**NORMAL DISTRIBUTIONS**

**INTRODUCTION**

Random Variables

A random variable is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

Discrete random variable are those which may take on only a countable number of distinct values such as 0,1,2,3,4,…….. Examples of discrete random variable are-

* Number of children in the family
* Number of patients in a hospital
* Number of students in this workshop

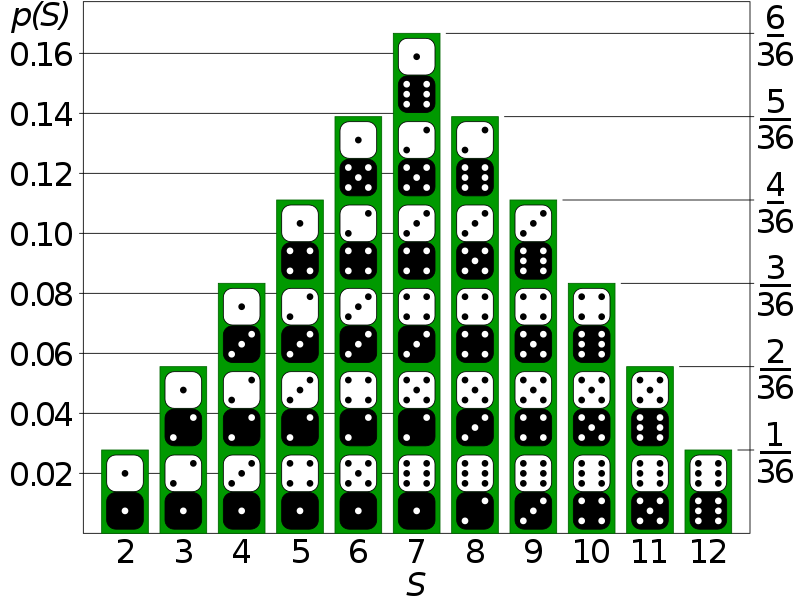
Continuous random variable are those which takes an infinite number of possible values. Examples of continuous random variable are-

* Height/Weight of a person
* Time required to run a 5km race
* Time at which any individual can enter this workshop

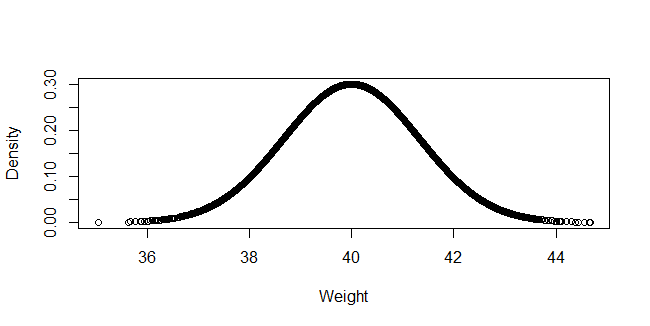
Probability Distribution

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

Example - let us look at the number observed when rolling two standard six-sided dice. Each die has a 1/6 probability of rolling any single number, one through six, but the sum of two dice will form the probability distribution depicted in the image below. Seven is the most common outcome (1+6, 6+1, 5+2, 2+5, 3+4, 4+3). Two and twelve, on the other hand, are far less likely (1+1 and 6+6).



Example – Weight of 14-year-old girls during a study.



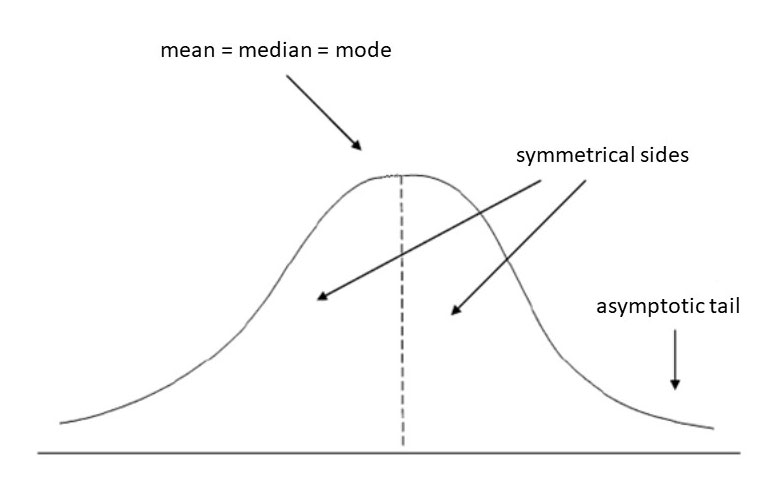
Normal Distribution

Normal distribution (or Gaussian or Gauss or Laplace-Gauss) is one of continuous probability distribution functions. In graph form, normal distribution will appear as a bell curve.

