is the data normal?

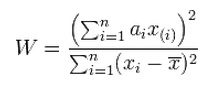
in t-tests we assumed normality

how do we test if data is normal or not

a crude test would be to plot a histogram

stat-test : shapiro- wilk test

Shapiro-Wilk test



→ do that in python

W- test stats

P- p-value

=scipy.stats.shapiro(data)

Interpret the same way an in t-test

Given null hypothesis that the data is drawn from a normal distribution, what is the likelihood that we would observe a value of w that was at least least extreme as the one that we see? The test rejects the null hypothesis of normality when P is less than or equal to 0.05.

Failing the normality test allows use to state with 95% confidence that the data does not fit normal distribution.

- non parametric test

We can run a non parametric test which is a statistical test that does not assume our data is drawn from any particular probability distribution.

Mann-whitney U test

It returns the u value and p value under the assumption of no normal distribution

= scipy.stats.mannwhitney(data1, data2)

Report the results with mean and median of the

Machine learning

Branch of AI focused on constructing

Systems that learn from large amounts of data to make predictions

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Read Leo Breiman pape: Statistical MOdeling: Two Cultures

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ML practical and hands on questions to be answered

Stich an buildings recommendation system or making “best” predictions or classifiers

Tpyes of ML:

-Supervised learning

-unsupervised learning

**Supervised Learning**:

There are labeled inputs that train our model on data (teaching the model what the correct answer would look like: needs known outputs)

We have to explain with examples of known inputs and outputs

Predict output for future inputs (only input data)

-classification

-regression

**Unsupervised learning**

-Don’t have training examples instead we have bunches of unlabeled data

-try to understand the structure of data by clustering similar data together