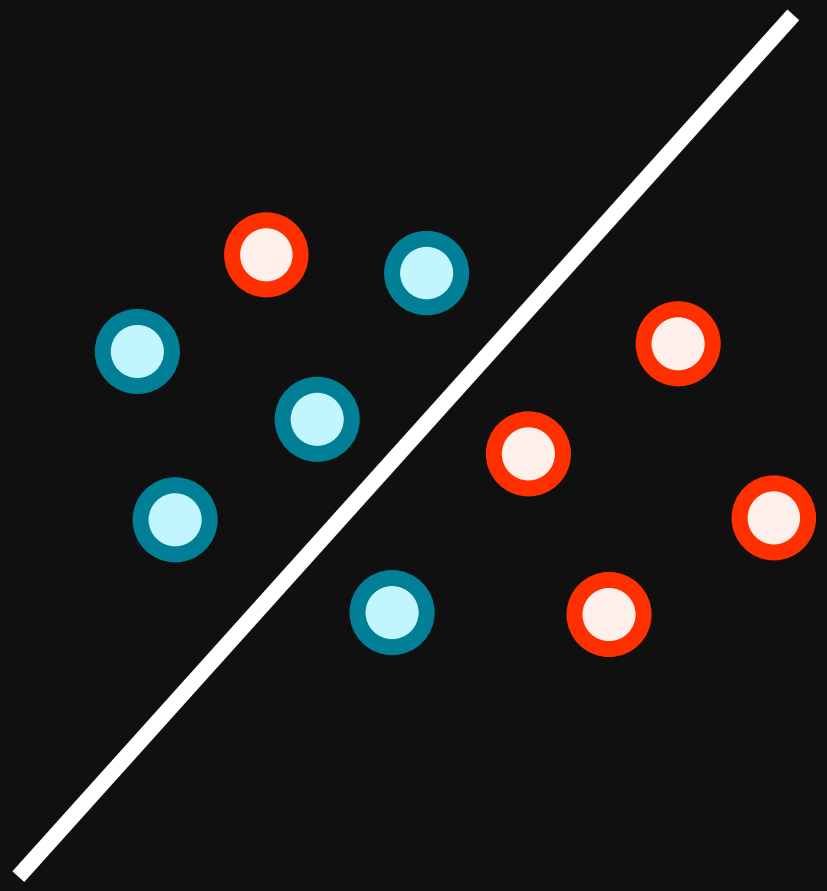
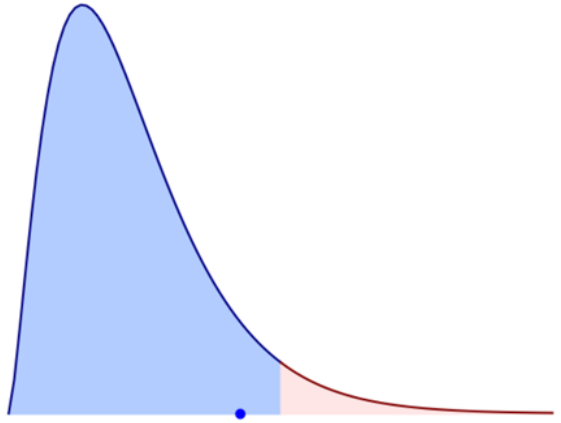
Basics of Machine Learning

Dmitry Ryabokon, github.com/dryabokon



Lesson 06
Statistical ML



Statistical ML

Summary

- P- Value and null hypothesis significance testing
- Chi-squared test
- T-test

- Feature importance
- Feature correlation
- Hashing
- Variance Inflation Factor

Statistical ML

Tutorials

- ex_06a_test_chi2.py
- ex_06a_test_normality.py
- ex_06a_test_student_one_var.py
- ex_06a_test_student_two_vars.py
- ex_06b_feature_correlation.py
- ex_06b_feature_importance.py

Hypothesis testing

Statistical significance

- The result/value has **statistical significance** if it is unlikely to occur under hypothesis H

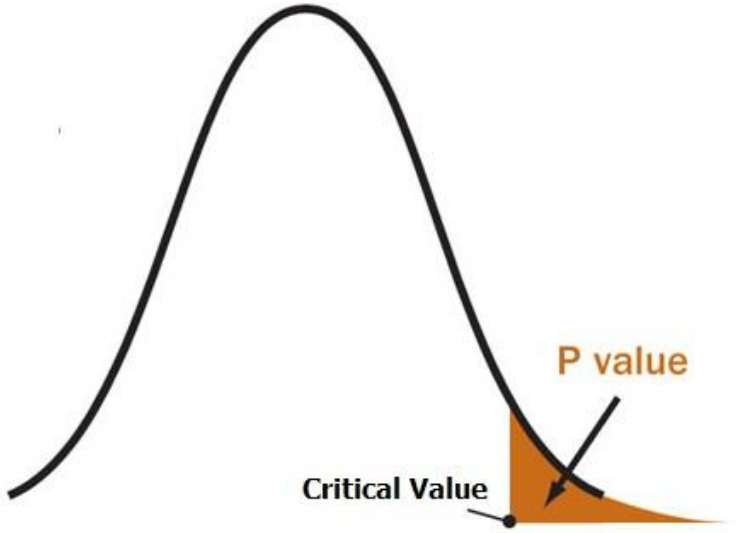
- **significance level α** is probability of false rejecting the hypothesis H (while it was true)

