William E. Wagner, Brian Joseph Gillespie.*
7. Statistical Significance, *Little Quick Fix.*
8. Statistical Method from the Viewpoint of Quality Control, *Walter A. Shewhart, W. Edwards Deming.*