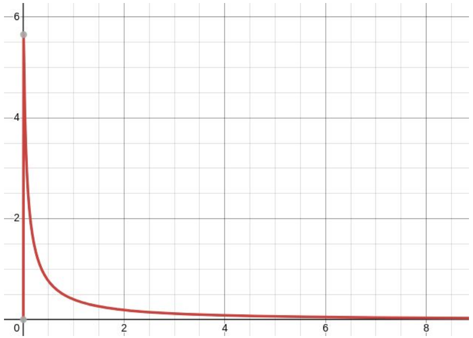
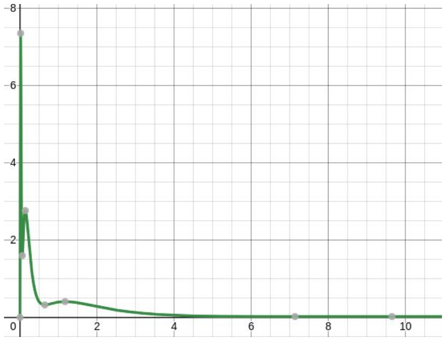
Normality Tests

Machine learning rocks!!! Why not just use it?

- The question is why not just train a machine learning model based on features such as mean, variance, skewness and kurtosis for a binary classification task to test if a distribution is normal
- Here is an example of two continuous random variables with the same moments but completely distinct distributions:

$$\frac{e^{-\frac{(\log x)^2}{2}}}{x \cdot (2\pi)^{0.5}}$$





$$\frac{e^{-\frac{(\log x)^2}{2}}}{x \cdot (2\pi)^{0.5}}\left(1 + \frac{\sin(2\pi\ \log\ x)}{2}\right)$$

- So it’s definitely possible to have two distributions with same moments which look different
- Sometimes we also need to tell how close or far a distribution is from a nearby normal distribution

When should you care about normality tests

- Drug development: In the pharmaceutical industry, normality tests are used to ensure that the distribution of a drug's active ingredient is normal. If the distribution is not normal, it can affect the drug's efficacy or lead to harmful side effects.
- Quality control: Normality tests are often used in quality control to ensure that a product's characteristics, such as weight or dimensions, are normally distributed. Deviations from normality can indicate manufacturing defects or problems in the production process.
- Environmental monitoring: Normality tests are used in environmental monitoring to check whether pollutant levels are normally distributed. Deviations from normality can indicate that there are sources of pollution that need to be investigated.
- And to some extent, your grades in this class will be calculated based on a normality assumption!!!!!!

Normality Tests

There are several methods of assessing whether data are normally distributed or not. They fall into two broad categories: graphical and statistical. The some common techniques are:

Graphical:

- Q-Q probability plots
- Cumulative frequency
- (P-P) plots

Statistical:

- W/S test
- Jarque-Bera test
- Shapiro-Wilks test
- Kolmogorov-Smirnov test
- D’Agostino test

Percentiles and Quantiles

The k-th percentile of a set of values divides them so that k % of the values lie below and (100 − k)% of the values lie above.

- The 25th percentile is known as the lower quartile.

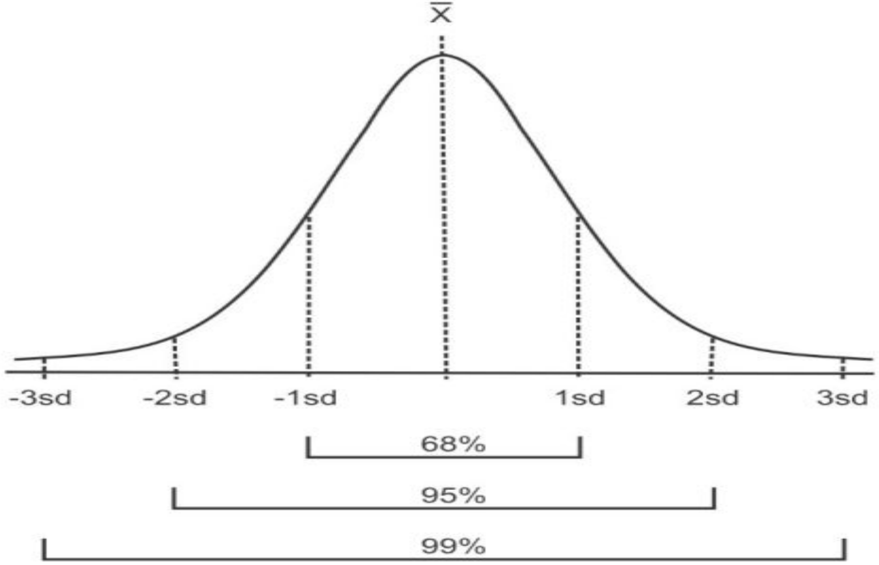
- The 50th percentile is known as the median.

