Course Code 009959 (Spring 2019)

Probability and Statistics with Programming

Test of Hypothesis Based On Two Sample

(Python Views)

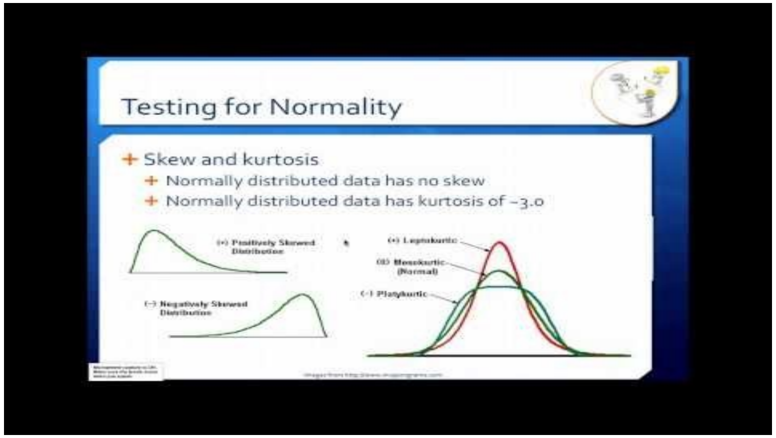
Prof. S. M. Riazul Islam, Dept. of Computer Engineering, Sejong University, Korea

E-mail: riaz@sejong.ac.kr

Several methods to test for normality

- Histogram
- Skew and kurtosis
- Probability plots
- Chi-square goodness of fit

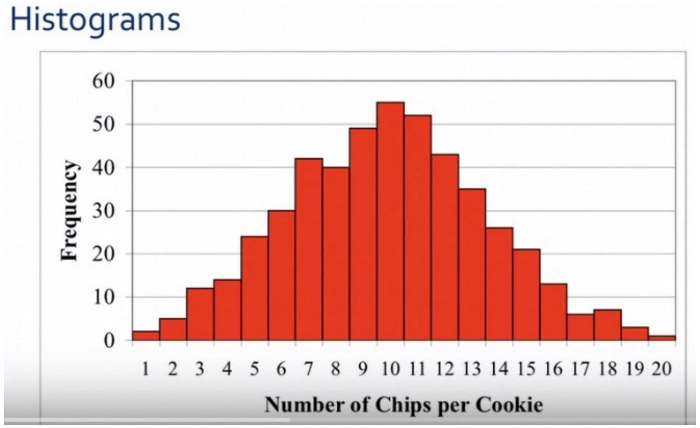
https://youtu.be/72WaWC7Lgjo



Testing for Normality

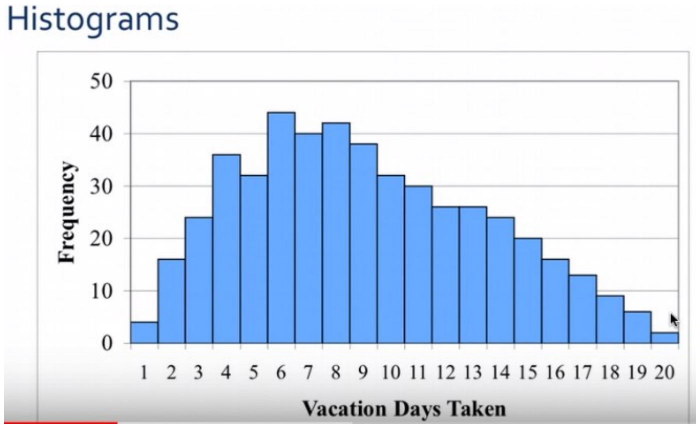
- Skew and kurtosis
  - Normally distributed data has no skew
  - Normally distributed data has kurtosis of –3.0

(+) Positively Skewed Distribution
(–) Negatively Skewed Distribution
(+) Leptokurtic
(0) Mesokurtic (Normal)
(–) Platykurtic