Hypothesis testing

P-Value: probability value or asymptotic significance

how likely your data/experiment is occurred under hypothesis H

A p-value is an area in the tail of a distribution
that tells you the probability of a result happening by chance



Hypothesis testing

P-Value: probability value or asymptotic significance

how likely your data/experiment is occurred under hypothesis H

The p-value gives us the probability of observing what we observed,
given a hypothesis is true

Small p-value has significance, it tells that hypothesis does not explain the experiment

If (p_value <= α)
    reject H
    data is statistically significant.
else:
    accept H

Hypothesis testing

P-Value: probability value or asymptotic significance

how likely your data/experiment is occurred under hypothesis H

Example: p-value == 0.05 => the experiment has 5% or lower chance of occurring under hypothesis H

p-value	significance	How likely experiment data is occurred under hypothesis H
[0 .. 0.01]	High significance	Highly unlikely to occur
[0.01 .. 0.05]	Strong significance	
[0.05 .. 0.10]	Marginal significance	
[0.10 ..	Low significance	Very likely to occur

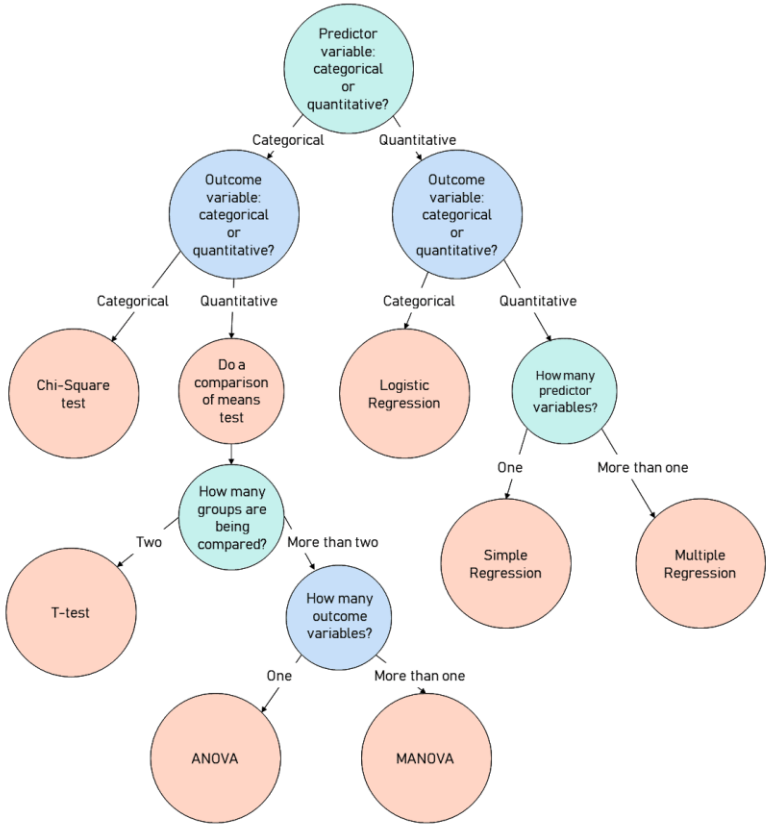
Hypothesis testing

P-Value: probability value or asymptotic significance

p-value depends on the statistical test you are using to test your hypothesis

There are four main test statistics you can use in a hypothesis test.

- Z-Test
- T-Test
- ANOVA
- Chi-Square Test



Chi2 test

Hypothesis testing

Chi-Square Test

Pearson's Chi-Square test for independence between categorical variables
Goal: conclude if two variables are related to each other

$$X^2 = \sum [(\text{observed-expected})^2 / \text{expected}]$$

Hypothesis testing

Chi-Square Test

Если гипотеза о согласии наблюдаемых и ожидаемых частот верна, то, при большом количестве наблюдений, выражение