- The 75th percentile is known as the upper quartile.

It is more common in statistics to refer to quantiles. These are the same as percentiles, but are indexed by sample fractions rather than by sample percentages.

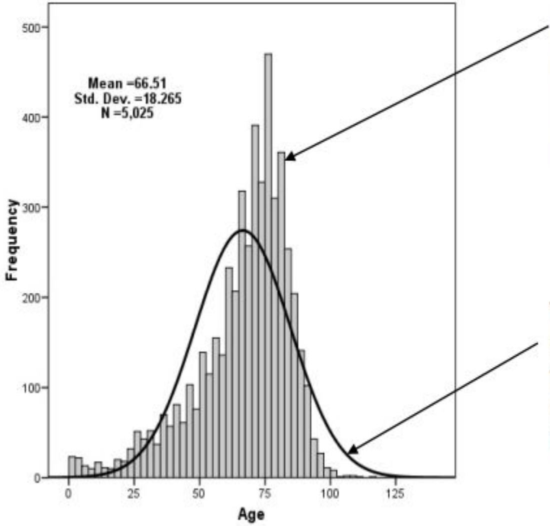
Graphical Normality Tests

- For each mean and standard deviation combination a theoretical normal distribution can be determined. This distribution is based on the proportions shown below.



Graphical Normality Tests

- This theoretical normal distribution can then be compared to the actual distribution of the data.



The ***actual*** data distribution that has a mean of 66.51 and a standard deviation of 18.265.

Theoretical normal distribution ***calculated*** from a mean of 66.51 and a standard deviation of 18.265.