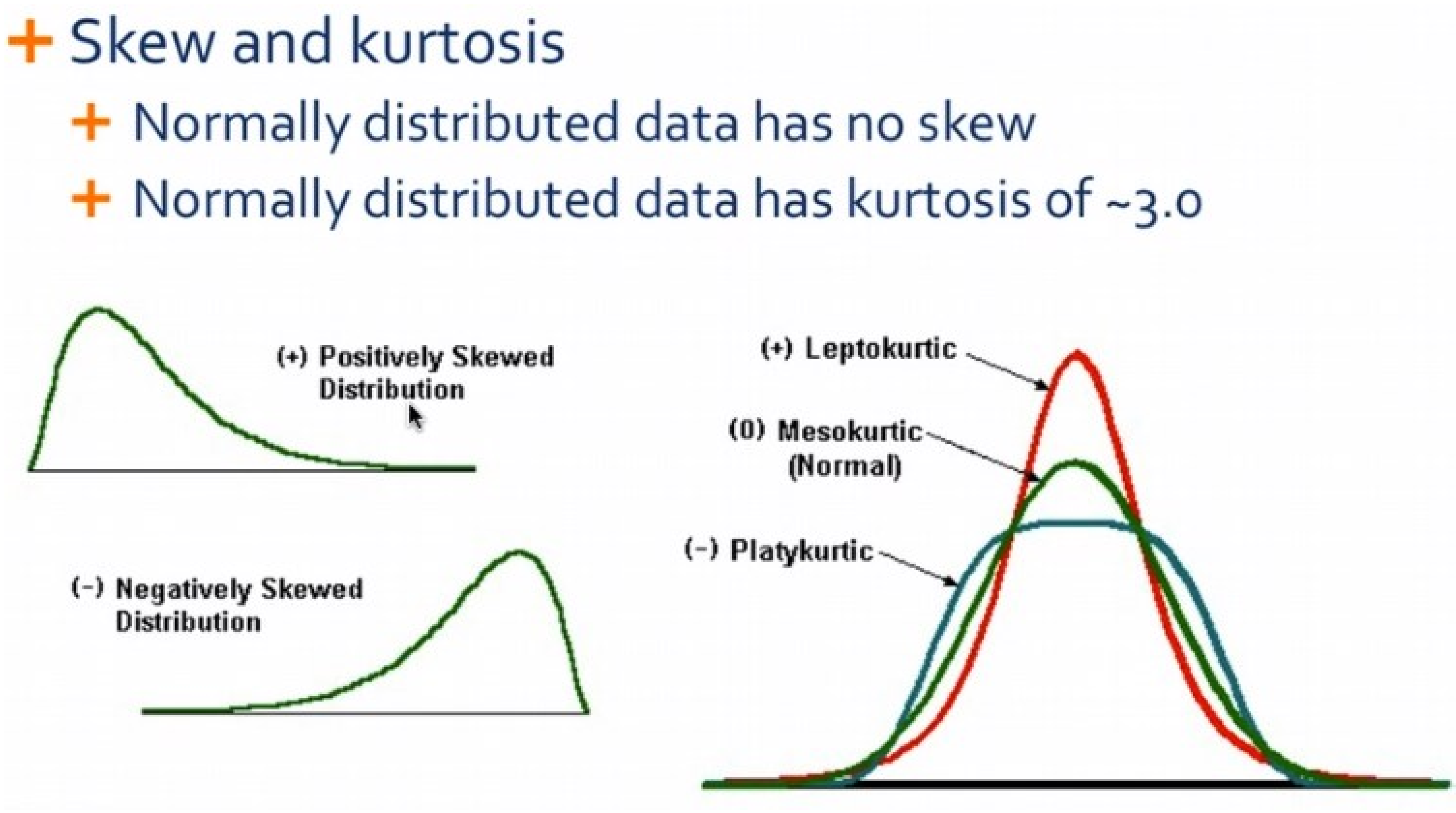


Histograms

data.plot(kind='hist')



