FRM Part I Exam

By AnalystPrep

Study Notes - Quantitative Analysis

Last Updated: Jul 20, 2022

Table of Contents

Reading 12: Fundamentals of Probability

After completing this reading, you should be able to:

- Describe an event and an event space.
- Describe independent events and mutually exclusive events.
- Explain the difference between independent events and conditionally independent events.
- Calculate the probability of an event for a discrete probability function.
- Define and calculate a conditional probability.
- Distinguish between conditional and unconditional probabilities.
- Explain and apply Bayes' rule.

Probability is the foundation of statistics, risk management, and econometrics. Probability quantifies the likelihood that some event will occur. For instance, we could be interested in the probability that there will be a defaulter in a prime mortgage facility.

Sample Space, Event Space, and Events

Sample Space (Ω)

A sample space is defined as a collection of all possible occurrences of an experiment. The outcomes are dependent on the problem being studied. For example, when modeling returns from a portfolio, the sample space is a set of real numbers. As another example, assume we want to model defaults in loan payment, we know that there can only be two outcomes: either the firm defaults or it doesn't. As such, the sample space is Ω = {Default, No Default}. To give yet another example, the sample space when a fair six-sided die is tossed is made of six different outcomes:
Ω = {1, 2, 3, 4, 5, 6}

Events (ω)

An event is a set of outcomes (which may contain more than one element). For example, suppose we tossed a die. A “6” would constitute an event. If we toss two dice simultaneously, a {6, 2} would constitute an event. An event that contains only one outcome is termed an elementary event.

Event Space (*F*)

The event space refers to the set of all possible outcomes and combinations of outcomes. For example, consider a scenario where we toss two fair coins simultaneously. The following would constitute our event space:

{HH, HT, TH, TT}

Note: If the coins are fair, the probability of a head, P(H), equals the probability of a tail, P(T).

Probability

The probability of an event refers to the likelihood of that particular event occurring. For example, the probability of a Head when we toss a coin is 0.5, and so is the probability of a Tail.

According to frequentist interpretation, the term probability stands for the number of times an event occurs if a set of independent experiments is performed. But this is what we call the frequentist interpretation because it defines an event’s probability as the limit of its relative frequency in many trials. It is just a conceptual explanation; in finance, we deal with actual, non-experimental events such as the return earned on a stock.

Independent and Mutually Exclusive Events

Mutually Exclusive Events

Two events, A and B, are said to be mutually exclusive if the occurrence of A rules out the occurrence of B, and vice versa. For example, a car cannot turn left and turn right at the same time.

Mutually Exclusive Events - Example

Mutually exclusive events are such that one event precludes the occurrence of all the other events. Thus, if you roll a dice and a 4 comes up, that particular event precludes all the other events, i.e., 1,2,3,5, and 6. In other words, rolling a 1 and a 5 are mutually exclusive events: they cannot occur simultaneously.

Furthermore, there is no way a single investment can have more than one arithmetic mean return. Thus, arithmetic returns of, say, 20% and 17% constitute mutually exclusive events.

Independent Events

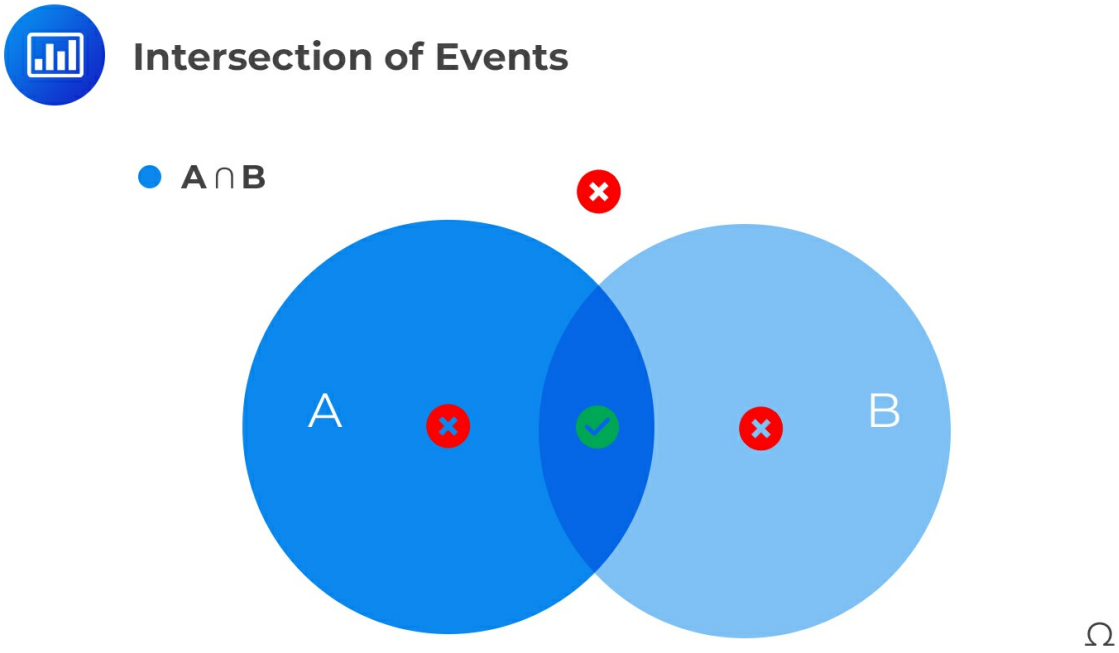
Two events, A and B, are independent if the fact that A occurs does not affect the probability of B occurring. When two events are independent, this simply means that both events can happen at the same time. In other words, the probability of one event happening does not depend on whether the other event occurs or not. For example, we can define A as the likelihood that it rains on March 15 in New York and B as the probability that it rains in Frankfurt on March 15. In this instance, both events can happen simultaneously or not.

Another example would be defining event A as getting tails on the first coin toss and B on the second coin toss. The fact of landing on tails on the first toss will not affect the probability of getting tails on

the second toss.

Intersection

The intersection of events, say A and B, is the set of outcomes occurring both in A **and** B. It is denoted as P(A∩B). Using the Venn diagram, this is represented as:



For independent events,

$$P(A \cap B) = P(A \text{ and } B) = P(A) \times P(B)$$

Independence can be extended to *n* independent events: Let $A_1, A_2, \ldots, A_n$ be independent events then:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \times P(A_2) \times \ldots \times P(A_n)$$

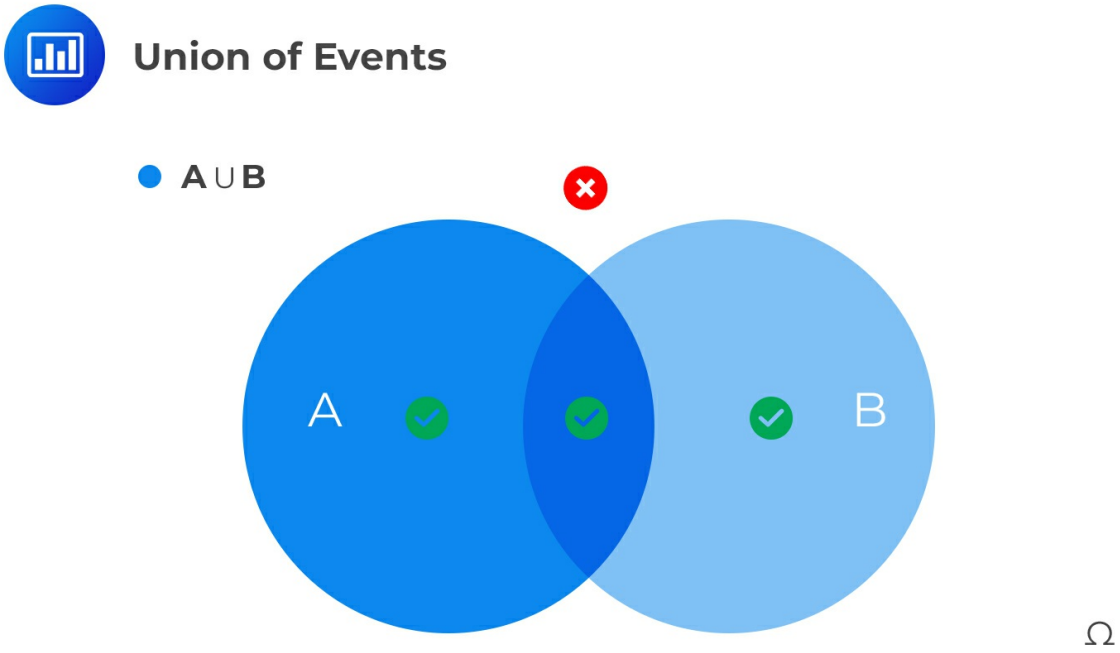
For mutually exclusive events,

$$P(A \cap B) = P(A \text{ and } B) = 0$$

This is because the occurrence of A rules out the occurrence of B. Remember that a car cannot turn left and turn right at the same time!

Union

The union of events say, A and B, is the set of outcomes occurring in at least one of the two sets – A **or** B. It is denoted as P(A∪B). Using the Venn diagram, this is represented as:



To determine the likelihood of any two **mutually exclusive events** occurring, we sum up their individual probabilities. The following is the statistical notation:

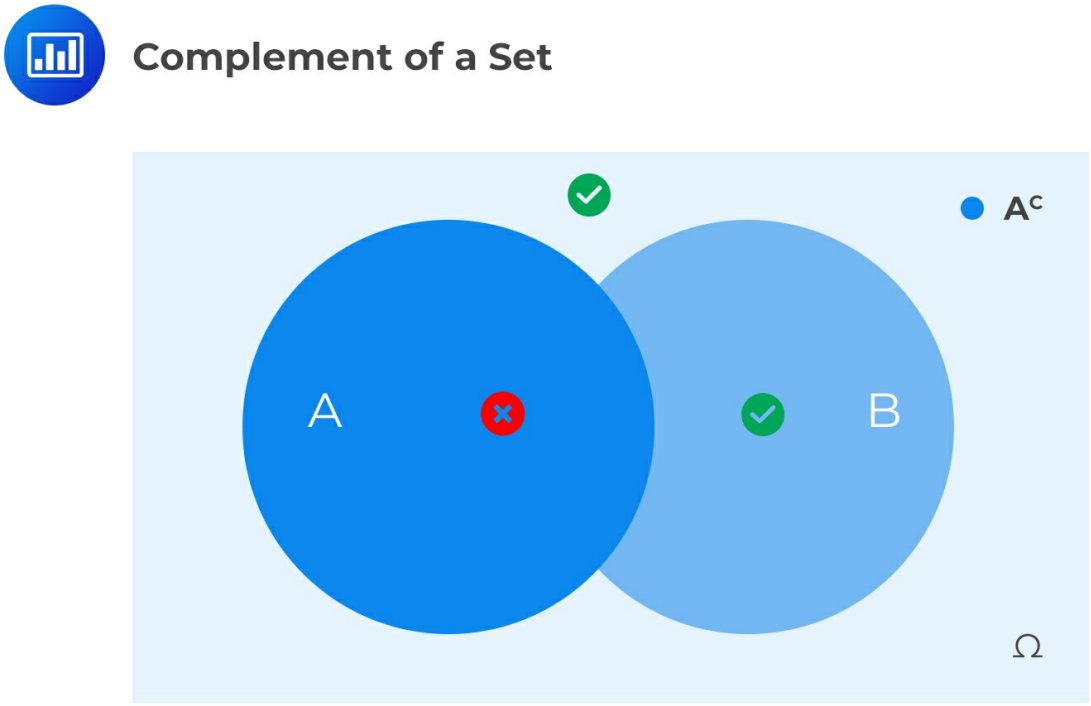
$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$$

Given two events A and B that are not mutually exclusive (**independent events**), the probability that **at least one** of the events will occur is given by:

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

The Complement of a Set

Another important concept under probability, is the **compliment of a set** denoted by $\mathbf{A^c}$ (where **A** can be any other event) which is the set of outcomes that are not in A. For example, consider the following Venn diagram:



This is the first axiom of probability, and it implies that:

$$P(A \cup A^c) = P(A) + P(A^c) = 1$$

Conditional Probability

Until now, we've only looked at unconditional probabilities. An **unconditional probability** (also known as a marginal probability) is simply the probability that an event occurs, without taking into account any other preceding events. In other words, unconditional probabilities are **not** conditioned

on the occurrence of any other events; they are 'stand-alone' events.

Conditional probability is the probability of one event occurring with some relationship to one or more other events. Our interest lies in the probability of an event 'A' **given** that another event 'B '**has already occurred.** Here's what you should ask yourself:

"What is the probability of one event occurring **if** another event has already taken place?"

We pronounce P(A | B) as "the probability of A given B.," and it is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The bar sandwiched between A and B simply indicates "given."

Example: Groups of Investors

In a group of 100 investors,

- 50 buy stocks,
- 30 purchase bonds, and
- 20 purchase stocks and bonds.

If an investor chosen at random, bought bonds, what is the probability they also bought stocks?

Event	Notation	Probability
Buys stocks	A	0.5(=50/100)
Buys bonds	B	0.3 (=30/100)
Buys stocks and bonds	A and B	0.2 (=20/100)

We want the probability of an investor buying stocks given that they have already bought bonds. This is P(A | B):

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{0.2}{0.3} = 0.67$$

Note that we can also make the numerator the subject so that:

$$P(A \cap B) = P(A|B)\,P(B)$$

For independent events, however,

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

This is also true for $P(B|A) = P(B)$.

Bayes' Theorem

Bayes' theorem describes the probability of an event based on prior knowledge of conditions that might be related to the event. Assuming that we have two random variables A and B, then according to Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Applying Bayes' Theorem

Supposing that we are issued with two bonds A and B. Each bond has a default probability of 10% over the year that follows. We are also told that there is a 6% chance that both the bonds will default, an 86% chance that none of them defaults, and a 14% chance that either of the bonds defaults. All of this information can be summarized in a probability matrix.

Often, there is a high correlation between bond defaults. This can be attributed to the sensitivity displayed by bond issuers when dealing with broad economic bonds. The 6% chances of both the bonds defaulting are higher than the 1% chances of default had the default events been independent.

The features of the probability matrix can also be expressed in terms of conditional probabilities.

For example, the likelihood that bond A will default given that B has defaulted is computed as:

$$P\,(A|B) = \frac{P\,[A \cap B]}{P\,[B]} = \frac{6\%}{10\%} = 60\%$$

This means that in 60% of the scenarios in which bond B will default, bond A will also default.

The above equation is often written as:

$$P\,[A \cap B] = P\,(A|B) \times P\,[B] \qquad \text{I}$$

Also:

$$P\,[A \cap B] = P\,(B|A) \times P\,[A] \qquad \text{II}$$

Both the right-hand sides of equations I, and II are combined and rearranged to give the Bayes' theorem:

$$P\,(B\,|A) \times P\,[A] = P\,(A\,|B) \times P\,(B)$$

$$\Rightarrow P\,(A|B) = \frac{P\,(B|A) \times P\,[A]}{P\,[B]}$$

When presented with new data, Bayes' theorem can be applied to update beliefs. To understand how the theorem can provide a framework for how exactly the new beliefs should be, consider the following scenario:

Example: Applying Baye's Theorem

Suppose that an analyst could group fund managers into two categories, star and non-star managers, after evaluating historical data. Given that the best managers are a star and there is a 75% likelihood that in a particular year, the market will be beaten by a star. On the other hand, there are equal chances that non-star managers will either beat the market or underperform it. Furthermore, there is a year-to-year independence in the probabilities that both types of managers will beat the market.

Only 16% of managers within a given cohort become stars. Three years have passed since a new

manager who was able to beat the market every single year was added to the portfolio of the analyst. Determine the chances of the manager being a star when he was first added to the portfolio. What are the chances of him being a star at present? What are his chances of beating the market in the year that follows, given that he has beaten it in the past three years?

Solution

We first summarize the data by introducing some notations as follows: The chances that a manager will beat the market on the condition that he is a star is:

$$P\,(B|S) = 0.75 = \frac{3}{4}$$

The chances of a non-star manager beating the market are:

$$P\,(B|\bar{S}) = 0.5 = \frac{1}{2}$$

The chances of the new manager being a star during the particular time he was added to the analyst's portfolio are exactly the chances that any manager will be made a star, which is unconditional:

$$P\,[S] = 0.16 = \frac{4}{25}$$

To evaluate the likelihood of him being a star at present, we compute the likelihood of him being a star given that he has beaten the market for three consecutive years, P (S|3B), using the Bayes' theorem:

$$P\,(S|3B) = \frac{P\,(3B|S) \times P\,[S]}{P\,[3B]}$$

$$P\,(3B|S) = \left(\frac{3}{4}\right)^3 = \frac{27}{64}$$

The unconditional chances that the manager will beat the market for three years is the denominator.

$$P\,[3B] = P\,(3B|S) \times P\,[S] + P\,(3B|\bar{S}) \times P\,[\bar{S}]$$

$$P[3B] = (\frac{3}{4})^3 \times \frac{4}{25} + (\frac{1}{2})^3 \frac{21}{25} = \frac{69}{400}$$

Therefore:

$$P(S|3B) = \frac{(\frac{27}{64})(\frac{4}{25})}{(\frac{69}{400})} = \frac{9}{23} = 39\%$$

Therefore, there are 39% chances that the manager will be a star after beating the market for three consecutive years, which happens to be our new belief and is a significant improvement to our old belief, which was 16%.

Finally, we compute the chances of the manager beating the market in the following year. This happens to be the summation of the chances of a star beating the market and the chances of a non-star beating the market, weighted by the new belief:

$$P[B] = P(B|S) \times P[S] + P(B|\bar{S}) \times P[\bar{S}]$$

$$P[B] = \frac{3}{4} \times \frac{9}{23} + \frac{1}{2} \times \frac{14}{23} = 60\% = \frac{3}{5}$$

We also have that:

$$P(S|3B) = \frac{P(3B|S) \times P[S]}{P[3B]}$$

The L.H.S of the formula is posterior. The first item on the numerator is the likelihood, and the second part is prior.