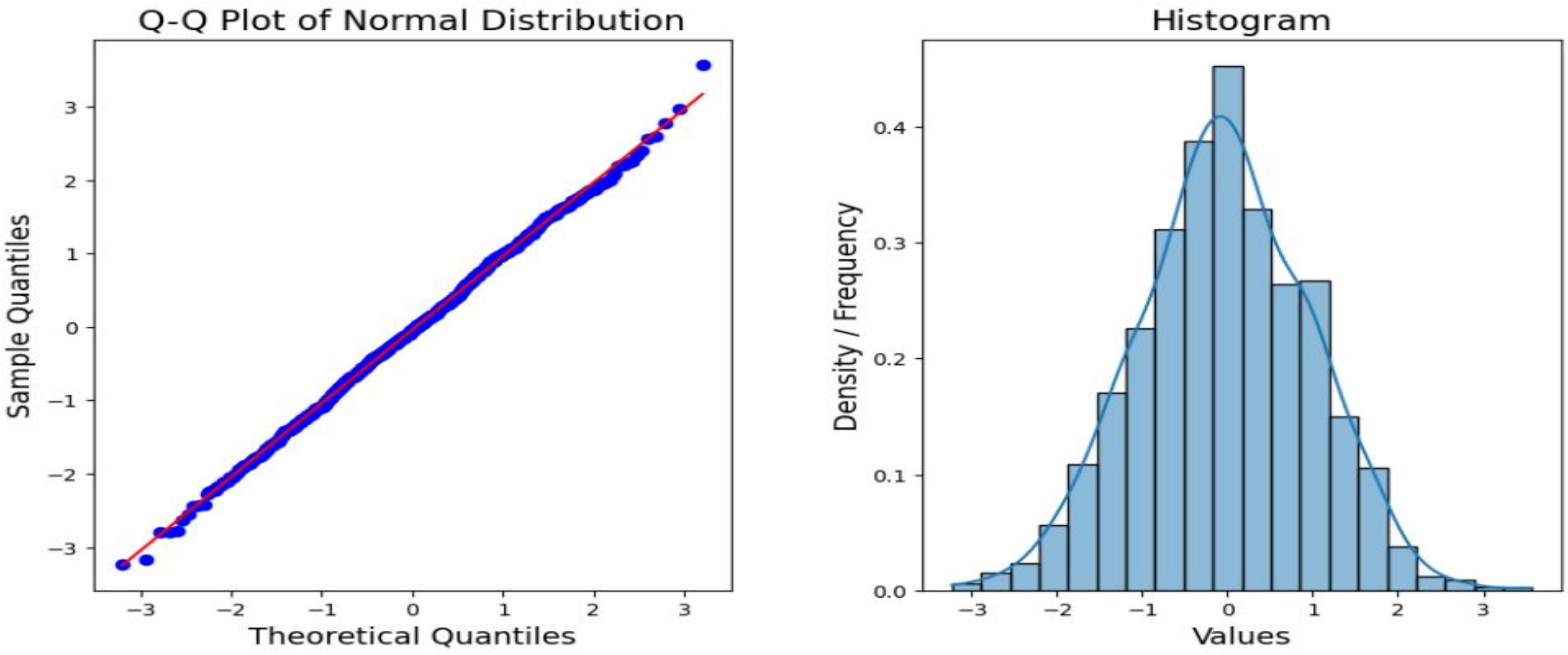
Theoretical Quantile Quantile Plots

- Quantile-quantile plots can be used to compare the distributions of two sets of numbers.
- They can also be used to compare the distributions of one set of values with some theoretical distribution.
- Most commonly, the yardstick distribution is the standard normal distribution:

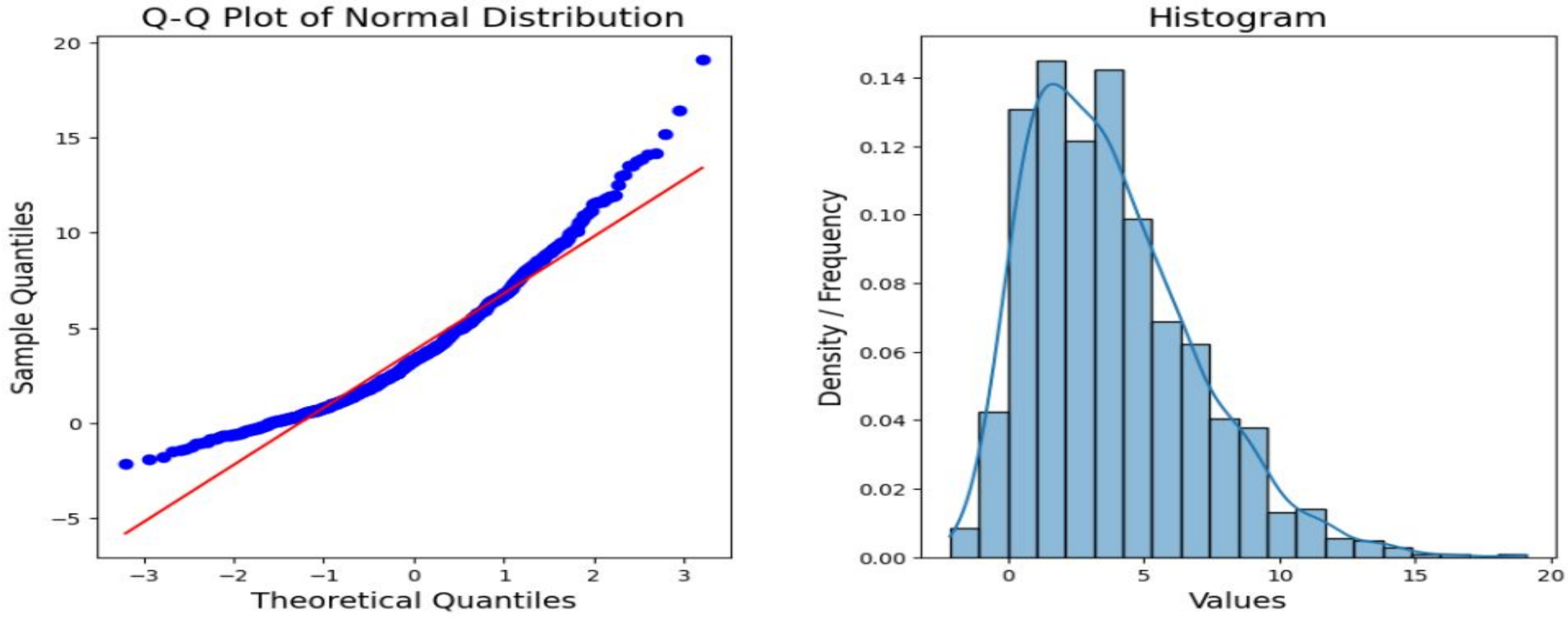
$$P[X \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