Histograms

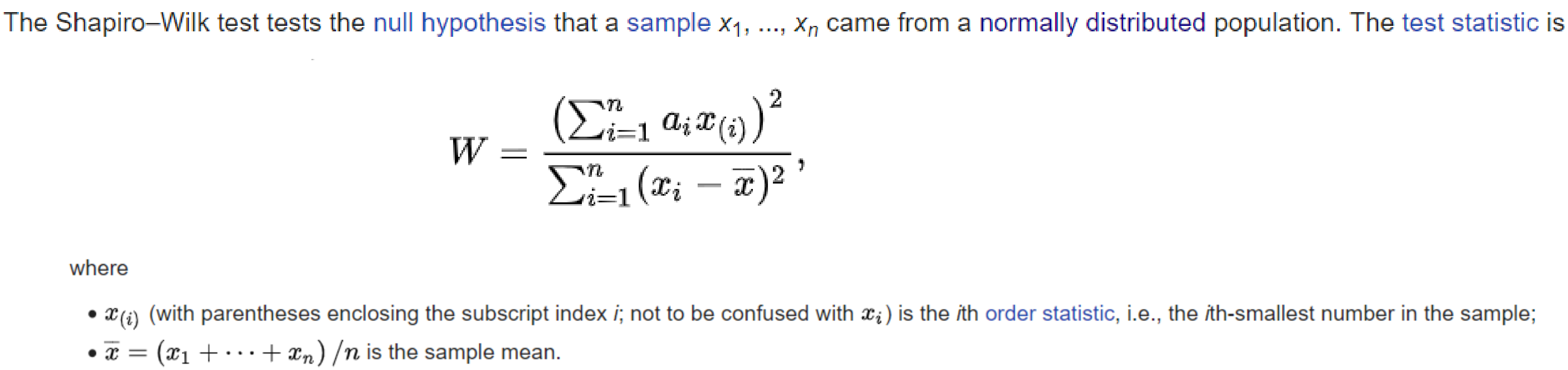


- Skew and kurtosis
  - Normally distributed data has no skew
  - Normally distributed data has kurtosis of ~3.0

(+) Positively Skewed Distribution
(−) Negatively Skewed Distribution
(+) Leptokurtic
(0) Mesokurtic (Normal)
(−) Platykurtic

The *excess kurtosis* is defined as kurtosis minus 3.

Shapiro–Wilk test

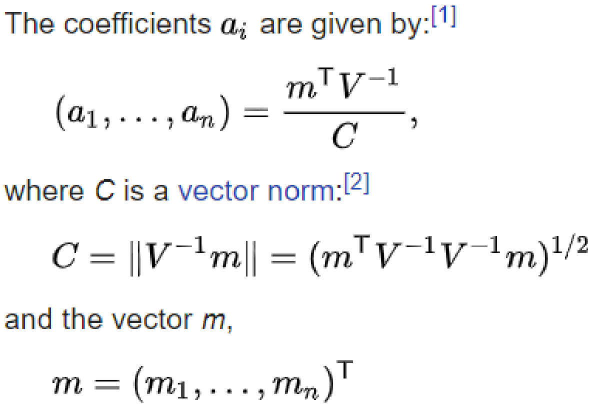


The Shapiro–Wilk test tests the null hypothesis that a sample $x_1, \ldots, x_n$ came from a normally distributed population. The test statistic is

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

where

- $x_{(i)}$ (with parentheses enclosing the subscript index *i*; not to be confused with x_i) is the *i*th order statistic, i.e., the *i*th-smallest number in the sample;
- $\overline{x} = \left(x_1 + \cdots + x_n\right)/n$ is the sample mean.



The coefficients a_i are given by:[1]