Question 1

The probability that the Eurozone economy will grow this year is 18%, and the probability that the European Central Bank (ECB) will loosen its monetary policy is 52%.

Assume that the joint probability that the Eurozone economy will grow and the ECB will loosen its monetary policy is 45%. What is the probability that either the Eurozone economy will grow ***or*** the ECB will loosen its monetary policy?

A. 42.12%

B. 25%

C. 11%

D. 17%

The correct answer is **B**.

The addition rule of probability is used to solve this question:
P(E) = 0.18 (the probability that the Eurozone economy will grow is 18%)
p(M) = 0.52 (the probability that the ECB will loosen the monetary policy is 52%)
p(EM) = 0.45 (the joint probability that Eurozone economy will grow and the ECB will loosen its monetary policy is 45%)

The probability that either the Eurozone economy will grow or the central bank will loosen its the monetary policy:
p(E or M) = p(E) + p(M) - p(EM) = 0.18 + 0.52 - 0.45 = 0.25

Question 2

A mathematician has given you the following conditional probabilities:

$p(O\|T) = 0.62$	Conditional probability of reaching the office if the train arrives on time
$p(O\|T^c) = 0.47$	Conditional probability of reaching the office if the train does not arrive on time
$p(T) = 0.65$	Unconditional probability of the train arriving on time
$p(O) = ?$	Unconditional probability of reaching the office

What is the unconditional probability of reaching the office, p(O)?

A. 0.4325

B. 0.5675

C. 0.3856

D. 0.5244

The correct answer is **B**.

This question can be solved using the total probability rule.

If $p(T) = 0.65$ (Unconditional probability of train arriving on time is 0.65), then the unconditional probability of the train not arriving on time $p(T^c) = 1 - p(T) = 1 - 0.65 = 0.35$.

Now, we can solve for

$$\begin{aligned} p(O) &= p(O|T) * p(T) + p(O|T^c) * p(T^c) \\ &= 0.62 * 0.65 + 0.47 * 0.35 \\ &= 0.5675 \end{aligned}$$

Note: p(O) is the unconditional probability of reaching the office. It is simply the addition of:

1. reaching the office if the train arrives on time, multiplied by the train arriving on time, and
2. reaching the office if the train does not arrive on time, multiplied by the train not arriving on time (or given the information, one minus the train arriving on time)

Question 3

Suppose you are an equity analyst for the XYZ investment bank. You use historical data to categorize the managers as excellent or average. Excellent managers outperform the market 70% of the time and average managers outperform the market only 40% of the time. Furthermore, 20% of all fund managers are excellent managers and 80% are simply average. The probability of a manager outperforming the market in any given year is independent of their performance in any other year.

A new fund manager started three years ago and outperformed the market all three years. What's the probability that the manager is excellent?

A. 29.53%

B. 12.56%

C. 57.26%

D. 30.21%

The correct answer is **C**.

The best way to visualize this problem is to start off with a probability matrix:

Kind of manager	Probability	Probability of beating market
Excellent	0.2	0.7
Average	0.8	0.4

Let E be the event of an excellent manager, and A represent the event of an average manager.

$P(E) = 0.2$ and $P(A) = 0.8$

Further, let O be the event of outperforming the market.

We know that:

$P(O|E) = 0.7$ and $P(O|A) = 0.4$

We want P(E|O):

$$\begin{aligned} P(E|O) &= \frac{P(O|E) \times P(E)}{P(O|E) \times P(E) + P(O|A) \times P(A)} \\ &= \frac{(0.7^3) \times 0.2}{(0.7^3) \times 0.2 + (0.4^3) \times 0.8} \\ &= 57.26\% \end{aligned}$$

Note: The power of three is used to indicate three consecutive years.

Reading 13: Random Variables

After completing this reading, you should be able to:

- Describe and distinguish a probability mass function from a cumulative distribution function and explain the relationship between these two.
- Understand and apply the concept of a mathematical expectation of a random variable.
- Describe the four common population moments.
- Explain the differences between a probability mass function and a probability density function.
- Characterize the quantile function and quantile-based estimators.
- Explain the effect of a linear transformation of a random variable on the mean, variance, standard deviation, skewness, kurtosis, median, and interquartile range.

Random Variables

A random variable is a variable whose possible values are outcomes of a random phenomenon. It is a function that maps outcomes of a random process to real values. It can also be termed as the realization of a random process.

Precisely, if ω is an element of a sample space Ω and x is the realization, then $X(\omega) = x$. Conventionally, random variables are given in upper case (such as X, Y, and Z) while the realized random values are represented in lower case (such as x, y, and z)

For example, let X be the random variable as a result of rolling a die. Therefore, x is the outcome of one roll, and it could take any of the values 1, 2, 3, 4, 5, or 6. The probability that the resulting random variable is equal to 3 can be expressed as:

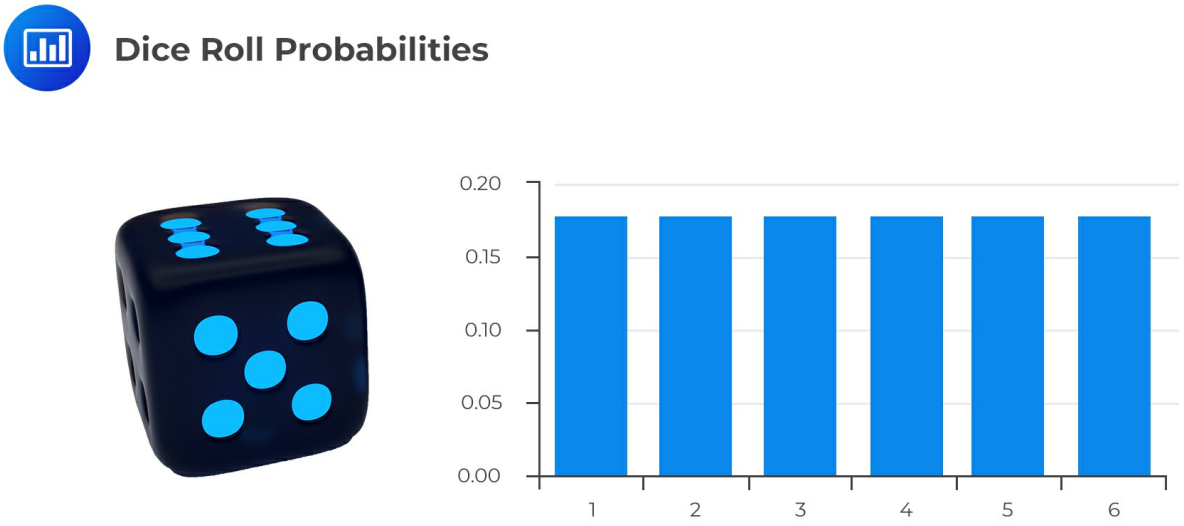
$$P(X = x) \text{ where } x = 3$$

Types of Random Variables

Discrete Random Variables

A discrete random variable is one that produces a set of distinct values. A discrete random variable manifests:

- If the range of all possible values is a **finite set**, e.g., {1,2,3,4,5,6} as in the case of a six-sided die or,
- If the range of all possible values is a **countably infinite set**: e.g. {1,2,3, ... }



Examples of discrete random variables include:

- Picking a random stock from the S&P 500.
- The number of candidates registered for the FRM level 1 exam at any given time.
- The number of study topics in a program.

Probability Functions under Discrete Random Variables

Since the possible values of a random variable are mostly numerical, they can be explained using

mathematical functions. A function $f_X(x) = P(X = x)$ for each x in the range of X is the probability function (PF) of X and explains how the total chance (which is 1) is distributed amongst the possible values of X.

There are two functions used when explaining the features of the distribution of discrete random variables: probability mass function (PMF) and cumulative distribution function (CDF).

Probability Mass Function (PMF)

This function gives the probability that a random variable takes a particular value. Since PMF outputs the probabilities, it should possess the following properties:

1. $f_X(x) \geq 0 \quad \forall$ range of X (value returned must be a nonnegative)
2. $\sum_x f_X(x) = 1$ (sum across all value in support of a random variable should be equal to 1)

Example: Bernoulli Distribution

Assume that X is a Bernoulli random variable, the PMF of X is given by:

$$f_X(x) = p^x(1-p)^{1-x}, X = 0, 1$$

The Random variables in a Bernoulli distribution are 0 and 1. Therefore,

$$f_X(0) = p^0(1-p)^{1-0} = 1 - p$$

And

$$f_X(1) = p^1(1-p)^{1-1} = p$$

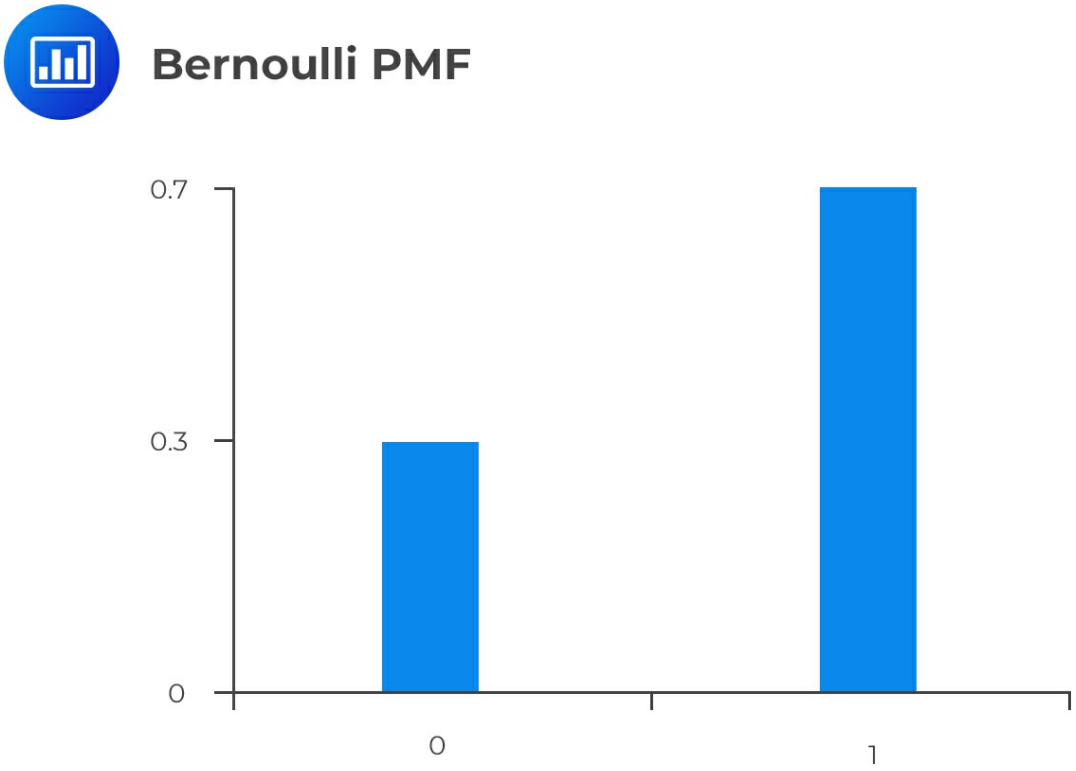
Looking at the above results, the first property $f_X(x) \geq 0$) of probability distributions is met. For the second property:

$$\sum_x f_X(x) = \sum_{x=0,1} f_X(x) = 1 - p + p = 1$$

Moreover, the probability that we observe random variable 0 is 1-p, and the probability of observing random variable 1 is p. More precisely,

$$F_X(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

The graph of the Bernoulli PMF is shown below, assuming the p=0.7. Note that PMF is only defined for X=0,1.



Cumulative Distribution Function (CDF)

CDF measures the probability of realizing a value less than or equal to the input x, $Pr(X \leq x)$. It is denoted by $F_X(x)$ and so,

$$F_X(x) = Pr(X \leq x)$$

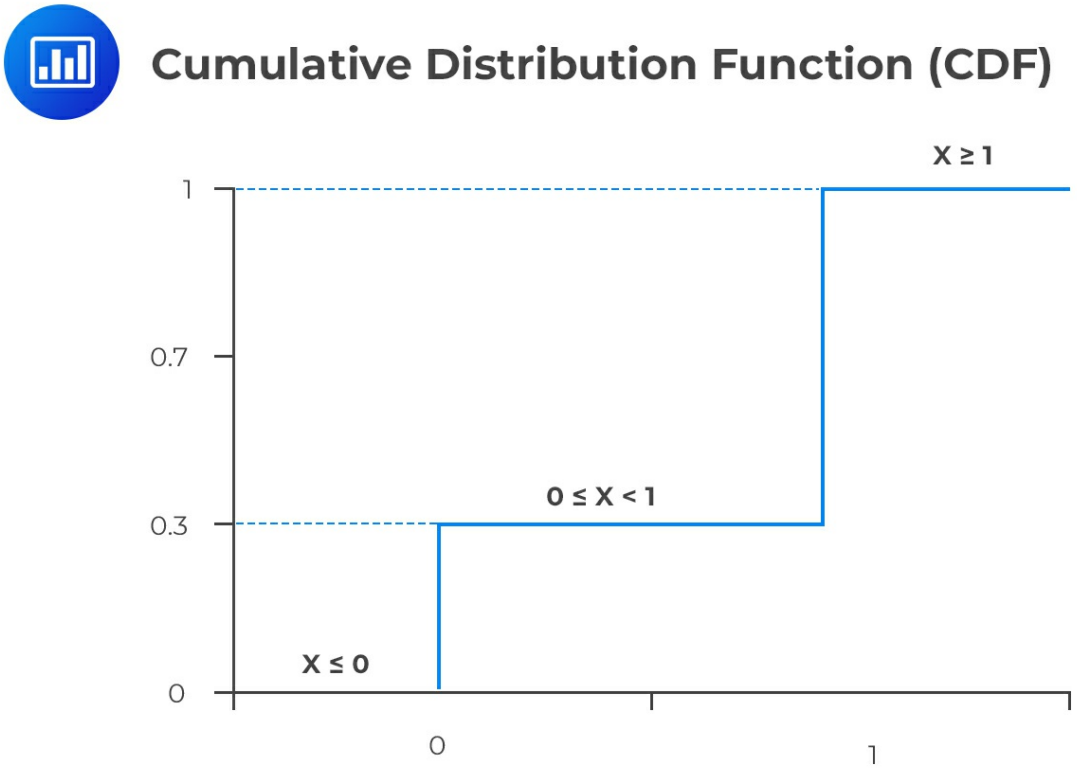
CDF is monotonic and increasing in x since it measures total probability. It is a continuous function (in contrast with PMF) because it supports any value between 0 and 1 (in the case of Bernoulli

random variables) inclusively.

For instance, the CDF of the Bernoulli random variable is:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - p, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$$

$F_X(x)$ is defined for all real values of x. The graph of $F_X(x)$ against x begins at 0 then rises by jumps as values of x are realized for which $p(X = x)$ is positive. The graph reaches its maximum value at 1. For the Bernoulli distribution with p=0.7, the graph is shown below:



Since CDF is defined for all values of x, the CDF for a Bernoulli distribution with a parameter p=0.7 is:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 0.3, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$$

The corresponding graph is as shown above

Relationship Between the CDF and PMF with Discrete Random Variables

The CDF can be represented as the sum of the PMF for all the values that are less than or equal to x. Simply put:

$$F_X(x) = \sum_{t \epsilon R(x), t \leq x} f_X(t)$$

Where R(x) is the range of realized values of X (X=x).

On the other hand, PMF is equivalent to the difference between the consecutive values of X. That is:

$$f_X(x) = F_X(x) - F_X(x-1)$$

Example: PMF and CDF under Discrete Random Variables

There are 8 hens with different weights in a cage. Hens1 to 3 weigh 1 kg, hens 4 and 5 weigh 2kg, and the rest weigh 3kg. We need to develop the PMF and the CDF.

Solution

The random variables (X = 1kg, 2kg, or 3kg) here are the weights of the chicken,

$$f_X(1) = Pr(X = 1) = \frac{3}{8}$$
$$f_X(2) = Pr(X = 2) = \frac{2}{8} = \frac{1}{4}$$
$$f_X(3) = Pr(X = 3) = \frac{3}{8}$$

So, the PMF is:

$$\begin{cases} \frac{3}{8}, & x = 1 \\ \frac{1}{4}, & x = 2 \\ \frac{3}{8}, & x = 3 \end{cases}$$

For the CDF, it includes all the realized values of the random variable. So,

$$F_X(0) = Pr(X \leq 0) = 0$$
$$F_X(1) = Pr(X \leq 1) = \frac{3}{8}$$
$$F_X(2) = Pr(X \leq 2) = \frac{3}{8} + \frac{2}{8} = \frac{5}{8} \left[\text{Using } F_X(x) = \sum_{t \epsilon R(x), t \leq x} f_X(t) \right]$$
$$F_X(3) = Pr(X \leq 3) = \frac{5}{8} + \frac{3}{8} = 1$$

So that the CDF is

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{3}{8}, & 1 \leq x < 2 \\ \frac{5}{8}, & 2 \leq x < 3 \\ 1, & 3 \leq x \end{cases}$$

Note that

$$f_X(x) = F_X(x) - F_X(x-1)$$

Which implies that:

$$f_X(3) = F_X(3) - F_X(2) = 1 - \frac{5}{8} = \frac{3}{8}$$

Which gives the same result as before.

Continuous Random Variables

A continuous random variable can assume **any value along a given interval of a number line**. For instance, $x > 0, (-\infty < x < \infty)$ and $0 < x < 1$. Examples of continuous random variables include the price of stock or bond, or the value at risk of a portfolio at a particular point in time.

The following relationship holds for a continuous random variable X:

$$P[r_1 < X < r_2] = p$$

This implies that p is the likelihood that the random variable X falls between r_1 and r_2.