- If the values being plotted resemble a sample from a normal distribution, they will lie on a straight line with intercept equal to the mean of the values and slope equal to the standard deviation
- Each point in the data is compared to the yardstick distribution, point by point, quantile by quantile and the theoretical value vs sample value is plotted

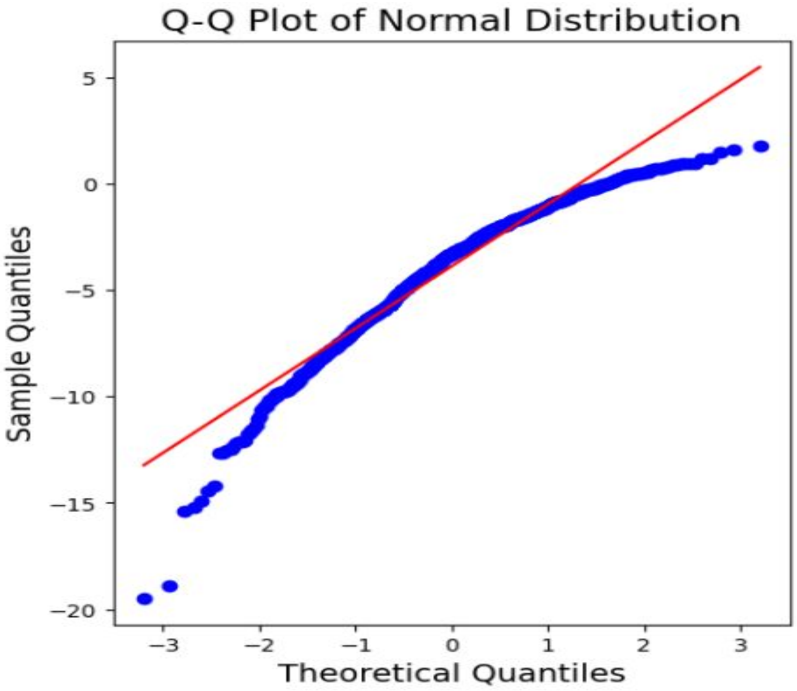
Q-Q plot for a normal distribution

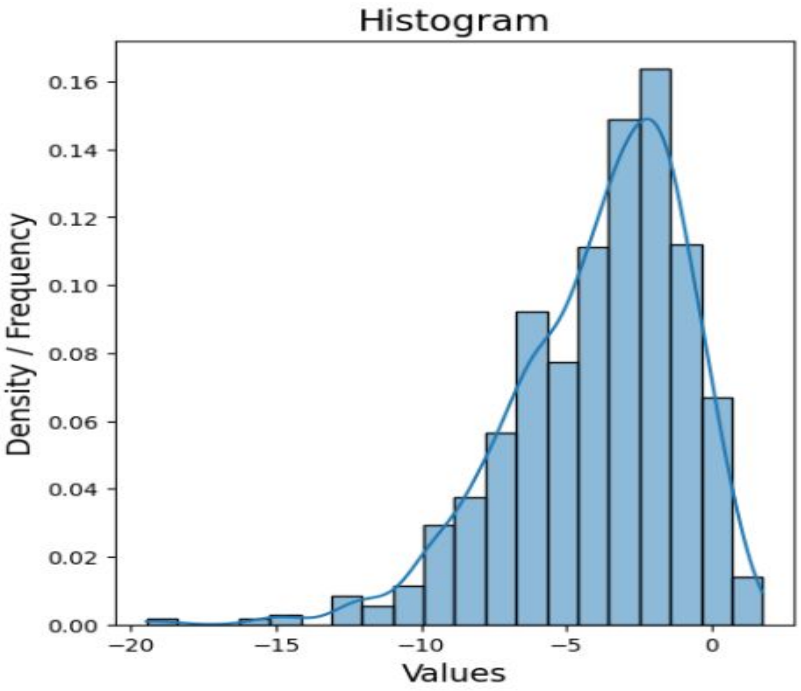


Departure from Normality - Right skewed distribution

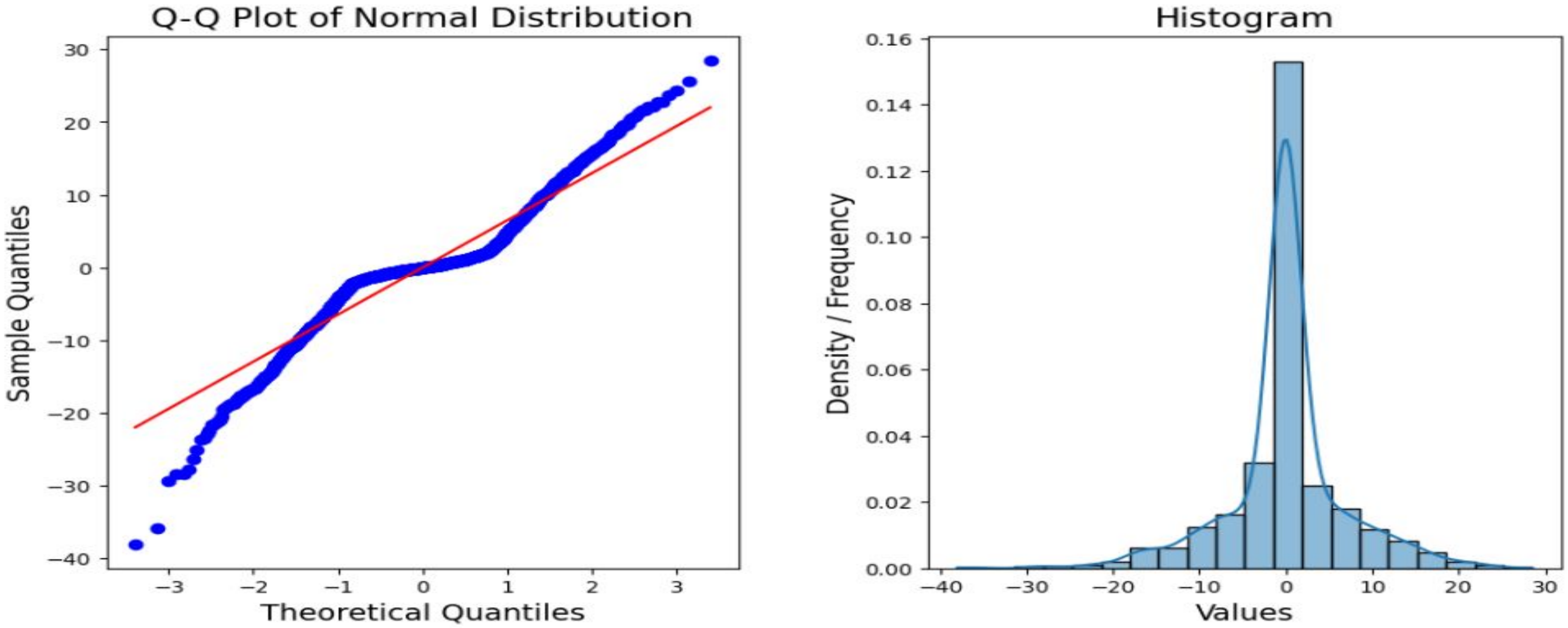


Departure from Normality - Left skewed distribution