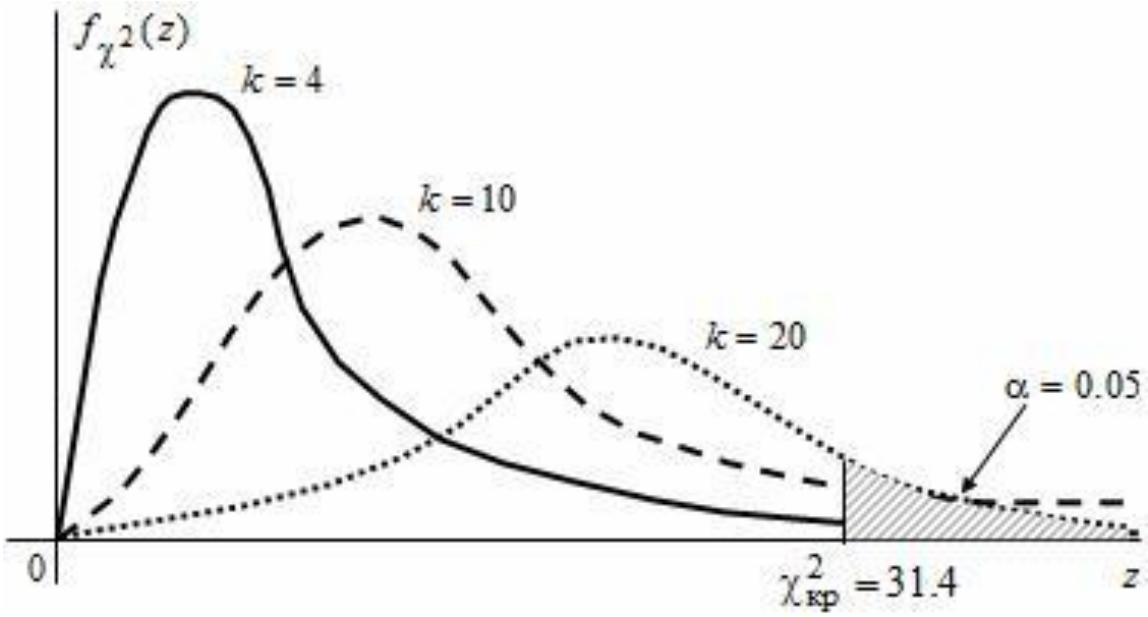
(observed-expected) / sqrt(expected)

имеет стандартное нормальное распределение

нормальность будет проявляться только при достаточно больших частотах. В статистике принято считать, что общее количество наблюдений (сумма частот) должна быть не менее 50 и ожидаемая частота в каждой группе должна быть не менее 5. Только в этом случае величина, показанная выше, имеет стандартное нормальное распределение.