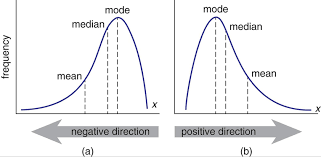
The normal distribution is the limiting case of a discrete binomial distribution as the sample size N becomes large. Binomial distribution is discrete probability distribution of the number of successes in a sequence of n independent questions, each asking a yes-no question.

Normal Distribution has –

* Mean = Median = Mode
* Symmetry around the center
* Total area under the curve is 1



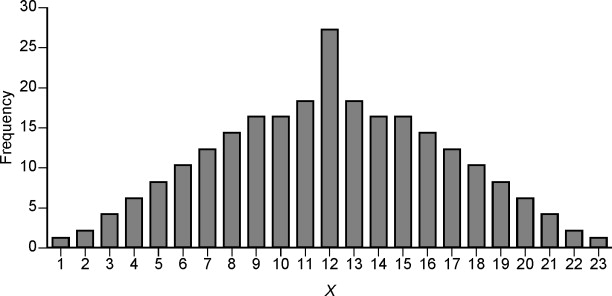
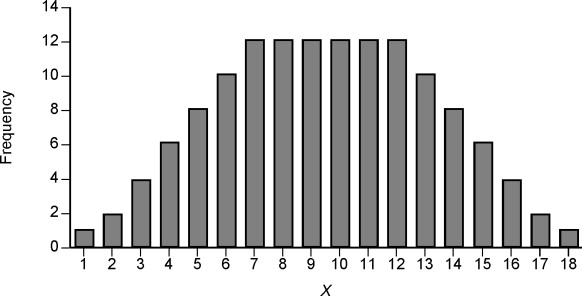
Skewness- It is the degree of distortion from the symmetrical bell curve or the normal distribution.



The negative skewness of the distribution indicates that an investor may expect frequent small gains and few large losses. In reality, many trading strategies employed by traders are based on negatively skewed distributions. Despite the fact that strategies based on negative skewness may provide stable profits, an investor or a trader should be aware that there is still a probability of large losses.

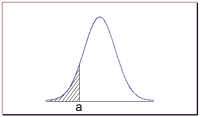
The positive skewness of a distribution indicates that an investor may expect frequent small losses and few large gains from the investment. The positively skewed distributions of investment returns are generally more desirable by investors since there is some probability to gain huge profits that can cover all the frequent small losses.

Kurtosis- It is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution.



Leptokurtic indicates a positive excess kurtosis. The leptokurtic distribution shows heavy tails on either side, indicating the large outliers. In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.

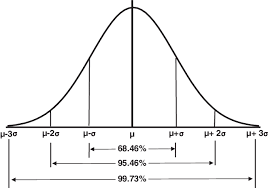
A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution. In the finance context, the platykurtic distribution of the investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.



Area under the curve –

* The total area under the normal curve is equal to 1.
* The probability that a normal random variable X equals any particular value is 0.
* The probability that X is greater than a equals the area under the normal curve bounded by a and plus infinity (as indicated by the non-shaded area in the figure below).
* The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (as indicated by the shaded area in the figure below).

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

* 

**Problem 1**  
An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

*Solution:* Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

* The value of the normal random variable is 365 days.
* The mean is equal to 300 days.
* The standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is: P( X < 365) = 0.90. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

Hypothesis Testing –

Hypothesis testing is a statistical method that is used in making decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

Example –

A person wants to test that a penny has exactly a 50% chance of landing on heads.

Null hypothesis: P(Head) = 50%

Alternative hypothesis: P(Head) != 50%

A random sample of 100 coin flips is taken from a random population of coin flippers, and the null hypothesis is then tested. If it is found that the 100 coin flips were distributed as 40 heads and 60 tails, the analyst would assume that a penny does not have a 50% chance of landing on heads and would reject the null hypothesis and accept the alternative hypothesis.

p-value- It is the probability with which we can either reject or fail to reject the null hypothesis.

The significance level is often referred to by the Greek lower case letter alpha. A common value used for alpha is 5% or 0.05 (or 95% confidence interval). A smaller alpha value suggests a more robust interpretation of the null hypothesis, such as 1% (or 99% confidence interval).

The p-value is compared to the pre-chosen alpha value.

If p-value > alpha: Fail to reject the null hypothesis (i.e. not significant result).

If p-value <= alpha: Reject the null hypothesis (i.e. significant result).

NOTE:

Type I Error: The incorrect rejection of a true null hypothesis or a false positive.

Type II Error: The incorrect failure of rejection of a false null hypothesis or a false negative.