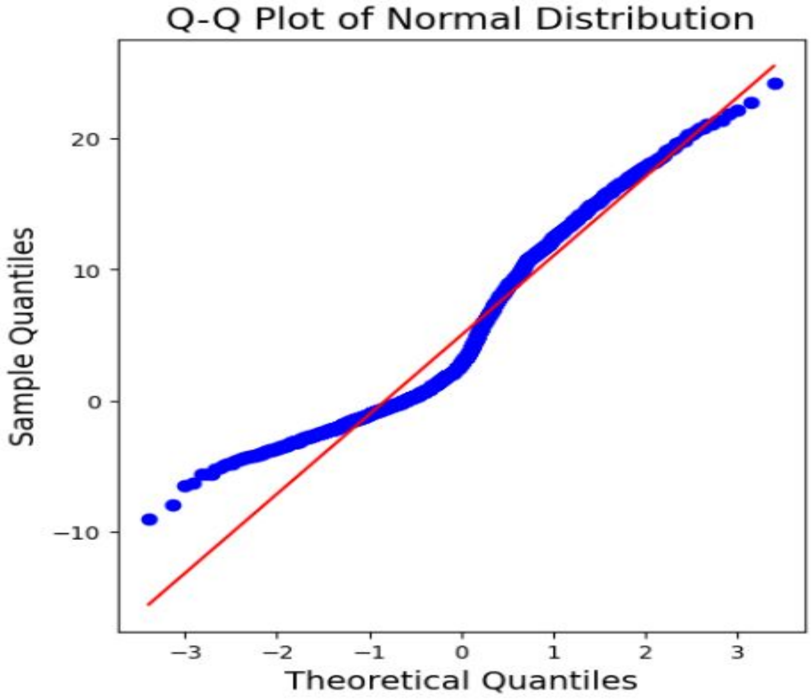


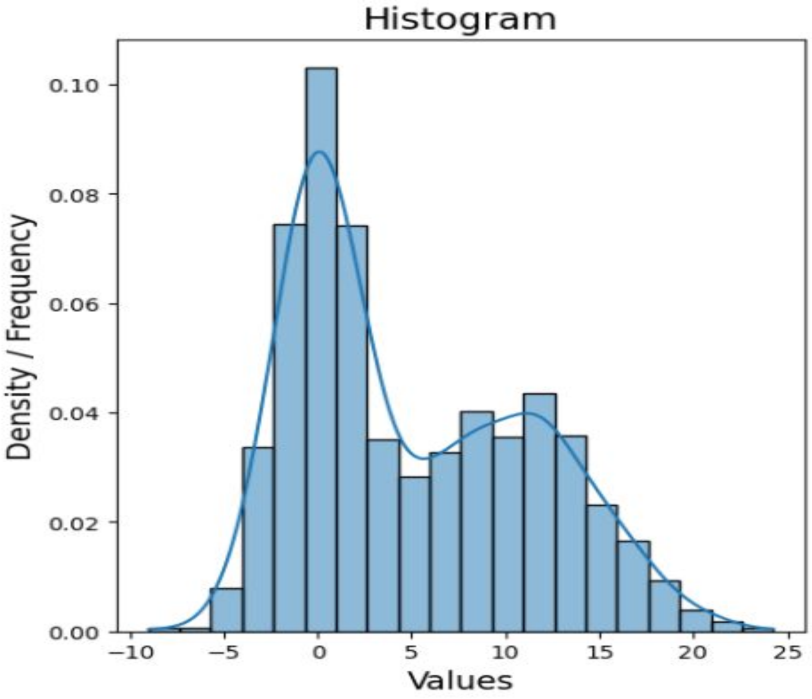


Departure from Normality - Heavy-tailed distribution



Departure from Normality - Bimodal distribution





Statistical Normality Tests

- Graphical methods are typically not very useful when the sample size is small.