Hypothesis testing

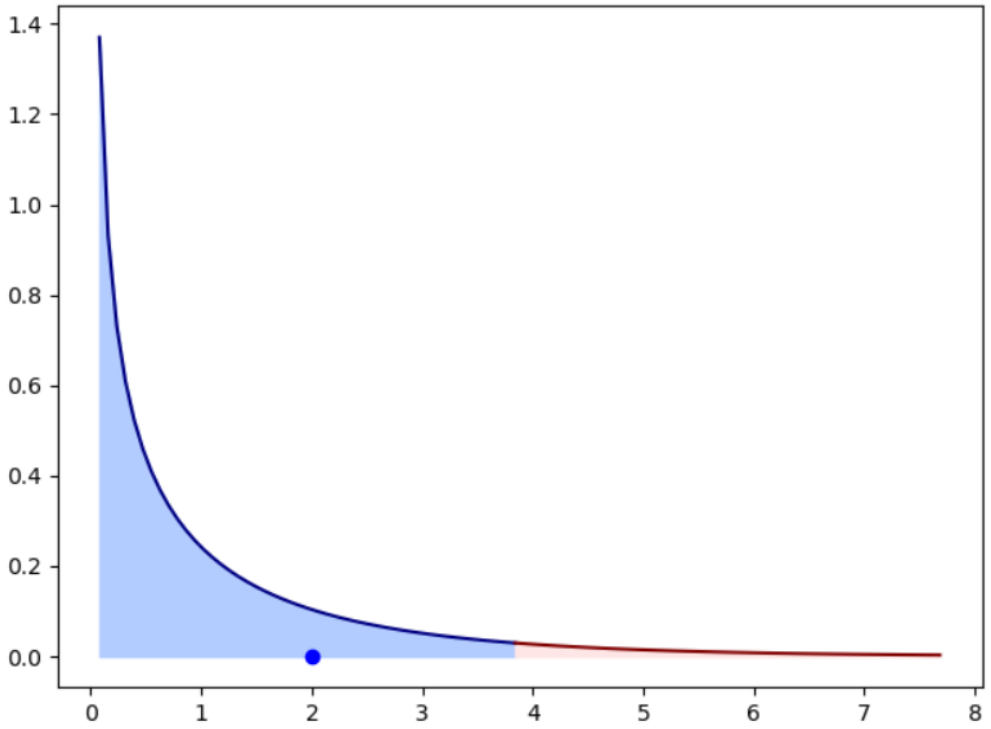
Chi-Square Test



Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Hypothesis testing

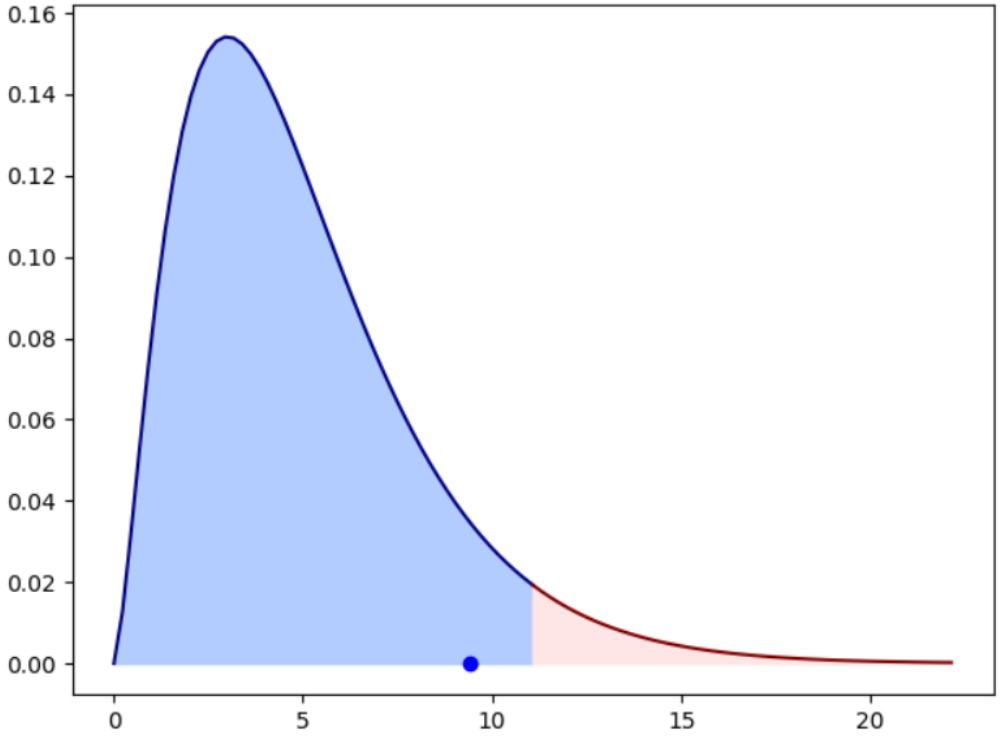
Chi-Square Test



```
X_obs = [110, 90]
X_exp = [100, 100]
```

Hypothesis testing

Chi-Square Test



```
X_obs = [8, 12, 13, 7, 12, 18]
X_exp = [10, 10, 10, 10, 10,
10]
```

T-test

Hypothesis testing

T-Test

Статистический метод, который позволяет сравнивать средние значения двух выборок и на основе результатов теста делать заключение о том, различаются ли они друг от друга статистически или нет.

Для проведения теста, необходимо, чтобы данные выборок имели распределение близкое к нормальному.

Для этого существуют методы оценки, которые позволяют сказать, допустимо ли в данном случае полагать, что данные распределены нормально или нет. Например ***график квантилей (qqplot) или тест Шапиро-Уилка***

Hypothesis testing

T-Test: one sample test

сравнение выборочного среднего с заданным значением

$$t = \frac{(\overline{x} - \mu)\sqrt{m}}{s}$$

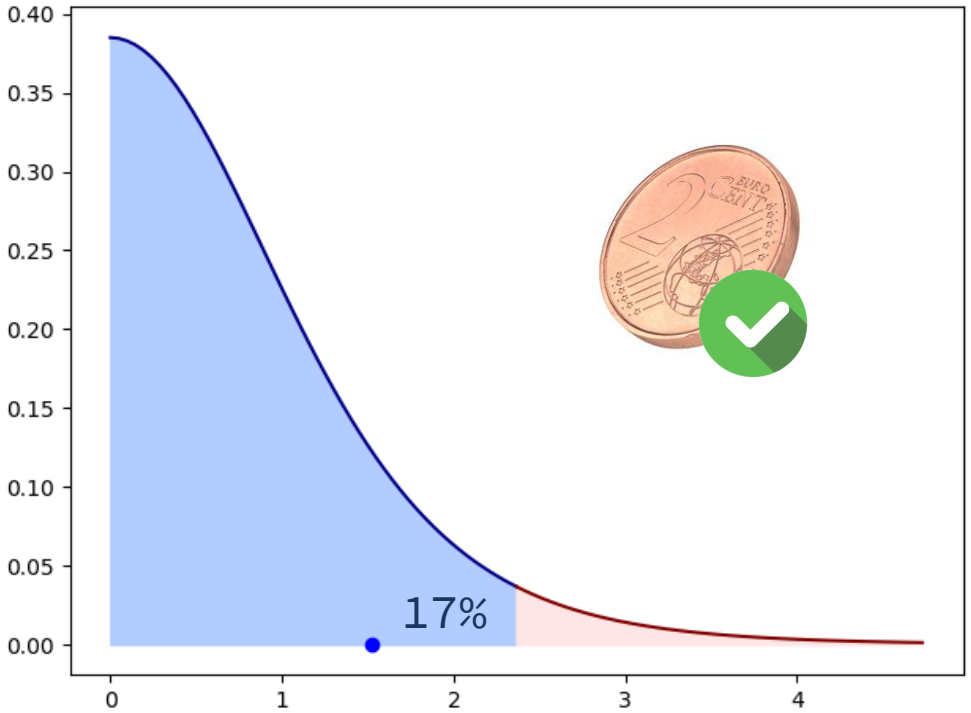
имеет распределение Стьюдента с $m-1$ степенями свободы, где