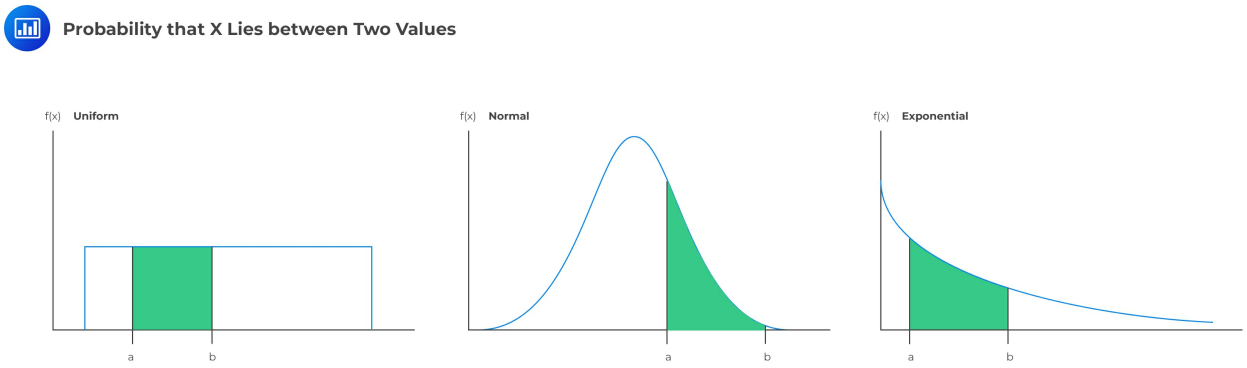
The Probability Density Function (PDF) under Continuous Random Variables

A probability density function (PDF) allows us to calculate the probability of an event.

Given a PDF f(x), we can determine the probability that x falls between a and b:

$$Pr(a < x \leq b) = \int_a^b f(x)\,dx$$

The probability that X lies between two values is the **area under** the density function graph between the two values:



Probability distribution function is another term used to refer to the probability density function. The properties of the PDF are the same as those of PMF. That is:

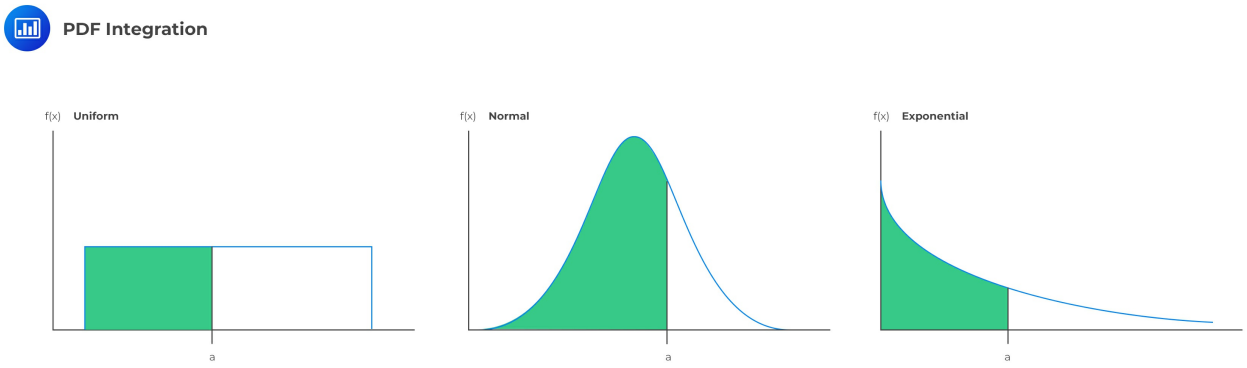
1. $f_X(x) \geq 0, -\infty \leq x \leq \infty$ (nonnegativity)
2. $\int_{r_{min}}^{r_{max}} f(x)dx = 1$(The sum of all probabilities must be equal to 1, just like in discrete random variables)

The upper and lower bounds of f(x) are defined by r_{min} and r_{max}

Cumulative Distribution Functions (CDF) under Continuous Random Variables

It is also called the cumulative density function and is closely related to the concept of a PDF. CDFA CDF defines the likelihood of a random variable falling below a specific value. To determine the CDF,

the PDF is integrated from its lower bound.



The corresponding density function's capital letter has traditionally been used to denote the CDF. The following computation depicts a CDF, F(x), of a random variable X whose PDF is f(x):

$$F(a) = \int_{-\infty}^{a} f(x)d(x) = P[X \leq a]$$

The region under the PDF is a depiction of the CDF. The CDF is usually non-decreasing and varies from zero to one. We must have a zero CDF at the minimum value of the PDF. The variable cannot be less than the minimum. The likelihood of the random variable is less than or equal to the maximum is 100%.

To obtain the PDF from the CDF, we have to compute the first derivative of the CDF. Therefore:

$$f(x) = \frac{dF(x)}{dx}$$

Next, we look at how to determine the probability that a random variable X will fall between some two values, a and b.

$$P[a < X \leq b] = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

Where a is less than b.

The following relationship is also true:

$$P[X > a] = 1 - F(a)$$

Example:Formulating the CDF of a Continuous Random Variable

The continuous random variable X has a pdf of $f(x) = 12x^2(1-x)$ for $0 < x < 1$. We need to find the expression for F(x).

Solution

We know that:

$$F(x) = \int_{-\infty}^{x} f(t)d(t)$$
$$F(x) = \int_{0}^{x} 12t^2(1-t)d(t) = [4t^3 - 3t^4]_0^x = x^3(4-3x)$$

So,

$$F(x) = x^3(4-3x)$$

Expected Values

The expected values are the numerical summaries of features of the distribution of random variables. Denoted by E[X] or μ, it gives the value of X that is the measure of average or center of the distribution of X. The expected value is the mean of the distribution of X.

For discrete random variables, the expected value is given by:

$$E[X] = \sum_{x} xf(X)$$

It is simply the sum of the product of the value of the random variable and the probability assumed by the corresponding random variable.

Example: Calculating the Expected Value in Discrete Random Variable

There are 8 hens with different weights in a cage. Hens 1 to 3 weigh 1 kg, hens 4 and 5 weigh 2kg,

and the rest weigh 3kg. We need to calculate the mean weight of the hens.

Solution

We had calculated the PDF as:

$$f(x) = \begin{cases} \frac{3}{8}, & x = 1 \\ \frac{1}{4}, & x = 2 \\ \frac{3}{8}, & x = 3 \end{cases}$$

Now,

$$E[X] = \sum_x xf(X) = 1 \times \frac{3}{8} + 2 \times \frac{1}{4} + 3 \times \frac{3}{8} = 2$$

So, the mean weight of the hens in the cage is 2kg.

For the continuous random variable, the mean is given by:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Basically, it is all about integrating the product of the value of the random variable and the probability assumed by the corresponding random variable.

Example: Calculating the Expected Value of a Continuous Random Variable

The continuous random variable X has a pdf of $f(x) = 12x^2(1-x)$ for $0 < x < 1$.

We need to calculate E[X].

Solution

We know that:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

So,

$$E(X) = \int_0^1 x12x^2(1-x)d(x) = [3x^4 - \frac{12}{5}x^5]_0^1 = 0.6$$

For random variables that are functions, we apply the same method as that of a "single" random variable. That is, summing or integrating the product of the value of the random variable function and the probability assumed by the corresponding random variable function.

Assume that the random variable function is g(x). Then:

$$E[g(x)] = \sum_x g(x)f(x)$$

for the discrete case and

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

for the continuous case.

Example: Calculating the Expected Values Involving Functions as Random Variable.

A random variable X has PDF of:

$$f_X(x) = \frac{1}{5}x^2, \text{ for } 0 < x < 3$$

Calculate E(2X + 1)

Solution

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$
$$= \int_{-\infty}^{\infty} \frac{1}{5}(2x+1)x^2dx = \frac{1}{5}[\frac{x^4}{2} + \frac{x^3}{3}]_0^3 = 9.9$$

Properties of Expectation

The expectation operator is a linear operator. Consequently, the expectation of a constant is a constant. That is, E(c)=c. Moreover, the expected value of a random variable is a constant and not a random variable.

For non-linear function g(x),E(g(x))≠ g(E(x)). For instance, $E\left(\frac{1}{X}\right) \neq \frac{1}{E(X)}$

The Variance of a Random Variable

The variance of random variable measures the spread (dispersion or variability) of the distribution about its mean. Mathematically,

$$Var(X) = E(X^2) - E(X)^2 = E[X - E(X)]^2$$

Intuitively, the standard deviation is the square root of the variance. Now, denoting $E(X) = \mu$, then:

$$Var(X) = E(X^2) - \mu^2$$

Example: Calculating the Variance of Random Variable

The continuous random variable X has a pdf of $f(x) = 12x^2(1-x)$ for $0 < x < 1$.

We need to calculate Var[X].

Solution

We know that:

$$Var(X) = E(X^2) - E(X)^2$$

We had calculated E(X)=0.6

We have to calculate:

$$E(X^2)$$

$$E(X) = \int_0^1 x.[12x^2(1-x)]dx = [3x^4 - \frac{12}{5}x^5]_0^1 = 0.6$$
$$E(X^2) = \int_0^1 12x^4 - 12x^5 dx = [\frac{12}{5}x^5 - 2x^6]_0^1 = 0.4$$

So,

$$Var(X) = 0.4 - 0.6^2 = 0.04$$

Moments

Moments are defined as the expected values that briefly describe the features of a distribution. The first moment is defined to be the expected value of X:

$$\mu_1 = E(X)$$

Therefore, the first moment provides the information about the average value. The second and higher moments are broadly divided into Central and Non-central moments

Central Moments

The general formula for the central moments is:

$$\mu_k = E([X - E(X)]^k), k = 2, 3 \ldots$$

Where k denotes the order of the moment. Central moments are moments about the mean.

Non-Central Moments

Non-central moments describe those moments about 0. The general formula is given by:

$$\mu_k = E(X^k)$$

Note that the central moments are constructed from the non-central moments and the first central and non-central moments are equal ($\mu_1 = E(X)$).

Population Moments

The four common population moments are: mean, variance, skewness, and kurtosis.

The Mean

The mean is the first moment and is given by:

$$\mu = E(X)$$

It is the average (also called the location of the distribution) value of X.

The Variance

This is the second moment. It is presented as:

$$\sigma^2 = E([X - E(X)]^2) = E[(X - \mu)^2]$$

The variance measures the spread of the random variable from its mean. The standard deviation (σ) is the square root of the variance. The standard deviation is more commonly quoted in the world of finance because it is easily comparable to the mean since they share the measurement units.

The Skewness

Skewness is a cubed standardized central moment given by:

$$\text{skew}(X) = \frac{E([X - E(X)])^3}{\sigma^3} = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right]$$

Note that $\frac{X - \mu}{\sigma}$ is a standardized X with a mean of 0 and a variance of 1.

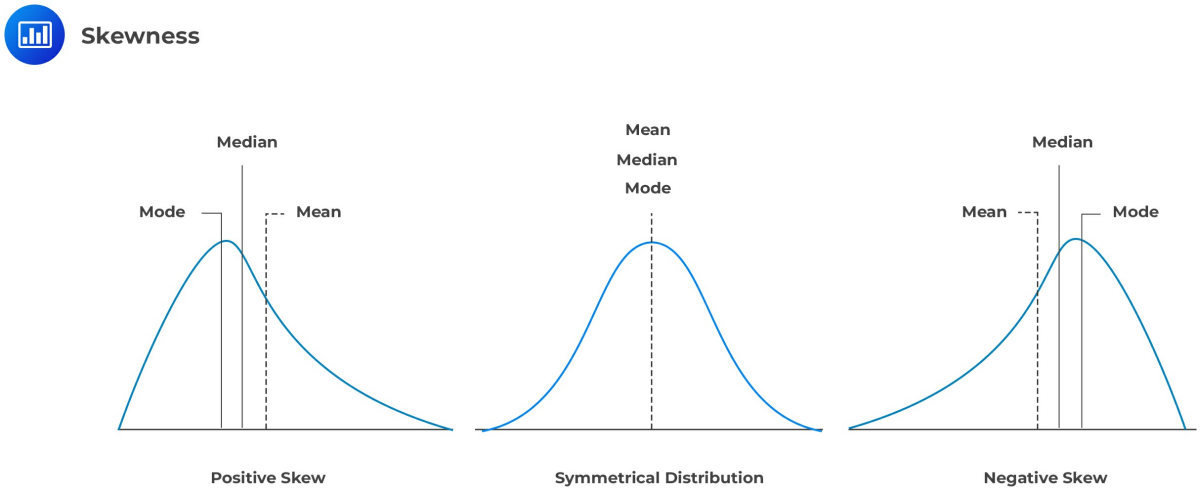
Skewness can be positive or negative.

Positive skew

- The right tail is longer
- The mass of the distribution is concentrated on the left
- There are a few relatively high values.
- In most cases (but not always), the mean is greater than the median, or equivalently, the mean is greater than the mode; in which case the skewness is greater than zero.

Negative skew

- The left tail is longer
- The mass of the distribution is concentrated on the right
- The distribution has a few relatively low values.
- In most cases (but not always), the mean is lower than the median, or equivalently, the mean is lower than the mode, in which case the skewness is lower than zero.

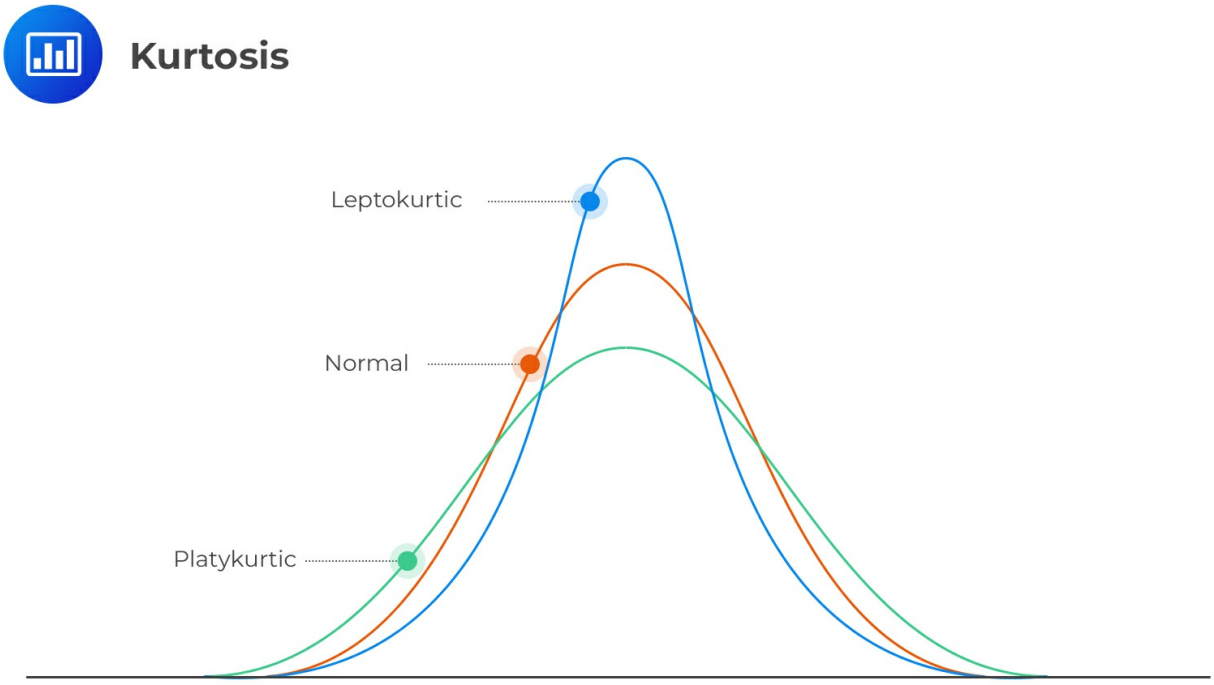


Kurtosis

The Kurtosis is defined as the fourth standardized moment given by:

$$\text{Kurt}(X) = \frac{E([X - E(X)]^4}{\sigma^4} = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right]$$

The description of kurtosis is analogous to that of the Skewness only that the fourth power of the Kurtosis implies that it measures the absolute deviation of random variables. The reference value of a normally distributed random variable is 3. A random variable with Kurtosis exceeding 3 is termed to be **heavily or fat-tailed**.



Effect of Linear Transformation on Moments

In very basic terms, a **linear transformation** is a change to a variable characterized by one or more of the major math operations:

- adding a constant to the variable,
- subtracting a constant from the variable,
- multiplying the variable by a constant,

- and/or dividing the variable by a constant.

Transformation results in the formation of a new random variable.

If X is a random variable and α and β are constants, then $\alpha + \beta x$ is a linear transformation of X. α is referred to as **the shift constant,** and β is the **scale constant**. The transformation *shifts* X by α and *scales* it by β. The process results in the formation of a new random variable, usually denoted by Y.

$$Y = \alpha + \beta x$$

Linear transformation of random variables is informed by the fact that many variables used in finance and risk management do not have a natural scale.

Example: Linear Transformation of Random Variables

Suppose your salary is α dollars per year, and you are entitled to a bonus of β dollars for every dollar of sales you successfully bring in. Let X be what you sell in a certain year. How much in total do you make?

Solution

We can linearly transform the sales variable X into a new variable Y that represents the total amount made.

$$Y = \alpha + \beta x$$

Where α serves as the shift constant and β as the scale constant.

Effect on Mean and Variance

If $Y = \alpha + \beta x$, where α and β are constants. The mean of Y is given by:

$$E(Y) = E(\alpha + \beta x) = \alpha + \beta E(X)$$

The variance is given by:

$$Var(Y) = Var(\alpha + \beta x) = \beta^2 Var(X) = \beta^2 \sigma^2$$

The shift parameter α does not affect the variance. Why? Because variance is a measure of spread from the mean; adding α does not change the spread but merely shifts the distribution to the left or right.

The standard deviation of Y is given by:

$$\sqrt{\beta^2 \sigma^2} = |\beta| \sigma$$

It also follows that α does not affect the standard deviation.