- Statistical tests for normality are more precise since actual probabilities are calculated. Tests for normality calculate the probability that the sample was drawn from a normal population.

- The hypotheses used are:
  - H_o: The sample data are not significantly different than a normal population.
  - H_a: The sample data are significantly different than a normal population

- Shapiro-Wilk (W) test is a very powerful omnibus test which has good power with symmetrical, short and long tails and is good with asymmetrical distros.

Order Statistics

- i-th order statistic: i-th smallest element
- Let X_1, X_2,…,X_n be a r.s. of size n from a distribution of continuous type having pdf f(x), a<x<b. Let $X_{(1)}$ be the smallest of X_i, $X_{(2)}$ be the second smallest of X_i,…, and $X_{(n)}$ be the largest of X_i.

$$a < X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} < b$$

- X(i) is the i-th order statistic

$$X_{(1)} = \min\left\{ X_1, X_2, \cdots, X_n \right\}$$

$$X_{(n)} = \max\left\{ X_1, X_2, \cdots, X_n \right\}$$

Order Statistics

- If X_1, X_2,…,X_n be a r.s. of size n from a population with continuous pdf f(x), then the joint pdf of the order statistics $X_{(1)}$, $X_{(2)}$,…,$X_{(n)}$ is:

$$g\left(x_{(1)}, x_{(2)}, \cdots, x_{(n)}\right) = n!\, f\left(x_{(1)}\right) f\left(x_{(2)}\right) \cdots f\left(x_{(n)}\right)$$

for $x_{(1)} \leq \ldots \leq x_{(n)}$

- Order statistics are not independent.
- The joint pdf of ordered sample is not same as the joint pdf of unordered sample

Shapiro-Wilk’s Test

- The Shapiro-Wilk’s test proposed in 1965, calculates a W-statistic that tests whether a r.s. X_1, X_2,…,X_n comes specifically from a normal distribution

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

- $X_{(i)}$ are the ordered sample values with $X_{(1)}$ being the smallest, and the coefficients ai is calculated as follows:

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

Where C is a vector norm : $C = \|V^{-1}m\| = (m^\mathsf{T} V^{-1} V^{-1} m)^{1/2}$

the vector m is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those normal order statistics

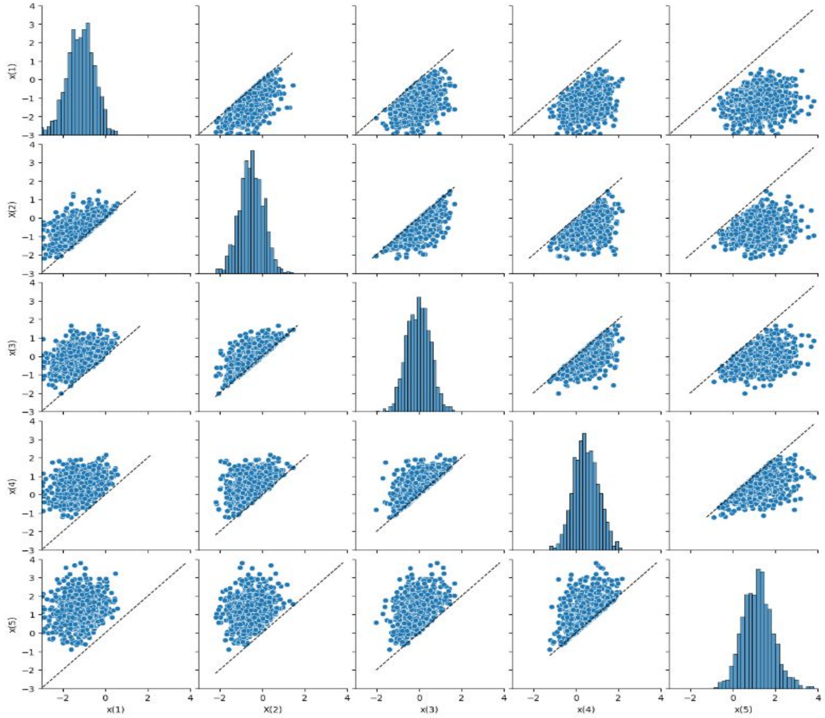
Interpretation of the Shapiro-Wilk’s statistic