$\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ — выборочное среднее,

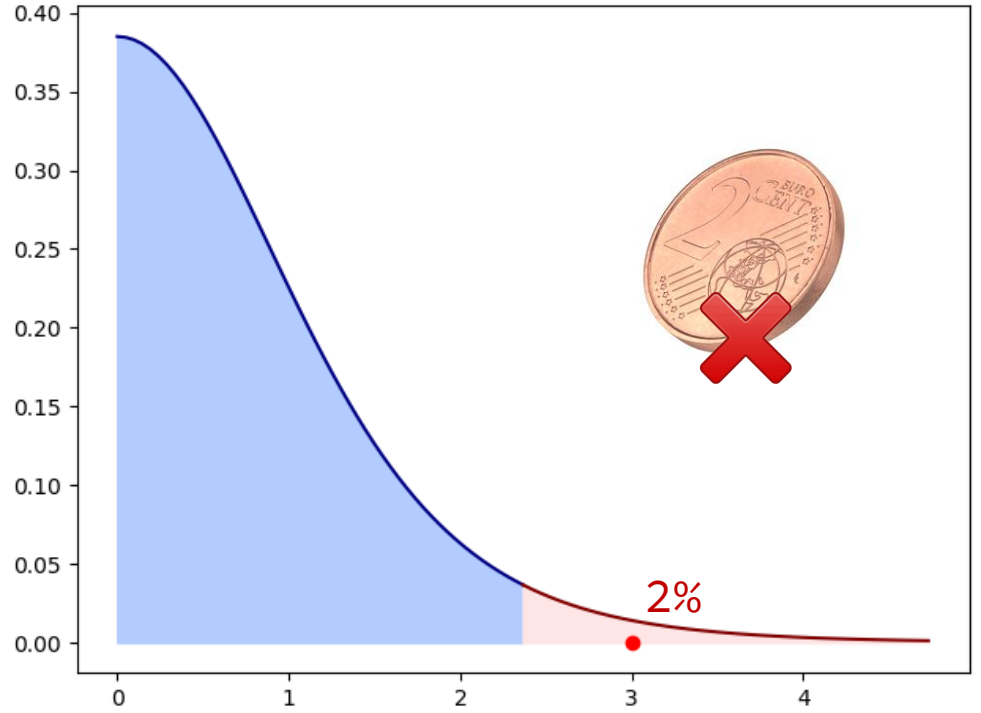
$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$ — выборочная дисперсия.

Hypothesis testing

T-Test: one sample test



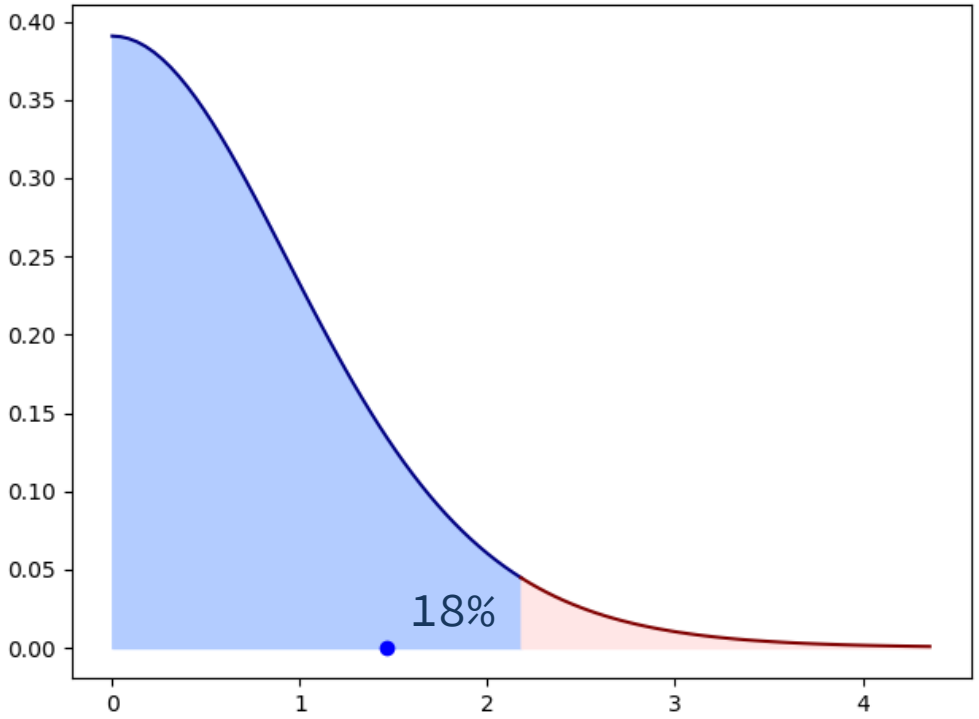

```
X_obs    = [0, 1, 1, 1, 1, 1, 1, 0]
mean_exp = 0.5
```