$$(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{C},$$

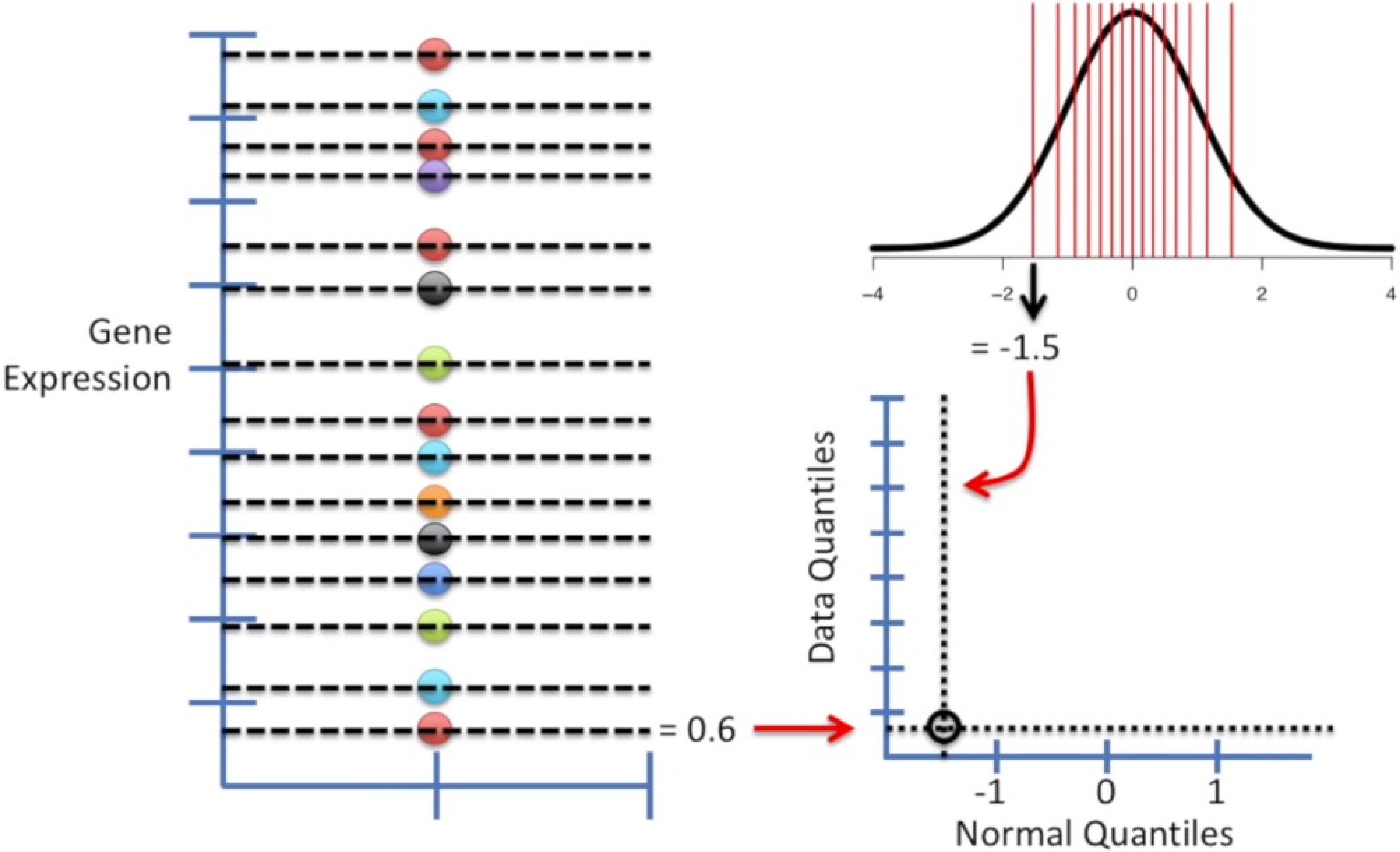
where *C* is a vector norm:[2]

$$C = \|V^{-1} m\| = (m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}$$

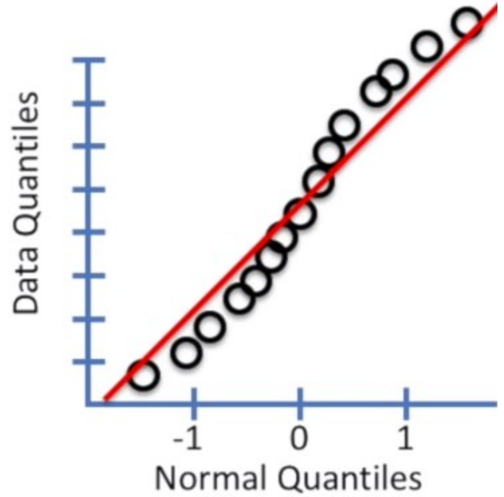
and the vector *m*,

$$m = (m_1, \ldots, m_n)^{\mathsf{T}}$$

stats.shapiro(data)