- The W statistic itself can be interpreted as a measure of how close the sample distribution is to a normal distribution. A value of W close to 1 indicates that the sample is very close to a normal distribution, while a value of W close to 0 indicates that the sample is very different from a normal distribution

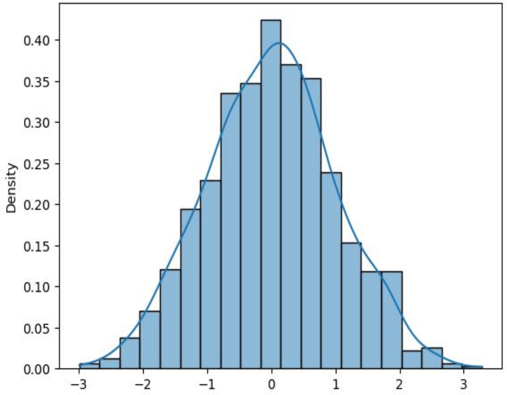
- What W captures is essentially, how far the actual distribution is compared to a theoretical normal distribution with mean and standard deviation set to the mean and standard deviation of the actual distribution

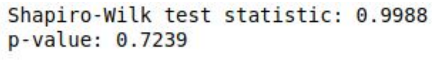
- The coefficients a_i capture the weighted statistical dispersion of theoretical normal distribution

Statistical dispersion of a standard normal distribution

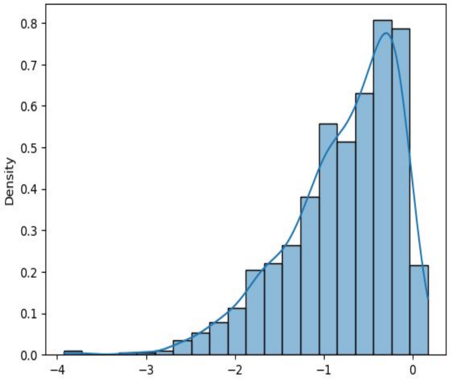


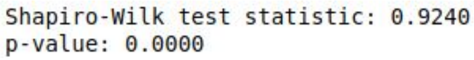
Shapiro Wilk’s test on a standard Normal distribution



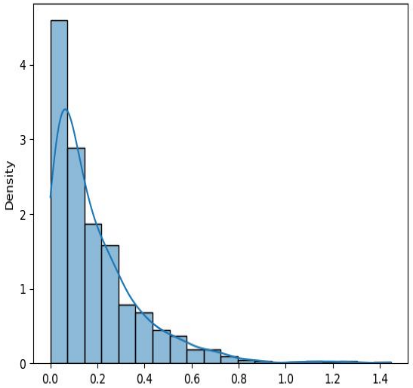


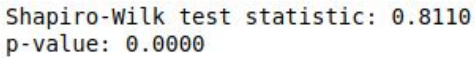
```
Shapiro-Wilk test statistic: 0.9988
p-value: 0.7239
```





```
Shapiro-Wilk test statistic: 0.9240
p-value: 0.0000
```





```
Shapiro-Wilk test statistic: 0.8110
p-value: 0.0000
```

- ○ H_o: The sample data are not significantly different than a normal population.
- ○ H_a: The sample data are significantly different than a normal population