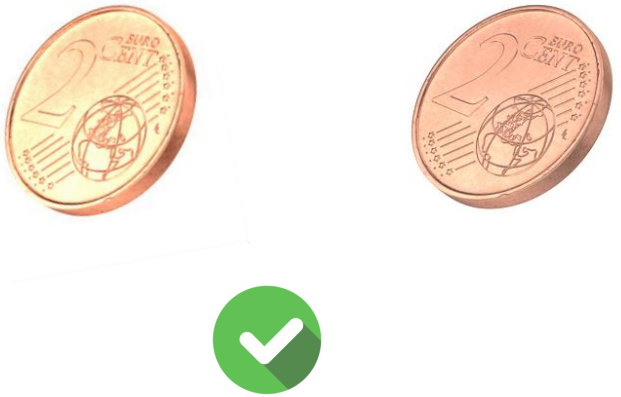



```
X_obs    = [0, 1, 1, 1, 1, 1, 1, 1]
mean_exp = 0.5
```

Hypothesis testing

T-Test: two samples test





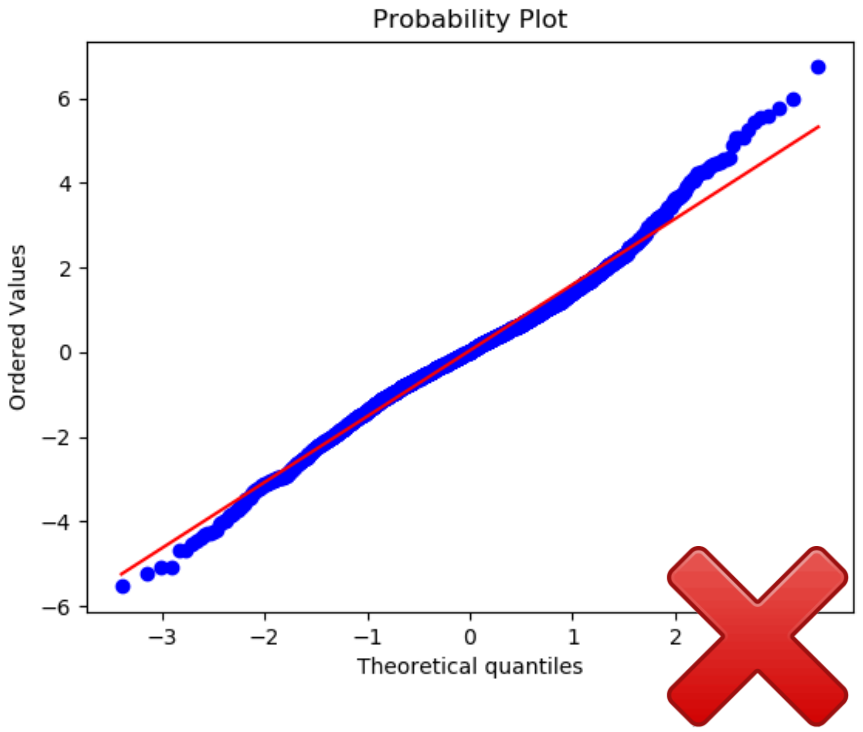
```
X_obs = [0, 1, 1, 1, 1, 1, 1,
1]
X_exp = [0, 0, 0, 1, 1, 1]
```