Effect on Skewness and Kurtosis

It can also be shown that if β is positive (so that $Y = \alpha + \beta x$ is an increasing transformation), then the skewness and kurtosis of Y are identical to the skewness and kurtosis of X. This is because both moments are defined on standardized quantities, which removes the effect of the shift constant α and the scaling factor β. This can be seen as follows:

We know that:

$$\text{skew}(X) = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right]$$

Now,

$$\text{skew}(Y) = \frac{E([Y - E(Y)])^3}{\sigma^3} = E\left[\left(\frac{Y - E(Y)}{\sigma}\right)^3\right]$$

$$= E\left[\left(\frac{\alpha + \beta X - (\alpha + \beta\mu)}{\beta\sigma}\right)^3\right]$$

$$= E\left[\left(\frac{\beta(X - \mu)}{\beta\sigma}\right)^3\right] = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] = \text{Skew}(X)$$

However, if $\beta < 0$, the magnitude of skewness of Y is the same as that of X but with the opposite sign because of the odd power (i.e., 3). On the other hand, the kurtosis is unaffected because it uses an even power (i.e., 4).

Quantiles and Modes

Just like any data, quantities such as the quantiles and the modes are used to describe the distribution.

The Quantiles

For a continuous random variable X, the α-quartile of X is the smallest number m such that:

$$Pr(X < m) = \alpha$$

Where $\alpha \epsilon [0, 1]$

For instance, if X is a continuous random variable, the median is defined to be the solution of:

$$P(X < m) = \int_{-\infty}^{m} f_X(x)dx = 0.5$$

Similarly, the lower and upper quartile is such that $P(X < Q_1) = 0.25$ and $P(X < Q_3) = 0.75$

The interquartile range (IQR), is an alternative measure of spread. It is given by:

$$IQR = Q_3 - Q_1$$

Example: Calculating the Quartiles of a PDF

The random variable X has a pdf given by:

$$f_X(x) = 3e^{-2x}, x > 0$$

. Calculate the median of the distribution.

Solution

Denote the median by m. Then m is such that:

$$P(X < m) = \int_0^m 3e^{-2x}dx = 0.5$$

So,

$$\begin{aligned} &= [-\frac{3}{2}e^{-2x}]_0^m = 0.5 \\ &= -\frac{3}{2}e^{-2m} + \frac{3}{2} = 0.5 \\ \Rightarrow m &= -\frac{1}{2} \times \ln\frac{2}{3} = 0.2027 \end{aligned}$$

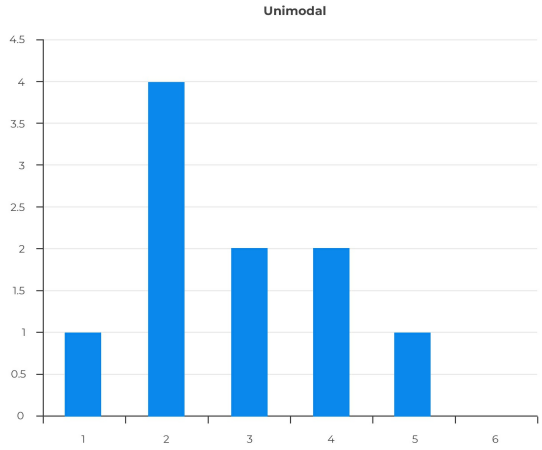
Mode

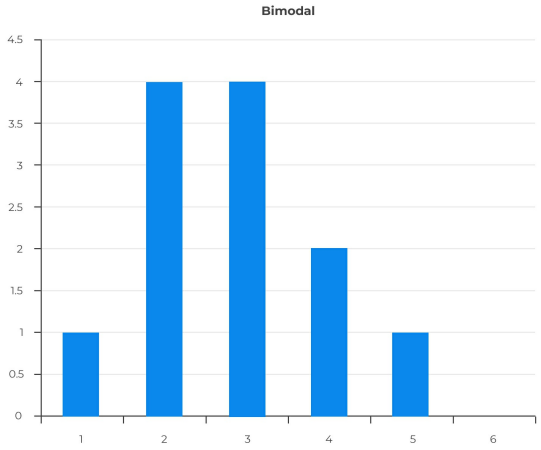
The mode measures the common tendency, that is, the location of the most observed value of a random variable. In a continuous random variable, the mode is represented by the highest point in the PDF.

Random variables can be unimodal if there's just one mode, bimodal if there are two modes, or multimodal if there are more than two modes.

The graph below shows the difference between unimodal and bimodal distributions.

Difference between Unimodal and Bimodal Distributions





Question 1

If a random variable X has a mean of 4 and a standard deviation of 2, calculate Var(3 - 4x)

A. 29

B. 30

C. 64

D. 35

Solution

The correct answer is C.

Recall that:

$$\text{Var}(\alpha + \beta x) = \beta^2 \text{Var}(Y)$$

So,

$$\text{Var}(3 - 4X) = (-4)^2 \text{Var}(X) = 16\ \text{Var}(X)$$

But we are given that the standard deviation is 2, implying that the variance is 4.

Therefore,

$$\text{Var}(3 - 4X) = 16 \times 4 = 64$$

Question 2

A continuous random variable has a pdf given by $f_X(x) = ce^{-3x}$ for all $x > 0$. Calculate Pr(X<6.5)

A. 0.4532

B. 0.4521

C. 0.3321

D. 0.9999

Solution

The correct answer is **D.**

We need to find the constant c first. We know that:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

So,

$$\int_{0}^{\infty} ce^{-3x}dx = 1 = c[-\frac{1}{3}e^{-3x}]_{0}^{\infty} = c\,[0 - -\frac{1}{3}] = 1$$
$$\Rightarrow c = 3$$

Therefore, the PDF is $f_X(x) = 3e^{-3x}$ so that $Pr(X > 6.5)$ is given by:

$$\int_{0}^{6.5} 3e^{-3x}dx = 3[-\frac{1}{3}e^{-3x}]_{0}^{6.5} = c\,[-\frac{1}{3}e^{-3\times 6.5} - -\frac{1}{3}]$$
$$= 0.9999$$

Reading 14: Common Univariate Random Variables

After completing this reading, you should be able to:

- Distinguish the key properties among the following distributions: uniform distribution, Bernoulli distribution, Binomial distribution, Poisson distribution, normal distribution, lognormal distribution, Chi-squared distribution, student's t, and F-distributions, and identify common occurrences of each distribution.
- Describe a mixture distribution and explain the creation and characteristics of mixture distributions.

Parametric Distributions

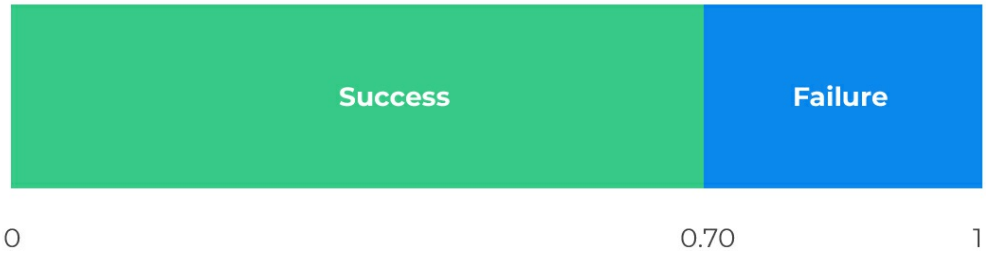
There are two types of distributions, namely parametric and non-parametric distributions. Functions mathematically describe parametric distributions. On the other hand, one cannot use a mathematical function to describe a non-parametric distribution. Examples of parametric distributions are uniform and normal distributions.

Discrete Random Variables

Bernoulli Distribution

Bernoulli distribution is a discrete random variable that takes on values of 0 and 1. This distribution is suitable for scenarios with binary outcomes, such as corporate defaults. Most of the time, 1 is always labeled "success" and 0 a "failure."

Bernoulli Distribution



The Bernoulli distribution has a parameter p which is the probability of success, i.e., the probability that X=1, then:

$$P[X = 1] = p \quad \text{and} \quad P[X = 0] = 1 - p$$

The probability mass function of the Bernoulli distribution stated as $X \sim \text{Bernoulli}(p)$ is given by:

$$f_X(x) = p^x(1 - p)^{1-x}$$

Therefore, the mean and variance of the distribution are computed as:

The PMF confirms that:

$$P[X = 1] = p \quad \text{and} \quad P[X = 0] = 1 - p$$

The CDF of a Bernoulli distribution is a step function given by:

$$F_X(x) = \begin{cases} 0, y < 0 \\ 1 - p, 0 \leq y < 1 \\ 1, y \geq 1 \end{cases}$$

Therefore, the mean and variance of the distribution are computed as:

$$E(X) = p \times 1 + (1 - p) \times 0 = p$$

$$V(X) = E(X^2) - [E(X)]^2 = [p \times 1^2 + (1 - p) \times 0^2] - p^2 = p(1 - p)$$

Example: Bernoulli Distribution

What is the ratio of the mean to variance for X~Bernoulli(0.75)?

Solution

We know that for Bernoulli Distribution,

$$E(X) = p$$

and

$$V(X) = p(1 - p)$$

So,

$$\frac{E(X)}{V(X)} = \frac{p}{p(1-p)} = \frac{1}{0.25} = 4$$

Thus, E(X): V(X)=4:1

Binomial Distribution

A binomial distribution is a collection of Bernoulli random variables. A binomial random variable quantifies the total number of successes from an independent Bernoulli random variable, with the probability of success being *p* and, of course, the failure being *1-p.* Consider the following example:

Suppose we are given two independent bonds with a default likelihood of 10%. Then we have the following possibilities:

- Both do not default,
- Both of them default, or

- Only one of them defaults.

Let X represent the number of defaults:

$$P[X = 0] = (1 - 10\%)^2 = 81\%$$

$$P[X = 1] = 2 \times 10\% \times (1 - 10\%) = 18\%$$

$$P[X = 2] = 10\%^2 = 1\%$$

If we possess three independent bonds having a 10% default probability then:

$$P[X = 0] = (1 - 10\%)^3 = 72.9\%$$

$$P[X = 1] = 3 \times 10\% \times (1 - 10\%)^2 = 24.3\%$$

$$P[X = 2] = 3 \times 10\%^2 \times (1 - 10\%) = 2.7\%$$

$$P[X = 3] = 10\%^3 = 0.1\%$$

Suppose now that we have n bonds. The following combination represents the number of ways in which k of the n bonds can default:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad \ldots\ldots\ldots \text{equation I}$$

If p is the likelihood that one bond will default, then the chances that any particular k bonds will default is given by:

$$p^x(1-p)^{n-x} \quad \ldots\ldots\ldots\ldots \text{equation II}$$

Combining equation I and II, we can determine the likelihood of k bonds defaulting as follows:

$$P[X = x] = \binom{n}{x} p^x(1-p)^{n-x} = \binom{n}{x} p^x(1-p)^{n-x} \quad \text{for} x = 0, 1, 2, \ldots n$$

This is the PDF for the binomial distribution.

Therefore, binomial distribution has two parameters: n and p and usually stated as X B(n, p).

The CDF of a binomial distribution is given by:

$$\sum_{i=1}^{|x|} \binom{n}{i} p^i (1-p)^{n-i}$$

Where |x| implies a random variable less than or equal to x.

The mean and variance of the binomial distribution can be evaluated using moments. The mean and variance are given by:

$E(X) = np$

And

$V(X) = np(1-p)$

The binomial can be approximated using a normal distribution (as will be seen later) if $np \geq 10$ and $n(1-p) \geq 10$

Example: Binomial Distribution

Consider a Binomial distribution $X \sim B(4,0.6)$. Calculate $P(X \geq 3)$.

Solution

We know that for binomial distribution:

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

In this case, $n = 4$ and $p = 0.6$

$$\Rightarrow P(X \geq 3) = P(X = 3) + P(X = 4) = \binom{4}{3} p^3 (1-p)^{4-3} + \binom{4}{4} p^4 (1-p)^{4-4}$$

$$= \binom{4}{3} 0.6^3 (1-0.6)^{4-3} + \binom{4}{4} 0.6^4 (1-0.6)^{4-4}$$

$$= 0.3456 + 0.1296 = 0.4752$$

Poisson Distribution

Events are said to follow a Poisson process if they happen at a constant rate over time, and the likelihood that one event will take place is independent of all the other events,for instance,the number of defaults that occur in each month.

Suppose that X is a Poisson random variable, stated as $X \sim Poisson(\lambda)$ then the PMF is given by:

$$P[X = x] = \frac{\lambda^x e^{-\lambda}}{x!}$$

The CDF of a Poisson distribution is given by:

$$\sum_{i=1}^{|x|} \frac{\lambda^i}{i!}$$

The Poisson parameter λ (lambda), termed as the hazard rate, represents the mean number of events in an interval. Therefore, the mean and variance of the Poisson distribution are given by:

$E(X) = \lambda$

And

$V(X) = \lambda$

Example: Poisson Distribution

A fixed income portfolio is made of a huge number of independent bonds. The average number of bonds defaulting every month is 10. What is the probability that there are exactly 5 defaults in one month?

Solution

For Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

For this question, we have that: $\lambda = 10$ and we need:

$$P(X = 5) = \frac{10^5 e^{-10}}{5!} = 0.03783$$

The notable feature of a Poisson distribution is that it is infinitely divisible. That is, if $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ and that $Y = X_1 + X_2$ then,

$$Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$$

Therefore, Poisson distribution is suitable for time series data since summing the number of events in the sampling interval does not distort the distribution.

Continuous Random Variables

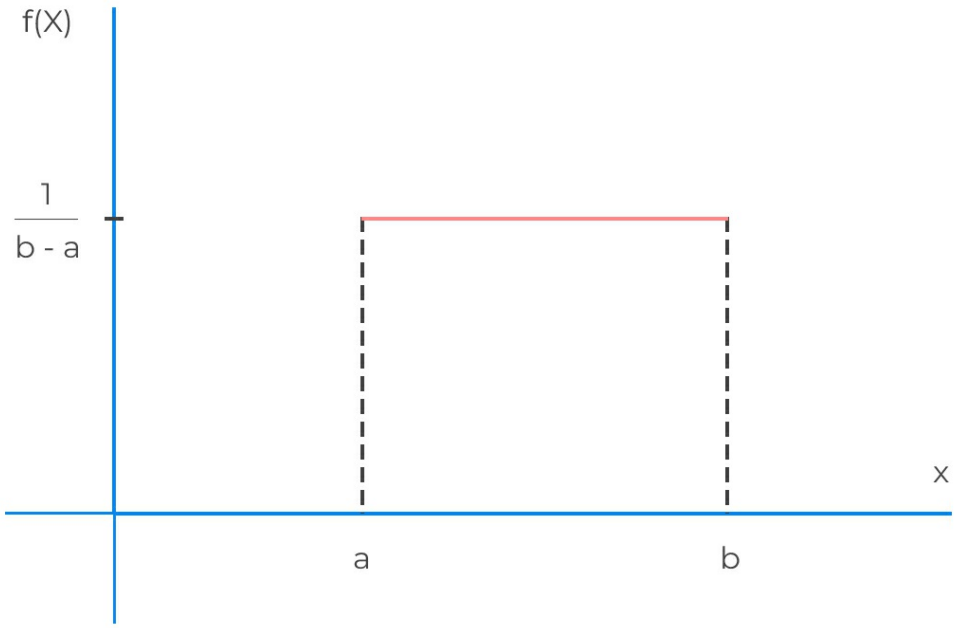
Uniform Distribution

A uniform distribution is a continuous distribution, which takes any value within the range [a,b], which is equally likely to occur.

The PDF of a uniform distribution is given by:

$$f_X(x) = \frac{1}{b-a}$$

PDF of a Uniform Distribution



Note that the PDF of a uniform random variable does not depend on x since all values are equally likely.

The CDF of the uniform distribution is:

$$F_X(x) = \begin{cases} 0, x < a \\ \frac{x-a}{b-a}, a \leq x \leq b \\ 1, x \geq b \end{cases}$$

When a=0 and b=1, the distribution is called the standard uniform distribution. From this distribution, we can construct any uniform distribution, U_2 and U_1 using the formula:

$$U_2 = a + (b - a)\, U_1$$

Where a and b are limits of U_2

The uniform distribution is denoted by $X \sim U(a, b)$, and the mean and variance are given by:

$$E(X) = \frac{a+b}{2}$$

$$V(X) = \frac{(b-a)^2}{12}$$

For instance, the variance of the standard uniform distribution $U_1 \sim N(0,1)$ is given by:

$$E(X) = \frac{(0+1)}{2} = \frac{1}{2}$$

And

$$V(X) = \frac{(1-0)^2}{12} = \frac{1}{12}$$

Assume that we want to calculate the probability that X falls in the interval $l < X < u$ where l is the lower limit and u is the upper limit. That is, we need $P(l < X < u)$ given that $X \sim U(a,b)$. To compute this, we use the formula:

$$P(l < X < u) = \frac{\min(u,b) - \max(l,a)}{b-a}$$

Intuitively, if $l \geq a$ and $u \leq b$, the formula above simplifies into:

$$\frac{u-l}{b-a}$$

Given the uniform distribution $X\ U(-5,10)$, calculate the mean, variance, and $P(-3 < X < 6)$.