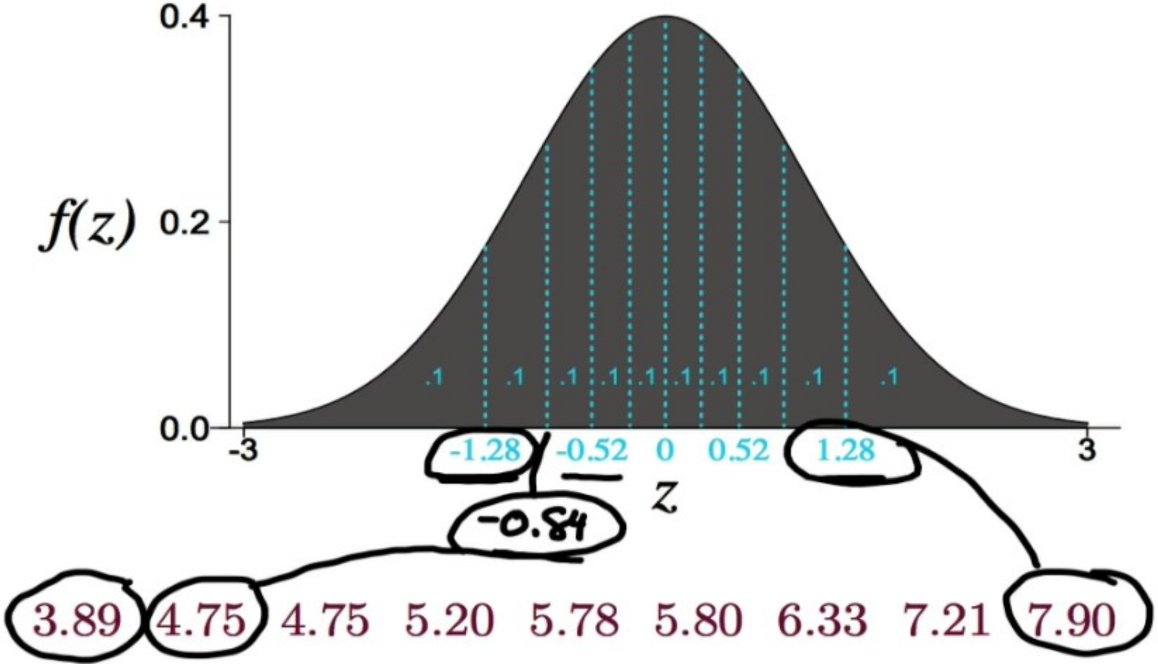


Quantile-Quantile Plots (QQ plots)



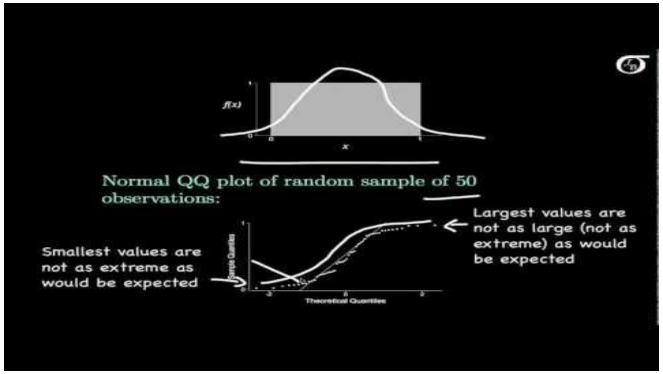
https://youtu.be/okjYjClSjOg

Quantile-Quantile Plots (QQ plots)



```
stats.probplot(data,dist=st.norm,sparams=(0,1),plot=plt);
```

https://youtu.be/X9_ISJ0YpGw


Normal QQ plot of random sample of 50 observations:

Smallest values are not as extreme as would be expected

Largest values are not as large (not as extreme) as would be expected

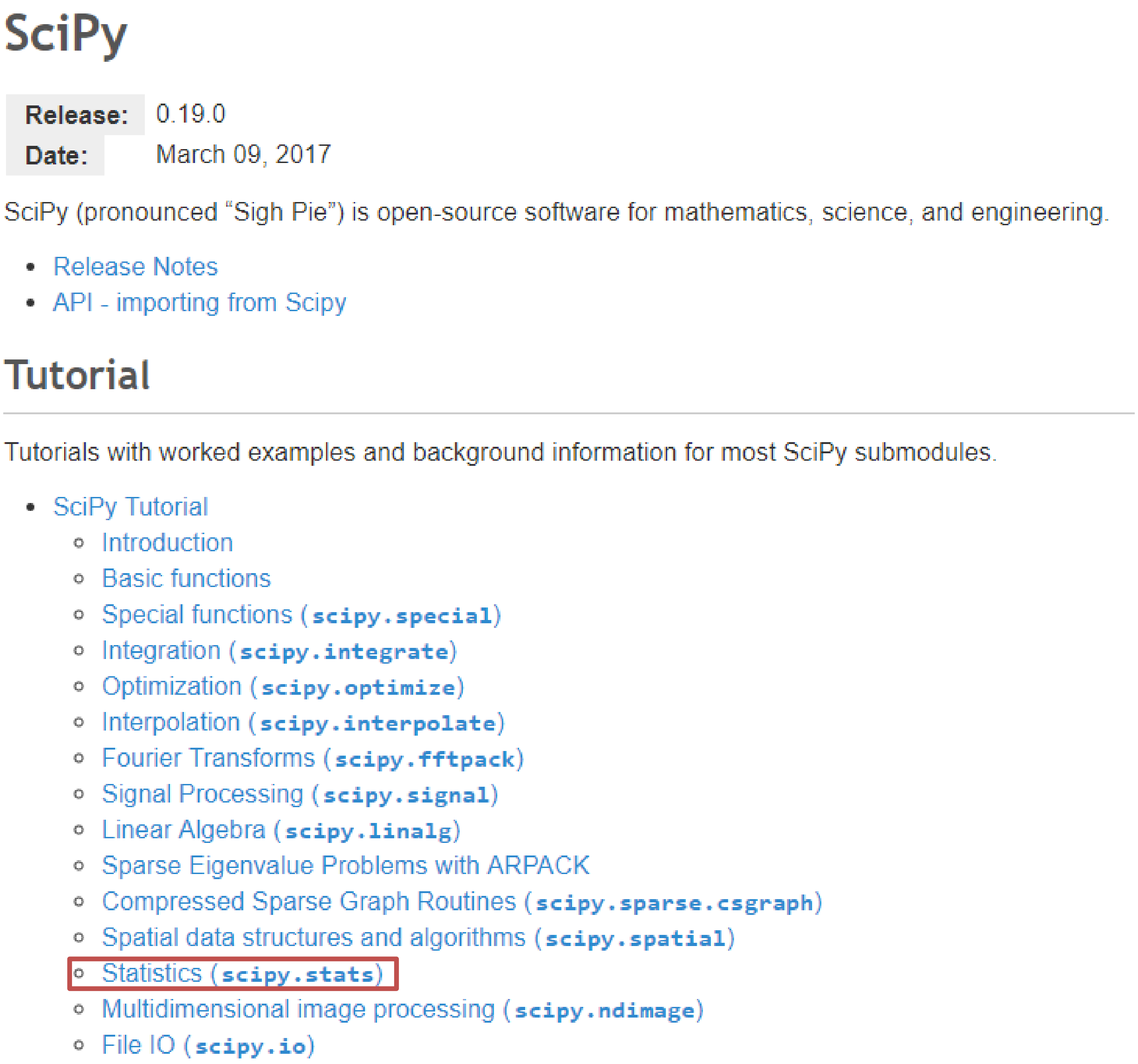
The Unpaired t Test and The Paired t Test requires data to be normally distributed

So, we will perform normality test of the data before we perform t Test

```
stats.ttest_ind(data1, data2)
```

```
stats.ttest_rel(data1, data2)
```

https://pythonfordatascience.org/paired-samples-t-test-python/



SciPy

Release:	0.19.0
Date:	March 09, 2017

SciPy (pronounced “Sigh Pie”) is open-source software for mathematics, science, and engineering.