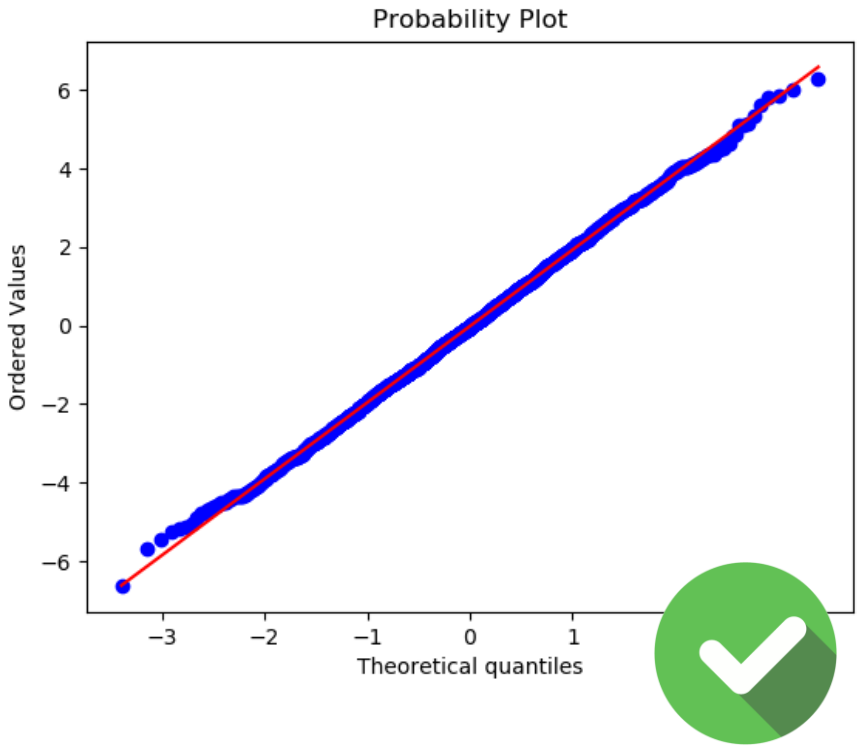
Normality test

- Shapiro-Wilk W Test
- Anderson-Darling Test
- Martinez-Iglewicz Test
- Kolmogorov-Smirnov Test
- D'Agostino Skewness Test

Hypothesis testing

Normality test





Feature Selection

Feature Selection

F-test

F-Test does a hypothesis testing against models X and Y. It tells you if X is significantly better than Y, with respect to a p-value
- **X** is a model created by just a constant
- **Y** is the model created by a constant and a feature

The least square errors in both models are compared and checks if the difference in errors between model **X** and **Y** are significant or introduced by chance

The goal of the F-test is to provide significance level. If you want to make sure the features your are including are significant with respect to your pp-value, you use an F-test.

If you just want to include the kk best features, you can use the correlation only

$$\rho_i = \frac{(X[:,i] - mean(X[:,i])) * (y - mean(y))}{std(X[:,i]) * std(y)}.$$

$$F_i = \frac{\rho_i^2}{1-\rho_i^2} * (n-2)$$

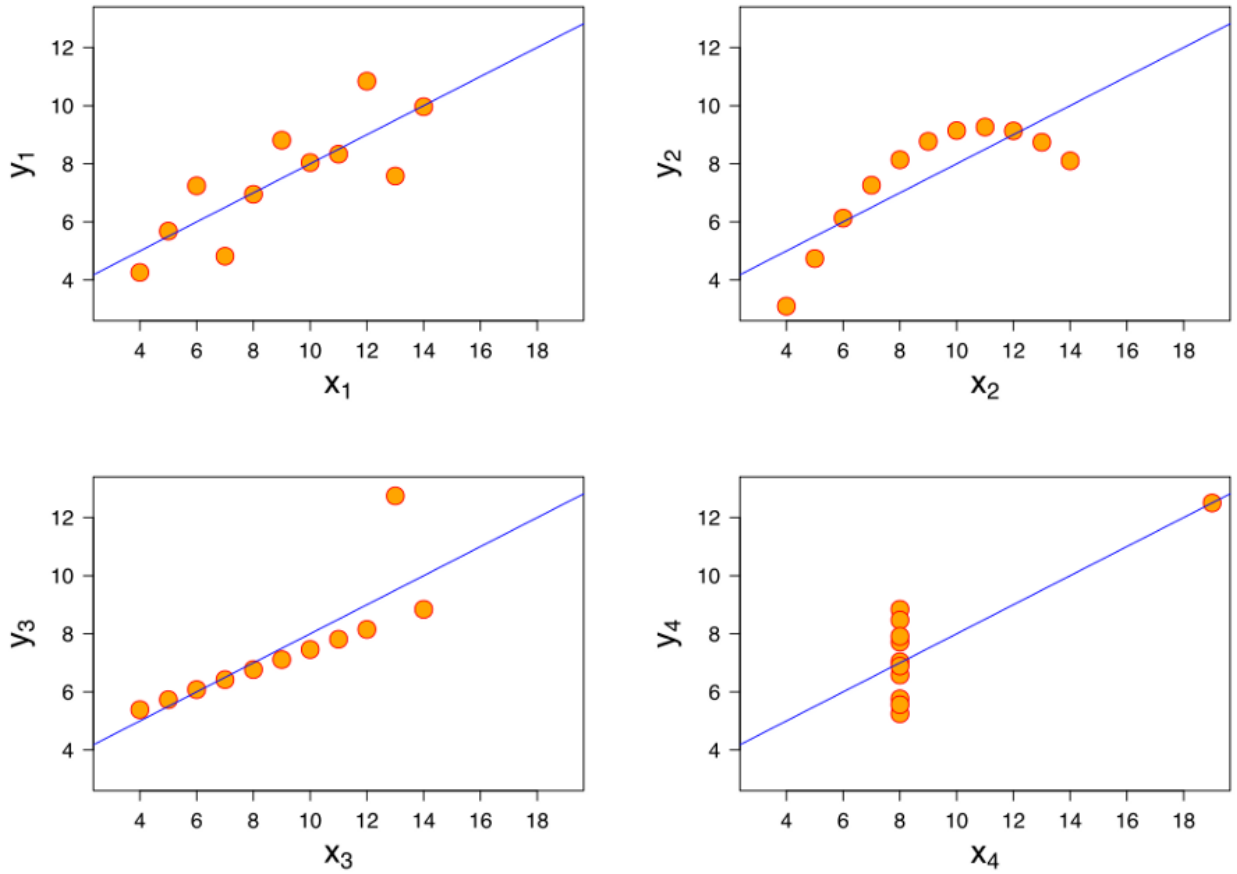
https://stats.stackexchange.com/questions/204141/difference-between-selecting-features-based-on-f-regression-and-based-on-r2