Solution

For uniform distribution,

$$E(X) = \frac{a+b}{2} = \frac{-5+10}{2} = 2.5$$

And

$$V(X) = \frac{(10 - -5)^2}{12} = \frac{225}{12} = 18.75$$

For $P(-3 < X < 6)$, using the formula:

$$P(l < X < u) = \frac{\min(u, b) - \max(l, a)}{b - a}$$

$$P(-3 < X < 6) = \frac{\min(6, 10) - \max(-3, -5)}{10 - -5} = \frac{6 - -3}{10 - -5} = \frac{9}{15} = 0.60$$

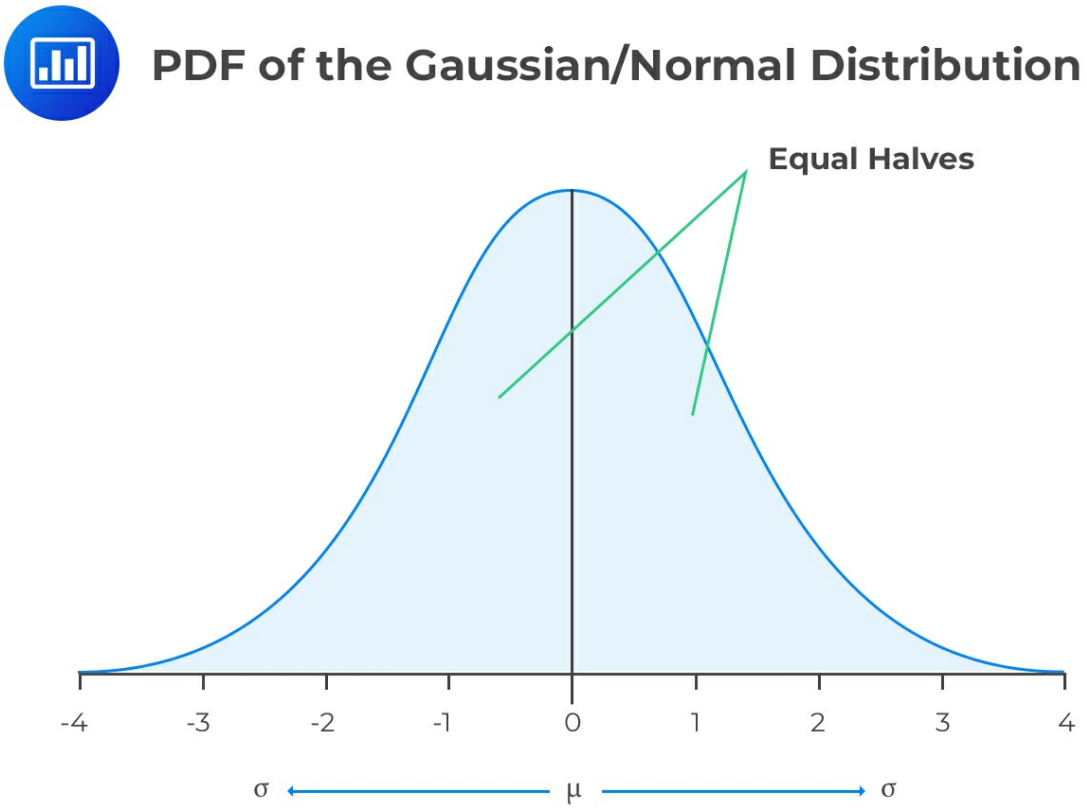
Alternatively, you can think of the probability as the area under the curve. Note that the height of the uniform distribution is \frac{1}{b-a}) and the length $u - l$.

That is:

$$\frac{1}{b - a} \times (u - l) = \frac{1}{10 - -5} \times (6 - -3) = \frac{9}{15} = 0.60$$

Normal Distribution

Also called the Gaussian distribution, the normal distribution has a symmetrical PDF, and the mean and median coincide with the highest point of the PDF. Furthermore, the normal distribution always has a skewness of 0 and a kurtosis of 3.



The following is the formula of a PDF that is normally distributed, for a given random variable X:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty$$

When a variable is normally distributed, it is often written as follows, for convenience:

$$X \sim N(\mu, \sigma^2)$$

Where $E(X) = \mu$ and $V(X) = \sigma^2$

We read this as X is normally distributed, with a mean, μ,and variance of σ^2. Any linear combination of independent normal variables is also normal. To illustrate this, assume X and Y are two variables that are normally distributed. We also have constants a and b. Then Z will be normally distributed such that:

$$Z = aX + bY, \quad \text{such that} \quad Z \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

For instance for $a = b = 1$, then $Z = X + Y$ and thus $Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

A standard normal distribution is a normal distribution whose mean is 0 and standard deviation is 1.It is denoted by N(0,1) and its PDF is as shown below:

$$\varnothing = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

To determine a normal variable whose standard deviation is σ and mean is μ, we compute the product of the standard normal variable with σ and then add the mean:

$$X = \mu + \sigma\varnothing \Rightarrow X \sim N(\mu, \sigma^2)$$

Three standard normal variables X_1, X_2, and X_3 are combined in the following way to construct two normal variables that are correlated:

$$X_A = \sqrt{\rho}X_1 + \sqrt{1-\rho}X_2$$

$$X_B = \sqrt{\rho}X_1 + \sqrt{1-\rho}X_3$$

Where X_A and X_B have a correlation of ρ, and are standard normal variables.

The z-value measures how many standard deviations the corresponding *x* value is above or below the mean. It is given by:

$$\Phi(z) = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

And

$$X \sim N(\mu\sigma^2)$$

Converting X normal random variables is termed as standardization. The values of z are usually tabulated.

For example, consider the normal distribution X~N(1,2). We wish to calculate P(X>2).

Solution

For

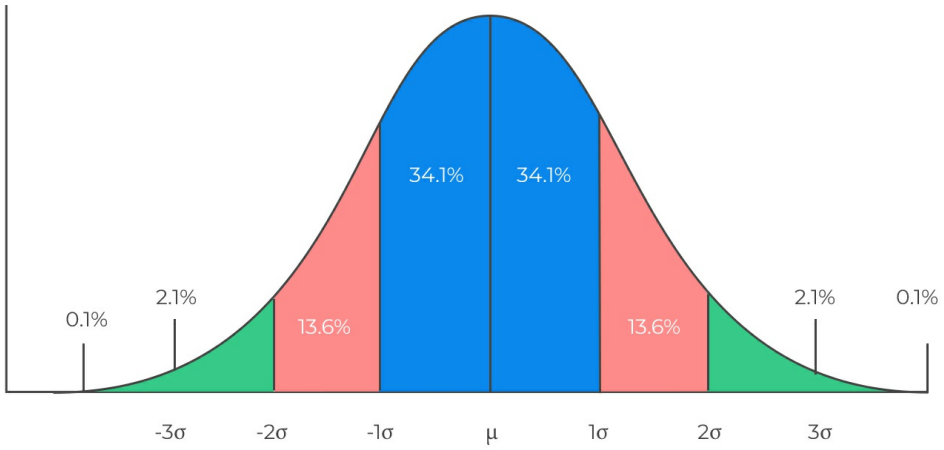
$$P(X > 2) = 1 - P(X \leq 2) = 1 - \frac{2-1}{\sqrt{2}} = 0.2929 \approx 0.29$$

We look up this value from the z-table.

$$\phi(0.29) \approx 61.41\%$$

x-value	z-value
μ	0
$\mu + 1\sigma$	1
$\mu + 2\sigma$	2
$\mu + n\sigma$	n

68 - 95 - 99.7 Rule



Recall that for a binomial random variable, if $np \geq 10$ and $n(1-p) \geq 10$, then the binomial distribution is normally distributed as:

$$X \sim N(np, np(1-p))$$

Also, Poisson distribution is normally approximated as $\lambda \geq 1000$ so that:

$$X \sim N(\lambda, \lambda)$$

We then calculate the probabilities while maintaining the normal distribution principles. The normal distribution is very popular as compared to other distributions because:

- Many discrete and continuous random variables distributions can be approximated using the normal distribution.
- The normal distribution is widely used in Central Limit Theorem (CLT), which is utilized in hypothesis testing.
- The normal distribution is closely related to other important distributions, such as the chi-squared and the F distributions.
- The notable property of the normal random variables is that they are infinitely divisible, which makes the normal distribution suitable for modeling asset prices.
- The normal distributions are closed under linear operations. In other words, the weighted sum of the normal random variables is also normally distributed.

Lognormal Distribution

A variable X is said to be lognormally distributed if the variable Y is normally distributed such that:

$$Y = \ln X$$

This also can be treated as:

$$X = e^{Y}$$

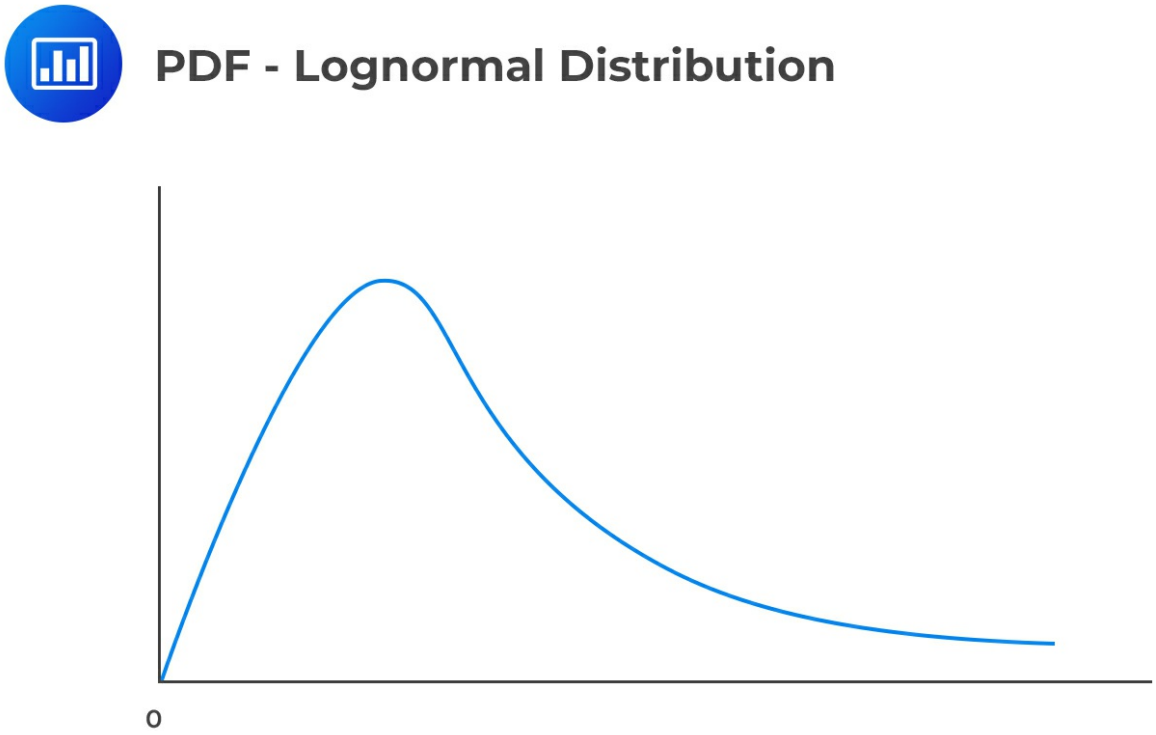
Where

$$Y \sim N(\mu, \sigma^2)$$

Since $Y \sim N(\mu, \sigma^2)$,then the PDF of a log-normal random variable is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}, x \geq 0$$

A variable is said to have a lognormal distribution if its natural logarithm has a normal distribution. The lognormal distribution is undefined for negative values, unlike the normal distribution that has a range of values between negative infinity and positive infinity.



If the above equation of the density function of the lognormal distribution is rearranged, we obtain an equation that has a similar form to the normal distribution. That is:

$$f(x) = e^{\frac{1}{2}\sigma^2-\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x-(\mu-\sigma^2)}{\sigma}\right)^2}$$

From the above, we notice that the lognormal distribution happens to be asymmetrical. It's not symmetrical around the mean as is the case under the normal distribution. The lognormal distribution peaks at $\exp(\mu - \sigma^2)$.

The following is the formula for the mean:

$$E[X] = e^{\mu+\frac{1}{2}\sigma^2}$$

This yields to an expression that closely resembles the Taylor expansion of the natural logarithm around 1. Recall that:

$$r \approx R - \frac{1}{2}R^2$$

where R is a standard return and r is the corresponding log return.

The following is the formula for the variance of the lognormal distribution:

$$V(X) = E[(X - E[X]^2)] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

Example: Lognormal Distribution

Consider a lognormal distribution given by $X \sim LogN(0.08, 0.2)$. Calculate the expected value.

Solution

For the lognormal distribution, the expected value is given by:

$$E[X] = e^{\mu+\frac{1}{2}\sigma^2} = e^{0.08+\frac{1}{2}\times 0.2} = 1.19721$$

Chi-Squared Distribution, χ^2

Assume we've got *k* independent standard normal variables ranging from Z_1 to Z_k. The sum of their squares will then have a Chi-Square distribution, written as follows:

$$S = \sum_{1=1}^{k} Z_i^2$$

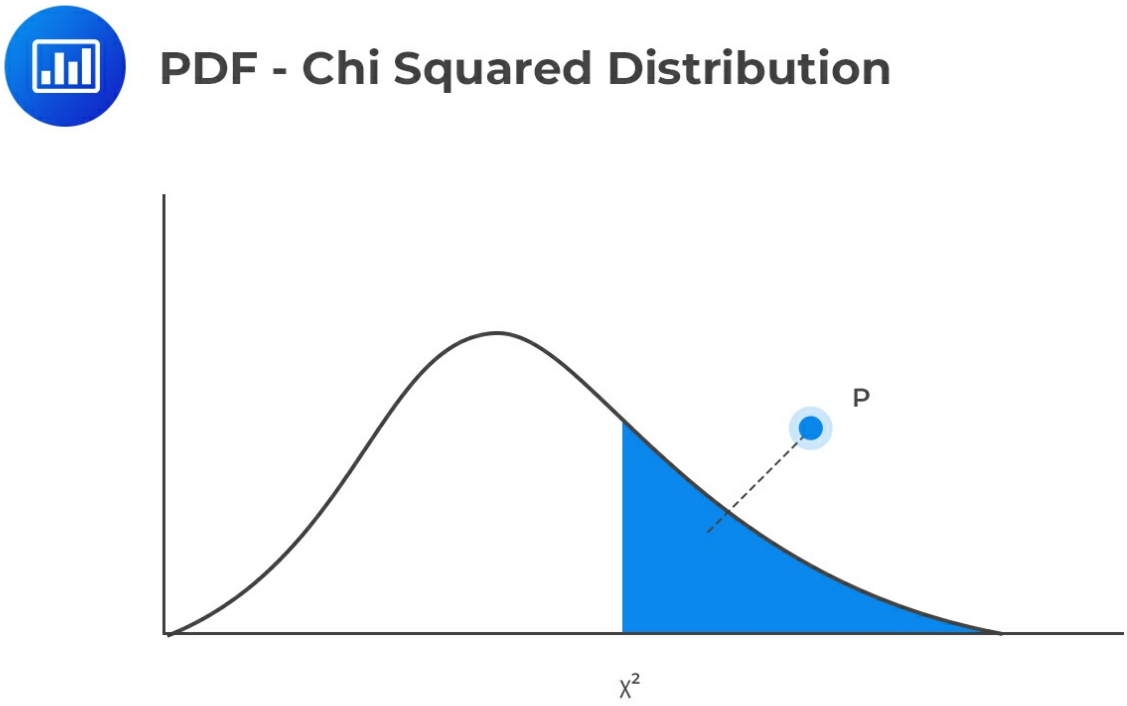
So, we can denote chi-distribution as:

$$S \sim X_k^2$$

k is called the degree of freedom. It is important to note that two chi-squared variables that are

independent, with degrees of freedom as k_1 and k_2, respectively, have a sum that is chi-square distributed with $(k_1 + k_2)$ degrees of freedom.

The chi-squared variable is usually asymmetrical and takes on non-negative values only. The distribution has a mean of k and a standard deviation of 2k.



The distribution has a mean and variance given by:

$$E(S) = k$$

and

$$V(S) = 2k$$

The chi-squared distribution takes the following PDF, for positive values of x:

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

The gamma function, Γ, is such that:

$$\Gamma(n) = \int_0^{\infty} x^{n-1} e^{-x} dx$$

Note also that the gamma function,Γis such that:

$$\Gamma(n) = (n-1)!$$

For instance:

$$\Gamma(3) = (3-1)! = 2 \times 1 = 2$$

This distribution is widely applicable in statistics and risk management when testing hypotheses. The chi-distribution is approximated using normal distribution when n is large. This implies that:

$$\chi_k^2 \sim N(k, 2k)$$

This is true because as the number of degrees of freedom increases, the skewness reduces. Degrees of freedom measure the amount of data required to test model parameters. If we have a sample size n, the degrees of freedom are given by n – p, where p is the number of parameters estimated..

Student's t Distribution

This distribution is often called the t distribution. Let Z be the standard normal variable, and U a chi-square variable with k degrees of freedom. Also, assume that U is independent of Z. Then, a random variable X that follows a t distribution is such that:

$$X = \frac{Z}{\sqrt{\frac{U}{k}}}$$

The following formula represents its PDF:

$$f(x) = \frac{\Gamma(k + \frac{1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}(1 + x^2/k)^{-\frac{k+1}{2}}$$

The mean of the t distribution is usually zero, and the distribution is symmetrical around it.

That is:

$$E(X) = 0$$

The variance is given by:

$$V(X) = \frac{k}{k-2}$$

The kurtosis is also given by:

$$Kurt(X) = 3\frac{k-2}{k-4}$$

It is easy to see that the mean is valid for $k > 1$ and the variances finite for $v > 2$. The kurtosis is only definite if $k > 4$ and should always be higher than 3.

The distribution converges to a standard normal distribution as k tends towards infinity ($k \to \infty$). When $k > 2$, the variance of the distribution becomes: $\frac{k}{(k-2)}$, and it converges to one as k increases.