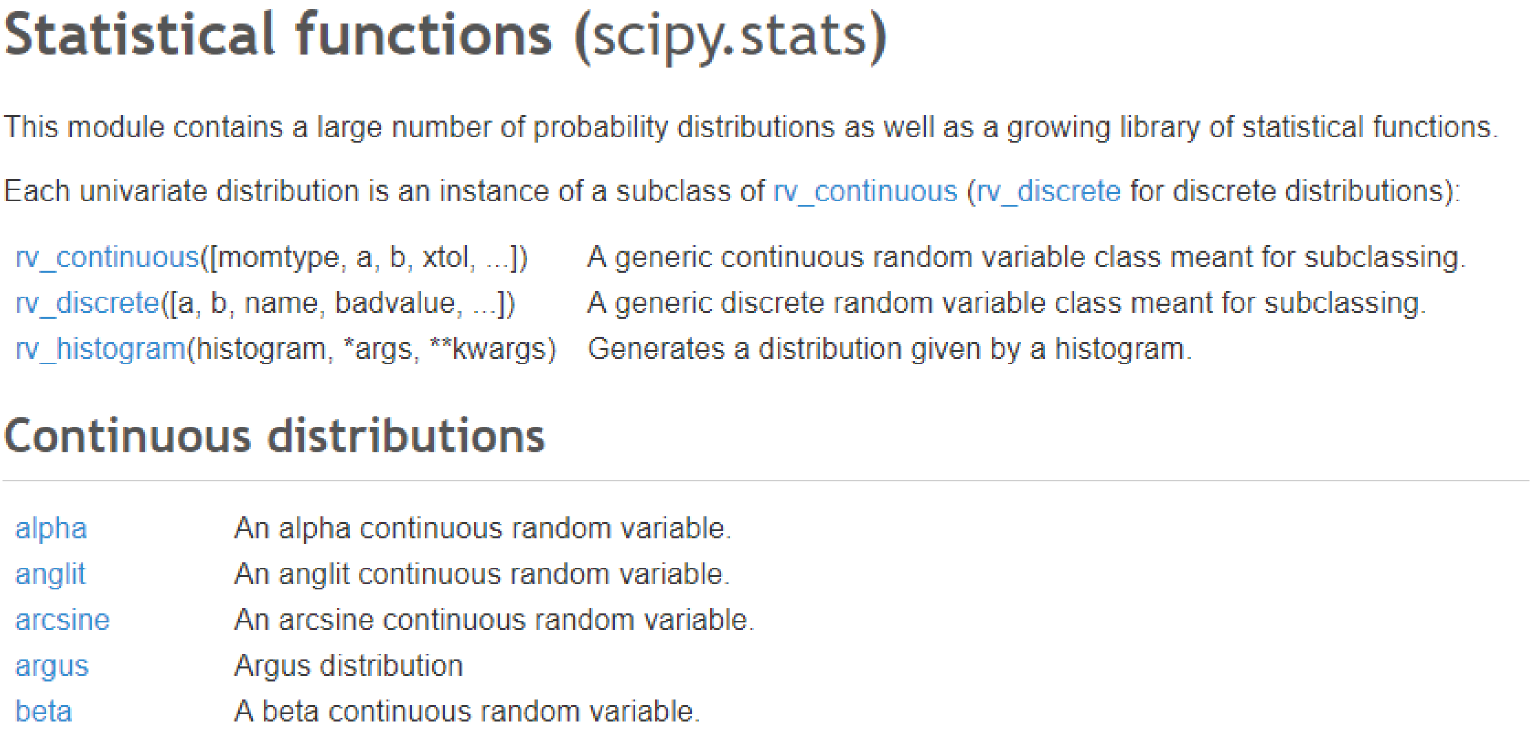
- Release Notes
- API - importing from Scipy

Tutorial

Tutorials with worked examples and background information for most SciPy submodules.

- SciPy Tutorial
  - Introduction
  - Basic functions
  - Special functions (`scipy.special`)
  - Integration (`scipy.integrate`)
  - Optimization (`scipy.optimize`)
  - Interpolation (`scipy.interpolate`)
  - Fourier Transforms (`scipy.fftpack`)
  - Signal Processing (`scipy.signal`)
  - Linear Algebra (`scipy.linalg`)
  - Sparse Eigenvalue Problems with ARPACK
  - Compressed Sparse Graph Routines (`scipy.sparse.csgraph`)
  - Spatial data structures and algorithms (`scipy.spatial`)
  - Statistics (`scipy.stats`)
  - Multidimensional image processing (`scipy.ndimage`)
  - File IO (`scipy.io`)

https://docs.scipy.org/doc/scipy-0.19.0/reference/index.html



Statistical functions (scipy.stats)

This module contains a large number of probability distributions as well as a growing library of statistical functions.

Each univariate distribution is an instance of a subclass of rv_continuous (rv_discrete for discrete distributions):

rv_continuous([momtype, a, b, xtol, ...])	A generic continuous random variable class meant for subclassing.
rv_discrete([a, b, name, badvalue, ...])	A generic discrete random variable class meant for subclassing.
rv_histogram(histogram, *args, **kwargs)	Generates a distribution given by a histogram.

Continuous distributions

alpha	An alpha continuous random variable.
anglit	An anglit continuous random variable.
arcsine	An arcsine continuous random variable.
argus	Argus distribution
beta	A beta continuous random variable.

https://docs.scipy.org/doc/scipy-0.19.0/reference/stats.html#module-scipy.stats

Q&A