Feature Selection

F-test: drawbacks

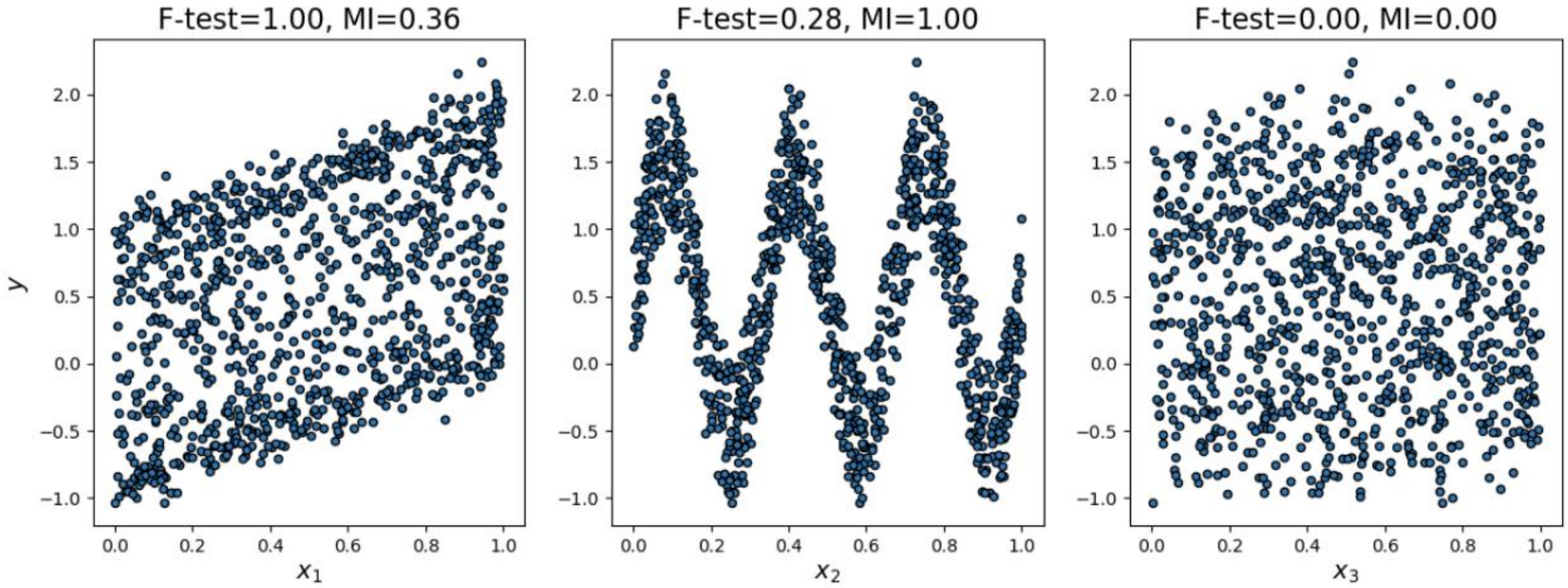
F-Test checks for and only captures linear relationships between features and labels.

A highly correlated feature is given higher score and less correlated features are given lower score.



Feature Selection

Mutual information



Variance Inflation Factor

Variance Inflation Factor

Multicollinearity

Multicollinearity is problem that you can run into when you're fitting a regression model, or other linear model. It refers to predictors that are correlated with other predictors in the model. Unfortunately, the effects of multicollinearity can feel murky and intangible, which makes it unclear whether it's important to fix.

Moderate multicollinearity may not be problematic. However, severe multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity saps the statistical power of the analysis, can cause the coefficients to switch signs, and makes it more difficult to specify the correct model

Upon correlation matrix - **remove high correlated variables** (eg above 75%).

In addition, calculate **VIF (variance inflation factor) to check the presence of multicollinearity**. VIF value <= 4 suggests no multicollinearity whereas
a value of >= 10 implies serious multicollinearity.