We can also separate the degrees of freedom from variance to get what we called the **standardized student's t.** Using the formula:

$$V(aX) = a^2V(X))$$

Using this result, it is easy to see that :

$$V\left[\sqrt{\frac{v-2}{v}}Y\right] = 1$$

Where

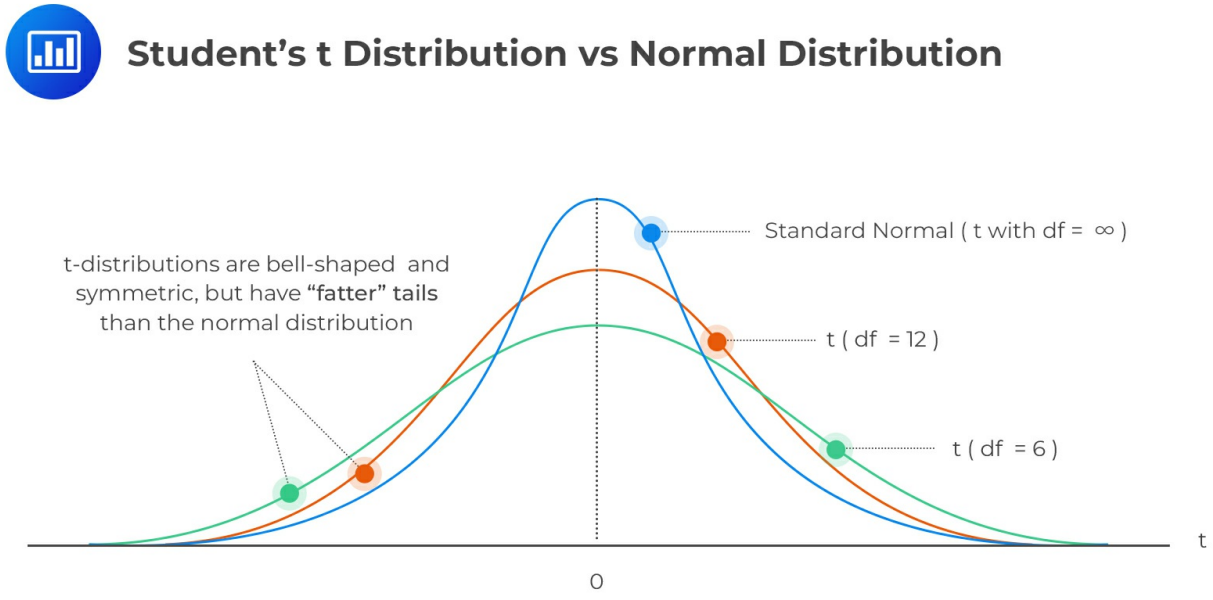
$$X \sim t_k$$

The generalized student's t is called standardized student's t because it has a mean of 0 and a variance

of 1. Note that we still rescale it to have any variance for k>2.

A generalized student's t is stated by the mean, variance, and the number of degrees of freedom. It is stated as Gen. $t_k(\mu, \sigma^2)$

This distribution is widely applicable in hypotheses testing, and modeling the returns of financial assets due to the excess kurtosis it displays.



Example: Standardized Student's t

The kurtosis of some returns on a bond portfolio with three parameters to be estimated is 6. What are the degrees of freedom if the parameters were generated using student's t_k?

Solution

We know that for t-distribution:

$$\text{Kurt}(X) = 3\frac{k-2}{k-4}$$

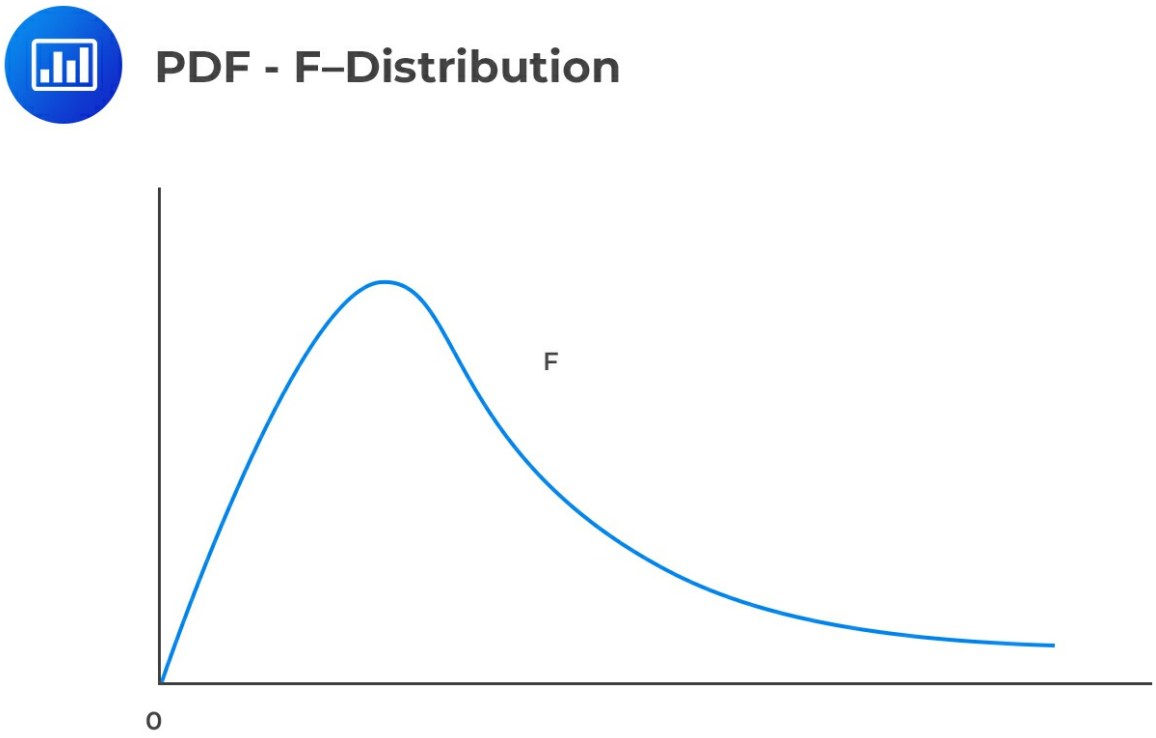
$$\therefore 6 = 3\frac{k-2}{k-4} \Rightarrow \frac{5}{3}(k-4)$$

So that

$$k = 6$$

F–Distribution

The F-distribution is often used in the analysis of variance (ANOVA). The F distribution is an asymmetric distribution that has a minimum value of 0, but no maximum value. Notably, the curve approaches but never quite touches the horizontal axis.



X is said to follow an F-distribution with parameters k_1 and k_2 if:

$$X = \frac{U_1/k_1}{U_2/k_2} \sim F(k_1, k_2)$$

Provided that U_1 and U_2 are chi-squared distributions that are independent having k_1 and k_2 as their degrees of freedom.

The F-distribution has the following PDF:

$$f(x) = \frac{\sqrt{\frac{(k_1 X)^{k_1} k_2^{k_2}}{(k_1 X + k_2)^{k_1 + k_2}}}}{x B\left(\frac{k_1}{2}, \frac{k_2}{2}\right)}$$

B(x,y) is a beta function such that:

$$B(x,y) = \int_0^1 z^{x-1}(1-z)^{y-1} dz$$

The distribution has the following mean and variance respectively:

$$E(X) = \frac{k_2}{k_2 - 2} \text{for} \quad k_2 > 2$$

$$\sigma^2 = \frac{2k_2^2 (k_1 + k_2 - 2)}{k_1 (k_2 - 2)^2 (k_2 - 4)} \text{for} \quad k_2 > 4$$

Suppose that X is a random variable with a t-distribution, and it has k degrees of freedom, then X^2 is said to have an F-distribution with 1 and k degrees of freedom, i.e.,

$$\chi^2 \sim F(1,k)$$

The Beta Distribution

The beta distribution applies to continuous random variables in the range of 0 and 1. This distribution is similar to the triangle distribution in the sense that they are both applicable in the modelling of default rates and recovery rates. Assuming that a and b are two positive constants, then the PDF of the beta distribution is written as:

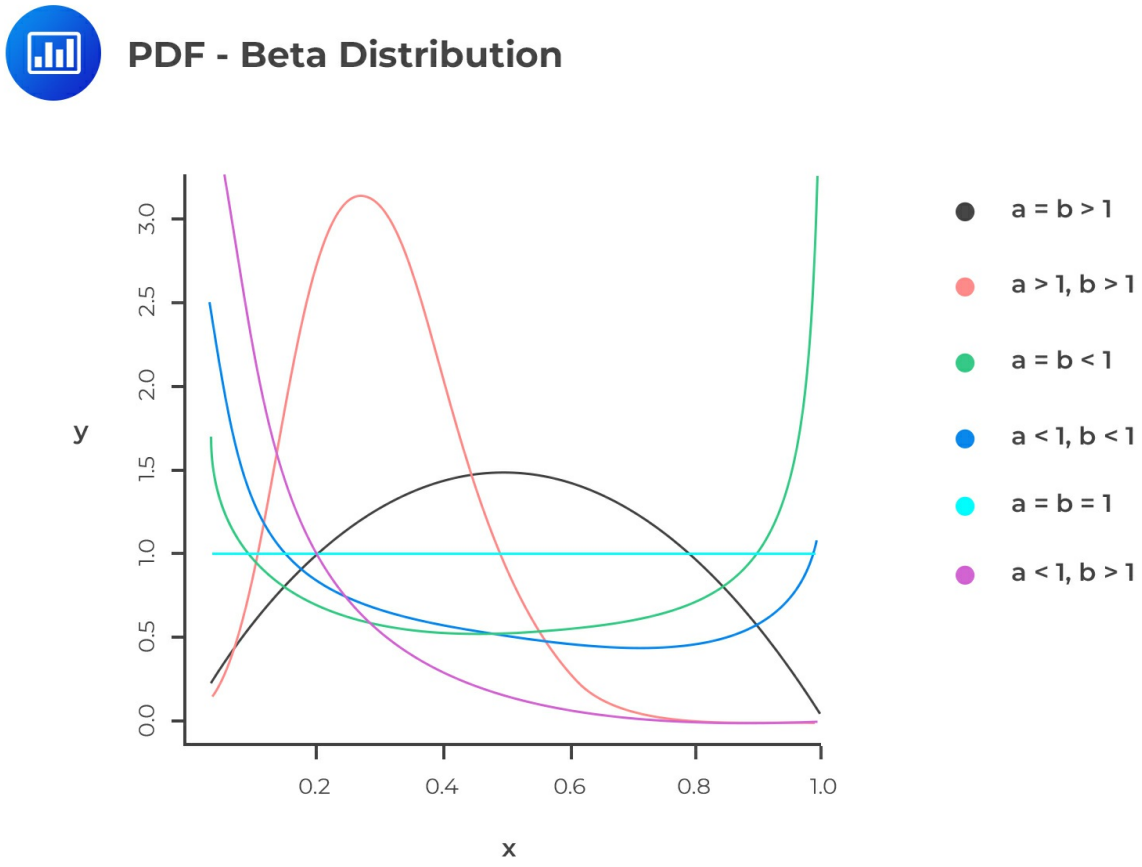
$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \qquad 0 \le x \le 1$$

Where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

The following two equations represent the mean and variance of the beta distribution:

$$\mu = \frac{a}{a+b}$$

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$



Exponential Distribution

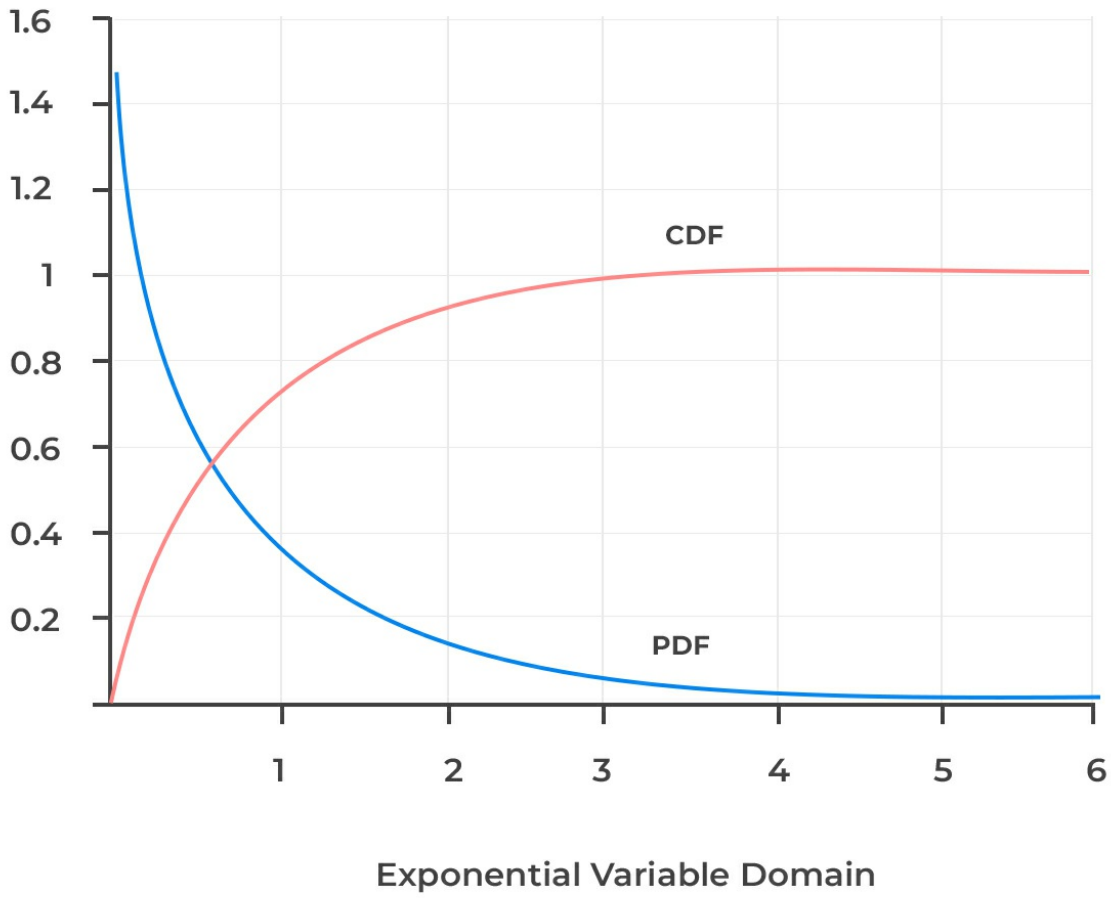
The exponential distribution is a continuous distribution with a parameter , whose PDF is:

$$f_X(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, x \geq 0$$

The CDF is also given by:

$$F_X(x) = 1 - e^{-\frac{x}{\beta}}$$

Exponential Distribution



The parameter of the exponential distribution determines the mean and variance of the distribution. That is:

$$E(X) = \beta$$

And

$$V(X) = \beta^2$$

Notably, exponential distribution is a close 'cousins' of a Poisson. The time intervals between one and subsequent Poisson random variables are exponentially distributed. Another feature of the exponential distribution is that it is memoryless. That is, its distributions are independent of their histories.

Example: Exponential Distribution

Assume that the time to default for a specific segment of mortgage consumers is exponentially distributed with a β of ten years. What is the probability that a borrower will not default before year 11?

Solution

To find the probability that the borrower will not default before year eleven, we start by calculating the cumulative distribution until year eleven and then subtract this from 100%::

$$P(X > 11) = 1 - P(X \leq 11) = 1 - F_X(x = 11)$$

$$= 1 - e^{-\frac{11}{10}} = 1 - 0.3329 = 0.6671 = 66.7\%$$

The Mixture Distribution

Mixture distributions are complex, and new distributions built using two or more distributions. In this summary, we shall concentrate on the two distributions.

Generally, a mixture distribution comes from a weighted average distribution of density functions, and can be written as follows:

$$f(x) = \sum_{j=1}^{n} w_i f_i(x) \quad such \quad that : \sum_{i=1}^{n} w_i = 1$$

$f_i(x)$'s are the component distributions, with $w_i^{'}s$ as the weights or the mixing proportions. The component weights must all sum up to one, for the resulting mixture to be legitimately distributed. In other words, a two-distribution combination must draw value from Bernoulli random variables and

depending on the benefits (0 or 1), it then picks the component distributions. By doing this, it is possible to compute the CDF of the mixture when the component distributions are normal random variables. These distributions are very flexible as they fall between parametric and non-parametric distributions.

For example, consider $X_1 \sim F_{x_1}$ and $X_2 \sim F_{x_2}$ and $W_i \sim \text{Bernoulli}(p)$. So that the mixture distribution of X_1 and X_2 is given by:

$$Y = pX_1 + (1-p)X_2$$

Both of the PDF and the CDF of the mixture distribution are weighted average of the constituent CDFs and PDFs. That is:

$$F_Y(y) = pF_{X_1}(x_1) + (1-p)F_{X_2}(x_2)$$

And

$$f_Y(y) = pf_{X_1}(x_1) + (1-p)f_{X_2}(x_2)$$

Intuitively, the computation of the central moment is done in a similar way. That is:

$$E(Y) = pE(X_1) + (1-p)E(X_2)$$

And

$$V(Y) = E(Y^2) - (E(Y))^2$$

Where

$$E(Y^2) = pE(X_1^2) + (1-p)E(X_2^2)$$

Using the same logic, we can calculate the other higher central moments such as the kurtosis and skewness. However, note that the mixture distribution might have both the skewness and the kurtosis, while the components do not have (for example, normal random variables).

Moreover, mixing components with different means and variances leads to distribution that is both

skewed and heavy-tailed.

Example: Mixture Distributions

Consider two normal random variables $X_1 \sim N(0.15, 0.60)$ and $X_1 \sim N(-0.8, 3)$. . What is the mean of the resulting mixture distribution (Y) if the weight of X_1 is 0.6?

Solution

We know that:

$$\begin{aligned} E(Y) &= pE(X_1) + (1-p)E(X_2) \\ &= 0.6 \times 0.15 + (1-0.6)(-0.8) \\ &= -0.23 \end{aligned}$$

Question

The number of new clients that a wealth management company receives in a month is distributed as a Poisson random variable with mean 2. Calculate the probability that the company receives exactly 28 clients in a year.

A. 5.48%

B. 0.10%

C. 3.54%

D. 10.2%

The correct answer is **A**.