Normality tests-

1. Histogram
2. Q-Q Plot
3. Shapiro-Wilk Test
4. D’Agostino’s K^2 Test
5. Anderson-Darling Test

Histogram- It is a plot to show the frequency distribution of the data

import numpy as np

from matplotlib import pyplot

#Setting the seed

np.random.seed(100)

#Generating random numbers

data = 5 \* np.random.randn(100) +100

data2 = np.random.uniform(0,10,100)

data3 = np.array(list(data)+list(data2))

#Plot Histogram

pyplot.hist(data)

pyplot.show()

pyplot.hist(data2)

pyplot.show()

pyplot.hist(data3)

pyplot.show()

Quantile-Quantile Plot

Another popular plot for checking the distribution of a data sample is the quantile-quantile plot, Q-Q plot, or QQ plot for short. This plot generates its own sample of the idealized distribution that we are comparing with, in this case the Gaussian distribution. The idealized samples are divided into groups (e.g. 5), called quantiles. Each data point in the sample is paired with a similar member from the idealized distribution at the same cumulative distribution.

The resulting points are plotted as a scatter plot with the idealized value on the x-axis and the data sample on the y-axis.

A perfect match for the distribution will be shown by a line of dots on a 45-degree angle from the bottom left of the plot to the top right. Often a line is drawn on the plot to help make this expectation clear. Deviations by the dots from the line shows a deviation from the expected distribution.

from statsmodels.graphics.gofplots import qqplot

qqplot(data, line='s')

pyplot.show()

qqplot(data2, line='s')

pyplot.show()

qqplot(data3, line='s')

pyplot.show()

Shapiro-Wilk Test

It evaluates a data sample and quantifies how likely it is that the data was drawn from a Gaussian distribution. In practice, the Shapiro-Wilk test is believed to be a reliable test of normality, although there is some suggestion that the test may be suitable for smaller samples of data, e.g. thousands of observations or fewer.

from scipy.stats import shapiro

stat, p = shapiro(data)

print('Statistics=%.3f, p=%.3f' % (stat, p))

# interpret

alpha = 0.05

if p > alpha:

print('Sample looks Gaussian (fail to reject H0)')

else:

print('Sample does not look Gaussian (reject H0)')

stat, p = shapiro(data2)

print('Statistics=%.3f, p=%.3f' % (stat, p))

# interpret

alpha = 0.05

if p > alpha:

print('Sample looks Gaussian (fail to reject H0)')

else:

print('Sample does not look Gaussian (reject H0)')

stat, p = shapiro(data3)

print('Statistics=%.3f, p=%.3f' % (stat, p))

# interpret

alpha = 0.05

if p > alpha:

print('Sample looks Gaussian (fail to reject H0)')

else:

print('Sample does not look Gaussian (reject H0)')

D’Agostino’s K^2 Test

The D’Agostino’s K^2 test calculates summary statistics from the data, namely kurtosis and skewness, to determine if the data distribution departs from the normal distribution.

from scipy.stats import normaltest

stat, p = normaltest(data)

print('Statistics=%.3f, p=%.3f' % (stat, p))

# interpret

alpha = 0.05

if p > alpha:

print('Sample looks Gaussian (fail to reject H0)')

else:

print('Sample does not look Gaussian (reject H0)')

stat, p = normaltest(data2)

print('Statistics=%.3f, p=%.3f' % (stat, p))

# interpret

alpha = 0.05

if p > alpha:

print('Sample looks Gaussian (fail to reject H0)')

else:

print('Sample does not look Gaussian (reject H0)')

stat, p = normaltest(data3)

print('Statistics=%.3f, p=%.3f' % (stat, p))

# interpret

alpha = 0.05

if p > alpha:

print('Sample looks Gaussian (fail to reject H0)')

else:

print('Sample does not look Gaussian (reject H0)')

Anderson-Darling Test

It is a statistical test that can be used to evaluate whether a data sample comes from one of among many known data samples

from scipy.stats import anderson

result = anderson(data)

print('Statistic: %.3f' % result.statistic)

p = 0

for i in range(len(result.critical\_values)):

sl, cv = result.significance\_level[i], result.critical\_values[i]

if result.statistic < result.critical\_values[i]:

print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl, cv))

else:

print('%.3f: %.3f, data does not look normal (reject H0)' % (sl, cv))

result = anderson(data2)

print('Statistic: %.3f' % result.statistic)

p = 0

for i in range(len(result.critical\_values)):

sl, cv = result.significance\_level[i], result.critical\_values[i]

if result.statistic < result.critical\_values[i]:

print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl, cv))

else:

print('%.3f: %.3f, data does not look normal (reject H0)' % (sl, cv))

result = anderson(data3)

print('Statistic: %.3f' % result.statistic)

p = 0

for i in range(len(result.critical\_values)):

sl, cv = result.significance\_level[i], result.critical\_values[i]

if result.statistic < result.critical\_values[i]:

print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl, cv))

else:

print('%.3f: %.3f, data does not look normal (reject H0)' % (sl, cv))