The number of clients in a year (2×12) has a Poi(24) distribution.

$$P[X = n] = \frac{\lambda^n}{n!} e^{-\lambda}$$

$$P[X = 28] = \frac{24^{28}}{28!} e^{-24} = 5.48$$

Reading 15: Multivariate Random Variables

After completing this reading, you should be able to:

- Explain how a probability matrix can be used to express a probability mass function (PMF).
- Compute the marginal and conditional distributions of a discrete bivariate random variable.
- Explain how the expectation of a function is computed for a bivariate discrete random variable.
- Define covariance and explain what it measures.
- Explain the relationship between the covariance and correlation of two random variables and how these are related to the independence of the two variables.
- Explain the effects of applying linear transformations on the covariance and correlation between two random variables.
- Compute the variance of a weighted sum of two random variables.
- Compute the conditional expectation of a component of a bivariate random variable.
- Describe the features of an iid sequence of random variables.
- Explain how the iid property is helpful in computing the mean and variance of a sum of iid random variables.

Multivariate Random Variables

Multivariate random variables accommodate the dependence between two or more random variables. The concepts under multivariate random variables (such as expectations and moments) are analogous to those under univariate random variables.

Multivariate Discrete Random Variables

Multivariate random variables involve defining several random variables simultaneously on a sample space. In other words, multivariate random variables are vectors of random variables. For instance, a bivariate random variable X can be a vector with two components X_1 and X_2 with the corresponding realizations being x_1 and x_2, respectively.

The PMF or PDF for a bivariate random variable gives the probability that the two random variables each take a certain value. If we wish to plot these functions, we would need three factors: X_1, X_2, and the PMF/PDF. This is also applicable to the CDF.

The Probability Mass Function (PMF)

The PMF of a bivariate random variable is a function that gives the probability that the components of X=x takes the values $X_1 = x_1$ and $X_2 = x_2$. That is:

$$f_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

The PMF explains the probability of realization as a function of x_1 and x_2. The PMF has the following properties:

1. $f_{X_1,X_2}(x_1, x_2) \geq 0$
2. $\sum_{x_1} \sum_{x_2} f_{X_1,X_2}(x_1, x_2) = 1$

Example: Trinomial Distribution

The trinomial distribution is the distribution of n independent trials where each trial results in one of the three outcomes (a generalization of the binomial distribution). The first, second and the third components are X_1, X_2 and $n - X_1 - X_2$ respectively. However, the third component is redundant provided that we know X_1 and X_2.

The trinomial distribution has three parameters:

1. n, representing the total number of the trials
2. p_1, representing the probability of realizing X_1
3. p_2, representing the probability of realizing X_2

Intuitively, the probability of observing $n - X_1 - X_2$ is:

$$1 - p_1 - p_2$$

The PMF of the trinomial distribution, therefore, is given by:

$$f_{X_1,X_2}(x_1, x_2) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}$$

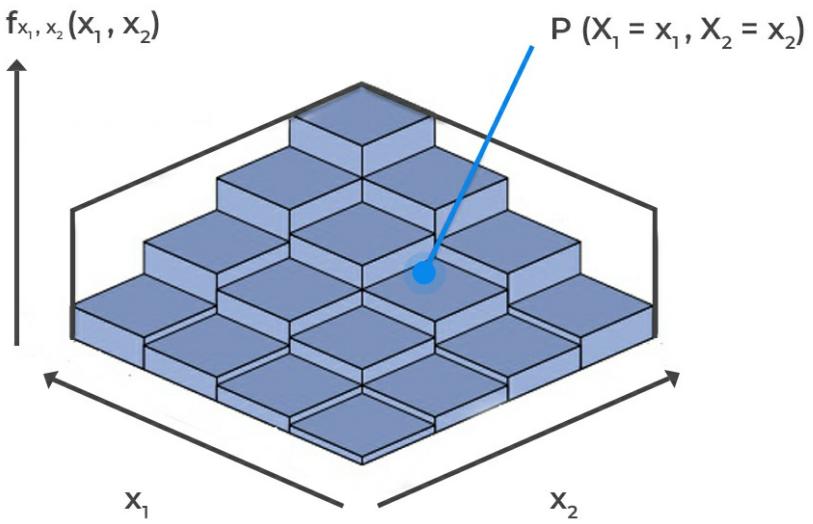
The Cumulative Distribution Function (CDF)

The CDF of a bivariate discrete random variable returns the total probability that each component is less than or equal to a given value. It is given by:

$$F_{X_1,X_2}(x_1, x_2) = P(X_1 < x_1, X_2 < x_2) = \sum_{\substack{t_1 \epsilon R(X_1) \\ t_1 \leq x_1}} \sum_{\substack{t_2 \epsilon R(X_2) \\ t_2 \leq x_2}} f_{(X_1,X_2)}(t_1, t_2)$$

In this equation, t_1 contains the values that X_1 may take as long as $t_1 \leq x_1$. Similarly, t_2 contains the values that X_2 may take as long as $t_2 \leq x_2$

CDF of a Bivariate Discrete Random Variable



Probability Matrices

The probability matrix is a tabular representation of the PMF.

Example: Probability Matrix

In financial markets, market sentiments play a role in determining the return earned on a security. Suppose the return earned on a bond is in part determined by the rating given to the bond by analysts. For simplicity, we are going to assume the following:

- There are only three possible returns :10%, 0%, or -10%
- Analyst ratings (sentiments) can be positive, neutral, or negative

We can represent this in a probability matrix as follows:

			Bond Return (X_1)		
			-10%	0%	10%
Analyst (X_2)	Positive	+1	5%	5%	30%
	Neutral	0	10%	10%	15%
	Negative	-1	20%	5%	0%

Each cell represents the probability of a joint outcome. For example, there's a 5% probability of a negative return (-10%) if analysts have positive views about the bond and its issuer. In other words, there's a 5% probability that the bond will decline in price with a positive rating. Similarly, there's a 10% chance that the bond's price will not change (and hence a zero return) given a neutral rating.

The Marginal Distribution

The marginal distribution gives the distribution of a single variable in a joint distribution. In the case of bivariate distribution, the marginal PMF of X_1 is computed by summing up the probabilities for X_1 across all the values in the support of X_2. The resulting PMF of X_1 is denoted by $f_{X_1}(x_1)$, i.e., the marginal distribution of X_1.

$$f_{X_1}(x_1) = \sum_{x_2 \in R(X_2)} f_{X_1,X_2}(x_1, x_2)$$

Intuitively, the PMF of X_2 is given by:

$$f_{X_2}(x_2) = \sum_{x_1 \in R(X_1)} f_{X_1,X_2}(x_1, x_2)$$

Example: Computing the Marginal Distribution

Using the probability matrix, we created above, we can come up with marginal distributions for both X_1 (return) and X_2 (analyst ratings) as follows:

For X_1,

$$P(X_1 = -10\%) = 5\% + 20\% + 10\% = 35\%$$
$$P(X_1 = 0\%) = 5\% + 10\% + 5\% = 20\%$$
$$P(X_1 = +10\%) = 30\% + 15\% + 0\% = 45\%$$

For X_2,

$$P(X_2 = +1) = 5\% + 5\% + 30\% = 40\%$$
$$P(X_2 = 0) = 10\% + 10\% + 15\% = 35\%$$
$$P(X_2 = -1) = 20\% + 5\% + 0\% = 25\%$$

We wish to compute the marginal distribution of the returns. Now,

In summary, for example, the marginal distribution of X_1 is given below:

$\text{Return}(X_1)$	-10%	0%	10%
$P(X_1 = x_1)$	35%	20%	35%

			Bond Return (X_1)			
			-10%	0%	10%	$f_{X_2}(x_2)$
Analyst (X_2)	Positive	+1	5%	5%	30%	40%
	Neutral	0	10%	10%	15%	35%
	Negative	−1	20%	5%	0%	25%
	$f_{X_1}(x_1)$		35%	20%	45%	

As you may have noticed, the marginal distribution satisfies the property of the ideal probability distribution. That is:

$$\sum_{\forall X_1} f_{X_1}(x_1) = 1$$

And

$$f_{X_1}(x_1) \geq 0$$

This is true because the marginal PMF is a univariate distribution.

We can, in addition, use the marginal PMF to compute the marginal CDF. The marginal CDF is such that, $P(X_1 < x_1)$. That is:

$$F_{X_1}(x_1) = \sum_{\substack{t_1 \in R(X_1) \\ t_1 \leq x_1}} f_{X_1}(t_1)$$

Independence of Random Variables

Recall that if the two events A and B are independent then:

$$P(A \cap B) = P(A)P(B)$$

This principle applies to bivariate random variables as well. If the distributions of the components of the bivariate distribution are independent, then:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

Example: Independence of Random Variables

Now let's use our earlier example on the return earned on a bond. If we assume that the two variables – return and ratings – are independent, we can calculate the joint distribution by the multiplying their marginal distributions. But are they really independent? Let's find out! We have already established the joint and the marginal distributions, as reproduced in the following table.

			Bond	Return	(X_1)	
			−10%	0%	10%	$f_{X_2}(x_2)$
Analyst	Positive	+1	5%	5%	30%	40%
(X_2)	Neutral	0	10%	10%	15%	35%
	Negative	−1	20%	5%	0%	25%
	$f_{X_1}(x_1)$		35%	20%	45%	

So assuming that our two variables are independent, our joint distribution would be as follows:

			Bond	Return	(X_1)
			−10%	0%	10%
Analyst	Positive	+1	14%	8%	18%
(X_2)	Neutral	0	12.25%	7%	15.75%
	Negative	−1	8.75%	5%	11.25%

We obtain the table above by multiplying the marginal PMF of the bond return by the marginal PMF of ratings. For example, the marginal probability that the bond return is 10% is 45% -- the sum of the third column. The marginal probability of a positive rating is 40% -- the sum of the first row. These

two values when multiplied give us the joint probability on the upper left end of the table (18%).

$$45\% * 40\% = 18\%$$

It is clear that the two variables are **not** independent because multiplying their marginal PMFs does not lead us back to the joint PMF.

The Conditional Distributions

The conditional distributions describe the probability of an outcome of a random variable conditioned on the other random variable taking a particular value.

Recall that, given any two events A and B, then:

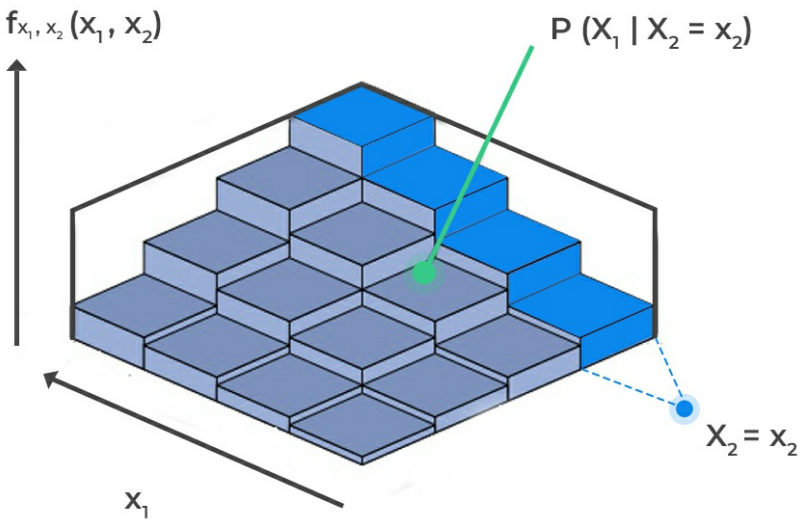
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This result can be applied in bivariate distributions. That is, the conditional distribution of X_1 given X_2 is defined as:

$$f_{X_1|X_2}(x_1|X_2 = x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

From the result above, the conditional distribution is joint distribution divided by the marginal distribution of the conditioning variable.

Conditional Distributions (Discrete)



Example: Calculating the Conditional Distribution

			Bond -10%	Return 0%	(X_1) 10%	$f_{X_2}(x_2)$
Analyst (X_2)	Positive	$+1$	5%	5%	30%	40%
	Neutral	0	10%	10%	15%	35%
	Negative	-1	20%	5%	0%	25%
	$f_{X_1}(x_1)$		35%	20%	45%	

Suppose we want to find the distribution of bond returns conditional on a positive analyst rating. The conditional distribution is:

$$f_{(X_1 \mid X_2)}(x_1 \mid X_2 = 1) = \frac{f_{X_1,X_2}(x_1, X_2 = 1)}{f_{X_2}(x_2 = 1)} = \frac{f_{X_1,X_2}(x_1, X_2 = 1)}{40\%}$$

With this, we can proceed to determine specific conditional probabilities:

$\text{Returns}(X_1)$	-10%	0%	10%
$f_{(X_1 \mid X_2)}(x_1 \mid X_2 = x_2) = P(X_1 = x_1 \mid X_2 = 1)$	$\frac{5\%}{40\%} = 12.5\%$	$\frac{5\%}{40\%} = 12.5\%$	$\frac{30\%}{40\%} = 75\%$

What we have done is to take the joint probabilities where there's a positive analyst rating and then divided these values by the marginal probability of a positive rating (40%) to produce the conditional distribution.

Note that the conditional PMF obeys the laws of probability, i.e.,

1. $f_{(X_1 \mid X_2)}(x_1 \mid X_2 = x_2) \geq 0 (\text{nonnegativity})$
2. $\sum_{\forall (X_1 \mid X_2)} f_{(X_1 \mid X_2)}(x_1 \mid X_2 = x_2) = 1$

Conditional Distribution for a Set of Outcomes

Conditional distributions can be computed for one variable, while conditioning on more than one variable.

For example, assume that we need to compute the conditional distribution of the bond returns given that analyst ratings are non-negative. Therefore, our conditioning set is {+1,0}:

$$X_2 \in \{+1, 0\}$$

The conditional PMF must sum across all outcomes in the set that is conditioned on S {+1,0}:

$$f_{(X_1 \mid X_2)}(x_1 \mid x_2 \epsilon S) = \frac{\sum_{X_2 \epsilon C} f_{(X_1, X_2)}(x_1, x_2)}{\sum_{X_2 \epsilon C} f_{(X_2)}(x_2)}$$

The marginal probability that $X_2 \in \{+1, 0\}$ is the sum of the marginal probabilities of these two outcomes:

$$f_{x_2}(+1) + f_{x_2}(0) = 75\%$$

			Bond Return (X_1)			
			-10%	0%	10%	$f_{X_2}(x_2)$
Analyst (X_2)	Positive	$+1$	5%	5%	30%	40%
	Neutral	0	10%	10%	15%	35%
	Negative	-1	20%	5%	0%	25%
	$f_{X_1}(x_1)$		35%	20%	45%	

Thus, the conditional distribution is given by:

$$f_{(X_1|X_2)}(x_1|x_2 \epsilon \{+1, 0\}) = \begin{cases} \frac{5\%+10\%}{75\%} = 20\% \\ \frac{5\%+10\%}{75\%} = 20\% \\ \frac{30\%+15\%}{75\%} = 60\% \end{cases}$$

Independence and Conditional Distribution of Random Variables

Recall that the conditional distribution is given by:

$$f_{(X_1|X_2)}(x_1|X_2 = x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

This can be rewritten into:

$$f_{(X_1,X_2)}(x_1, x_2) = f_{(X_1|X_2)}(x_1|X_2 = x_2) f_{X_2}(x_2)$$

Or

$$f_{(X_1,X_2)}(x_1, x_2) = f_{X_2|X_1}(x_2|X_1 = x_1) f_{X_1}(x_1)$$

Also, if the distributions of the components of the bivariate distributions are independent, then:

$$f_{(X_1,X_2)}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

If we substitute this in the above results we get:

$$f_{X_1}(x_1) f_{X_2}(x_2) = f_{(X_1|X_2)}(x_1|X_2 = x_2) f_{X_2}(x_2)$$
$$\Rightarrow f_{X_1}(x_1) = f_{(X_1|X_2)}(x_1|X_2 = x_2)$$

Applying again to

$$f_{X_1,X_2}(x_1,x_2) = f_{(X_2|X_1)}(x_2|X_1 = x_1)f_{X_1}(x_1)$$

we get:

$$f_{X_2}(x_2) = f_{(X_2|X_1)}(x_2|X_1 = x_1)$$

Expectations

The expectation of a function of a bivariate random variable is defined in the same way as that of the univariate random variable. Consider the function $g(X_1,X_2)$. The expectation is defined as:

$$E(g(X_1,X_2)) = \sum_{x_1 \epsilon R(X_1)} \sum_{x_2 \epsilon R(X_2)} g(x_1,x_2)f_{X_1,X_2}(x_1,x_2)$$

$g(x_1,x_2$ depends on both x_1 and $x_2)$ and it may be a function of one component only. Just like the univariate random variable,

$$E(g(X_1,X_2)) \neq g(E(X_1),E(X_2))$$

for a nonlinear function $g(x_1,x_2)$.

Example: Calculating the Expectation

Consider the following probability mass function:

		X_1	
		1	2
X_2	3	10%	15%
	4	70%	5%

Given that $g(x_1,x_2) = x_1^{x_2}$, calculate $E(g(x_1,x_2))$

Solution

Using the formula:

$$E(g(X_1, X_2)) = \sum_{x_1 \epsilon R(X_1)} \sum_{x_2 \epsilon R(X_2)} g(x_1, x_2) f_{X_1, X_2}(x_1, x_2)$$

In this case we need:

$$\begin{aligned} E(g(X_1, X_2)) &= \sum_{x_1 \epsilon \{1,2\}} \sum_{x_2 \epsilon \{3,4\}} g(x_1, x_2) f_{X_1, X_2}(x_1, x_2) \\ &= 1^3(0.10) + 1^4(0.7) + 2^3(0.15) + 2^4(0.05) \\ &= 2.80 \end{aligned}$$

Moments

Just like the univariate random variables, we shall use the expectations to define the moments.

The first moment is defined as:

$$E(X) = [E(X_1), E(X_2)] = [\mu_1, \mu_2]$$

The second moment involves the covariance between the components of the bivariate distribution X_1 and X_2. The second moment is given by:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1 X_2)$$

The Covariance between X_1 and X_2 is defined as:

$$\begin{aligned} Cov(X_1, X_2) &= E[(X_1 - E[X_1])]E[(X_2 - E[X_2])] \\ &= E[X_1 X_2] - E[X_1]E[X_2] \end{aligned}$$

Note that $Cov(X_1, X_1) = Var(X_1)$ and that if X_1 and X_2 are independent then $E[X_1 X_2] - E[X_1]E[X_2] = 0$ and thus:

$$\begin{aligned} Cov(X_1, X_2) &= E[X_1 X_2] - E[X_1]E[X_2] \\ &= E[X_1]E[X_2] - E[X_1]E[X_2] = 0 \end{aligned}$$

Most of the correlation between X_1 and X_2 is reported. Now let $Var(X_1) = \sigma_1^2, Var(X_2) = \sigma_2^2$ and

$Cov(X_1, X_2) = \sigma_{12}$ then the correlation is defined as:

$$Corr(X_1, X_2) = \rho_{X_1X_2} = \frac{Cov(X_1, X_2)}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

Therefore, we can write this in terms of covariance. That is:

$$\sigma_{12} = \rho_{X_1X_2}\sigma_1\sigma_2$$

Correlation gives the measure of the strength of the linear relationship between the two random variables, and it is always between -1 and 1. That is $-1 < Corr(X_1, X_1) < 1$

For instance, if $X_2 = \alpha + \beta X_1$ then:

$$Cov(X_1, X_2) = Cov(X_1, \alpha + \beta X_1) = \beta Var(X_1)$$

But we know that $Var(\alpha + \beta X_1) = \beta^2 Var(X_1)$. So,

$$Corr(X_1, X_2) = \rho_{X_1X_2} = \frac{\beta Var(X_1)}{\sqrt{Var(X_1)}\sqrt{\beta^2 Var(X_1)}} = \frac{\beta}{|\beta|}$$

it is now evident that if $\beta > 0$, then $\rho_{X_1X_2} = 1$ and when $\beta \leq 0$ then $\rho_{X_1X_2} = 0$

Similarly, if we consider two scaled random variables $a + bX_1$ and $c + dX_2$

Then,

$$Cov(a + bX_1, c + dX_2) = bdCov(X_1, X_2)$$

This implies that the scale factor in each random variable multiactivity affects the covariance. Using the above results, the corresponding correlation coefficient is given by:

$$Corr(a + bX_1, c + dX_2) = \frac{bdCov(X_1, X_2)}{\sqrt{a^2 Var(X_1)}\sqrt{b^2 Var(X_2)}} = \frac{ab}{|ab|}\frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)}\sqrt{Var(X_2)}}$$
$$= \frac{ab}{|ab|}\rho_{X_1X_2}$$

Application of Correlation: Portfolio Variance and Hedging

The variance of the underlying securities and their respective correlations are the necessary ingredients if the variance of a portfolio of securities is to be determined. Assuming that we have two securities whose random returns are X_A and X_B and their means are μ_A and μ_B with standard deviations of σ_A and σ_B. Then, the variance of X_A plus X_B can be computed as follows:

$$\sigma^2_{A+B} = \sigma^2_A + \sigma^2_B + 2\rho_{AB}\sigma_A\sigma_B$$

If X_A and X_B have a correlation of ρ_{AB} between them,

The equation changes to:

$$\sigma^2_{A+B} = 2\sigma^2(1 + \rho_{AB}),$$

Where:

$$\sigma^2_A = \sigma^2_B = \sigma^2$$

if both securities have an equal variance. If the correlation between the two securities is zero, then the equation can be simplified further. We have the following relation for the standard deviation:

$$\rho_{AB} = 0 \Rightarrow \sigma_{A+B} = \sqrt{2\sigma}$$

For any number of variables, we have that:

$$Y = \sum_{i=1}^{n} X_i = 1nX_i$$
$$\sigma^2_Y = \sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{ij}\sigma_i\sigma_j$$

In case all the X_i's are uncorrelated and all variances are equal to σ, then we have:

$$\sigma_Y = \sqrt{n\sigma} \text{ if } \rho_{ij} = 0 \quad \forall \quad i \neq j$$

This is what is called the square root rule for the addition of uncorrelated variables.

Suppose that Y, X_A, and X_B are such that:

$$Y = aX_A + bX_B$$

Therefore, with our standard notation, we have that:

$$\sigma_Y^2 = a^2\sigma_A^2 + b^2\sigma_B^2 + 2ab\rho_{AB}\sigma_A\sigma_B \ldots\ldots\ldots\ldots \text{Eq 1}$$

The major challenge during hedging is a correlation. Suppose we are provided with $1 of a security A. We are to hedge it with $h of another security B. A random variable p will be introduced to our hedged portfolio. h is, therefore, the hedge ratio. The variance of the hedged portfolio can easily be computed by applying Eq1:

$$P = X_A + hX_B$$
$$\sigma_P^2 = \sigma_A^2 + h^2\sigma_B^2 + 2h\rho_{AB}\sigma_A\sigma_B$$

The minimum variance of a hedge ratio can be determined by determining the derivative with respect to h of the portfolio variance equation and then equate it to zero:

$$\frac{d\sigma^2 P}{dh} = 2h\sigma_B^2 + 2\rho_{AB}\sigma_A\sigma_B = 0$$
$$\Rightarrow h^* = -\rho_{AB}\frac{\sigma_A}{\sigma_B}$$

To determine the minimum variance achievable, we substitute h* to our original equation:

$$\min[\sigma_P^2] = \sigma_A^2(1 - \rho_{AB}^2)$$

The Covariance Matrix

The covariance matrix is a 2x2 matrix that displays the covariance between the components of X. For instance, the covariance matrix of X is given by:

$$\text{Cov}(X) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

The Variance of Sums of Random Variables

The variance of the sum of two random variables is given by:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1X_2)$$

If the random variables are independent, then $Cov(X_1X_2) = 0$ and thus:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$$

In case of weighted random variables, the variance is given by:

$$Var(aX_1 + bX_2) = a^2Var(X_1) + b^2Var(X_2) + 2abCov(X_1X_2)$$

Conditional Expectation

A conditional expectation is simply the mean calculated after a set of prior conditions has happened. It is the value that a random variable takes "on average" over an arbitrarily large number of occurrences – given the occurrence of a certain set of "conditions." A conditional expectation uses the same expression as any other expectation and is a weighted average where the probabilities are determined by a conditional PMF.

For a discrete random variable, the conditional expectation is given by:

$$E(X_1 \mid X_2 = x_2) = \sum_i x_{1i} f(X_1 \mid X_2 = x_2)$$

Example: Calculating the Conditional Expectation

In the bond return/rating example, we may wish to calculate the expected return on the bond given a positive analyst rating, i.e., $E(X_1 \mid X_2 = 1)$

If you recall, the conditional distribution is as follows:

$Returns(X_1)$	-10%	0%	10%
$f_{(X_1 \mid X_2)}(x_1 \mid X_2 = 1) = P(X_1 = x_1 \mid X_2 = 1)$	$\frac{5\%}{40\%} = 12.5\%$	$\frac{5\%}{40\%} = 12.5\%$	$\frac{30\%}{40\%} = 75\%$

The conditional expectation of the return is determined as follows:

$$E(X_1|X_2 = 1) = -0.10 \times 0.125 + 0 \times 0.125 + 0.10 \times 0.75 = 0.0625 = 6.25\%$$

Conditional Variance

We can calculate the conditional variance by substituting the expectation in the variance formula with the conditional expectation.

We know that:

$$Var(X_1) = E[(X_1 - E(X_1))^2] = E(X_1)^2 - [E(X)]^2$$

Now the conditional variance of X_1 conditional on X_2 is given by:

$$Var(X_1|X_2 = x_2) = E(X_1^2|X_2 = x_2) - [E(X_1|X_2 = x_2)]^2$$

Returning to our example above, the conditional variance $Var(X_1|X_2 = 1)$ is given by:

$$Var(X_1|X_2 = 1) = E(X_1^2|X_2 = 1) - [E(X_1|X_2 = 1)]^2$$

Now,

$$E(X_1|X_2 = 1) = 0.0625$$

We need to calculate:

$$E(X_1^2|X_2 = 1) = (-0.10)^2 \times 0.125 + 0^2 \times 0.125 + 0.10^2 \times 0.75 = 0.00875$$

So that

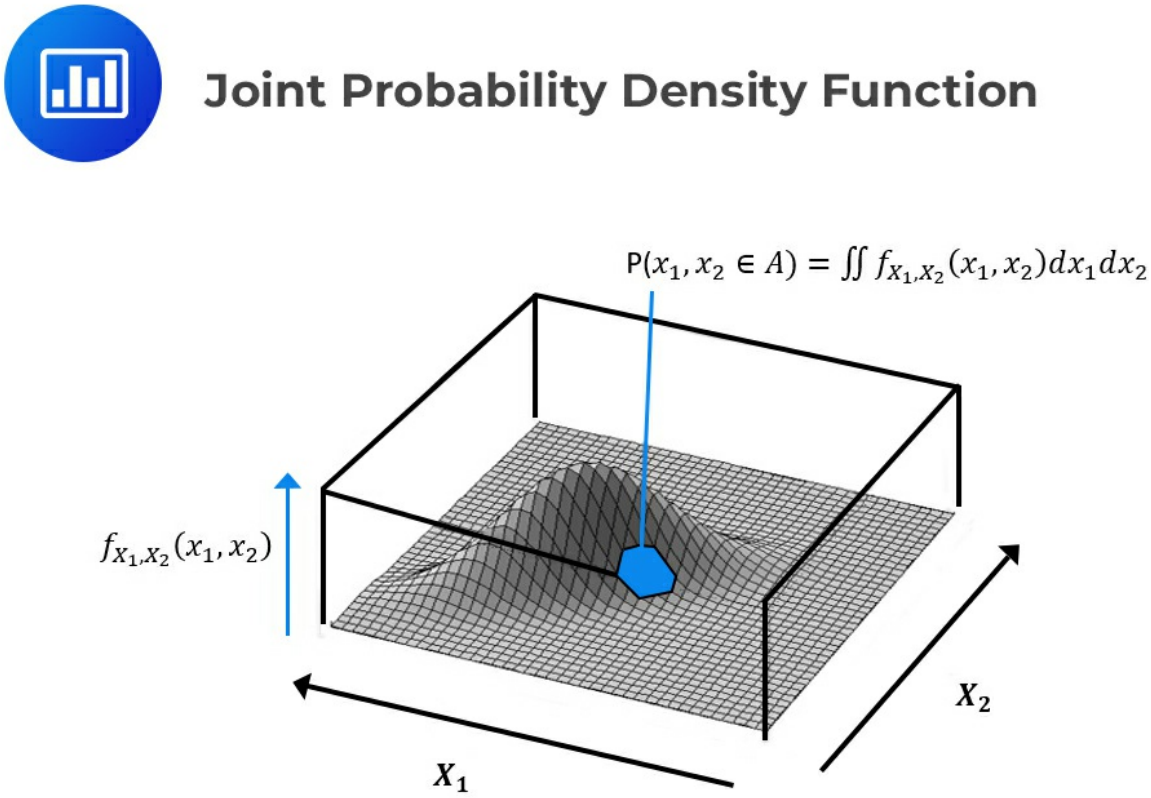
$$Var(X_1|X_2 = 1) = \sigma^2_{(X_1|X_2=1)} = 0.00875 - [0.0625]^2 = 0.004844 = 0.484\%$$

If we wish to find the standard deviation of the returns, we just find the square root of the variance:

$$\sigma_{(X_1|X_2=1)} = \sqrt{0.004844} = 0.06960 = 6.96\%$$

Continuous Random Variables

Before we continue, it is essential to note that continuous random variables make use of the same concepts and methodologies as discrete random variables. The main distinguishing factor is that instead of PMFs, continuous random variables use PDFs.



The Joint PDF

The joint (bivariate) distribution function gives the probability that the pair (X_1, X_2) takes values in a stated region A. It is given by:

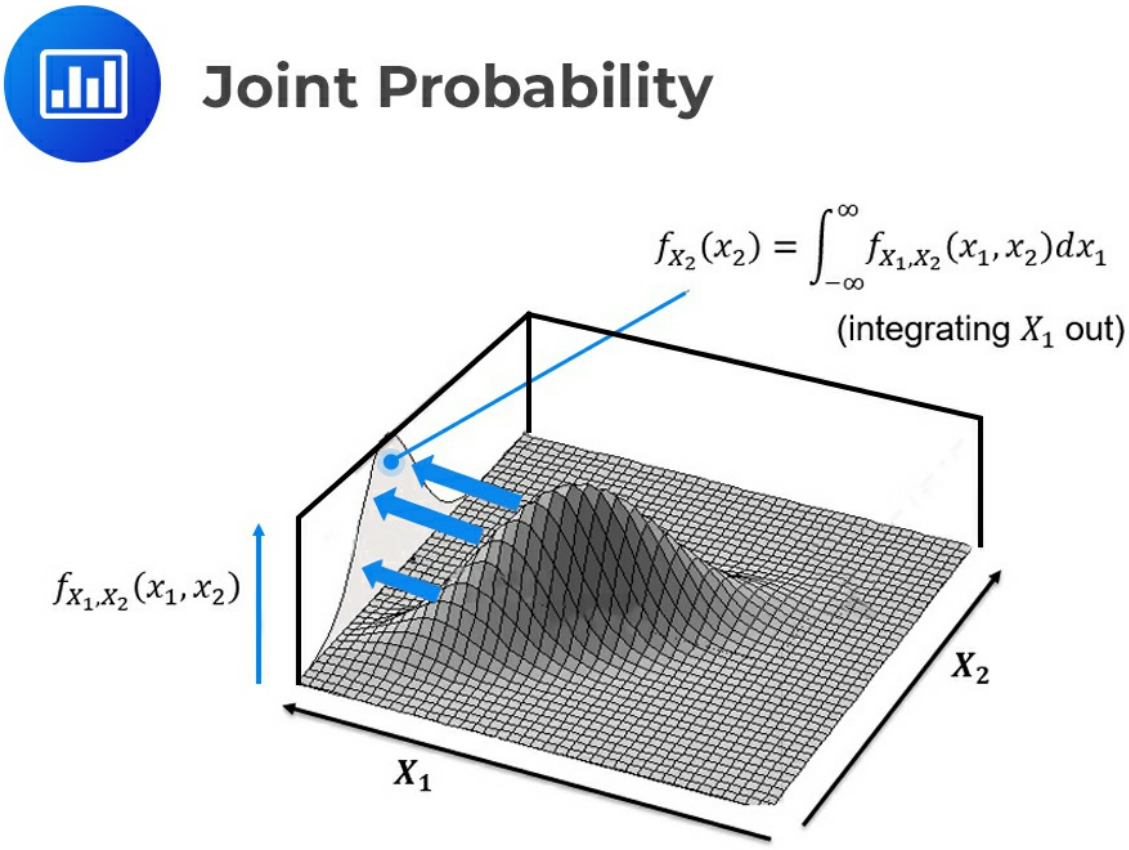
$$P(a < X_1 < b, c < X_2 < d) = \int_a^b \int_c^d f_{X_1,X_2}(x_1, x_2)dx_1dx_2$$

The joint pdf is always nonnegative, and the double integration yield a value of 1. That is:

$$f_{X_1,X_2}(x_1, x_2) \geq 0$$

And

$$\int_a^b \int_c^d f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 = 1$$



Example: Calculating the Joint Probability

Assume that the random variables (X_1) and (X_2) are jointly distributed as:

$$f_{X_1,X_2}(x_1, x_2) = k(x_1 + 3x_2) \quad 0 < x_1 < 2, 0 < x_2 < 2$$

Calculate the probability $P(X_1 < 1, X_2 > 1)$.

Solution

We need to first calculate the value of k.

Using the principle:

$$\int_a^b \int_c^d f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = 1$$

We have

$$\begin{aligned}\int_0^2 \int_0^2 k(x_1+3x_2)dx_1dx_2 &= \int_0^2 k[(\frac{1}{2}x_1^2+3x_1x_2)]_0^2 dx_2 = 1\\ &= \int_0^2 k(2+6x_2)dx_2 = k[2x_2+3x_2^2]_0^2 = 1\\ 16k = 1 &\Rightarrow k = \frac{1}{16}\end{aligned}$$

So,

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{16}(x_1+3x_2)$$

Therefore,

$$P(X_1<1, X_2>1) = \int_0^1 \int_1^2 \frac{1}{16}(x_1+3x_2)dx_1dx_2 = 0.3125$$

Joint Cumulative Distribution Function (CDF)

The joint cumulative distribution is given by:

$$F(X_1<x_1, X_2<x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1,X_2}(t_1,t_2)dt_1dt_2$$

Note that the lower bound of the integral can be adjusted so that it is the lower value of the interval.

Using the example above, we can calculate $F(X_1<1, X_2<1)$ in a similar way as above.

The Marginal Distributions

For the continuous random the marginal distribution is given by:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)dx_2$$

Similarly,

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)dx_1$$

Note that if we want to find the marginal distribution of X_1 we integrate X_2 out and vice versa.

Example: Computing the Marginal Distribution

Consider the example above. We have that

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{16}(x_1 + 3x_2) \quad 0 < x_1 < 2, 0 < x_2 < 2$$

We wish to find the marginal distribution of X_1. This implies that we need to integrate out X_2. So,

$$\begin{aligned} f_{X_1}(x_1) &= \int_0^2 \frac{1}{16}(x_1 + 3x_2)dx_2 = \frac{1}{16}[x_1x_2 + \frac{3}{2}x_2^2]_0^2 \\ &= \frac{1}{16}[2x_1 + 6] = \frac{1}{8}(x_1 + 3) \\ \Rightarrow f_{X_1}(x_1) &= \frac{1}{16}[2x_1 + 6] = \frac{1}{8}(x_1 + 3) \end{aligned}$$

Note that we can calculate $f_{X_2}(x_2)$ in a similar manner.

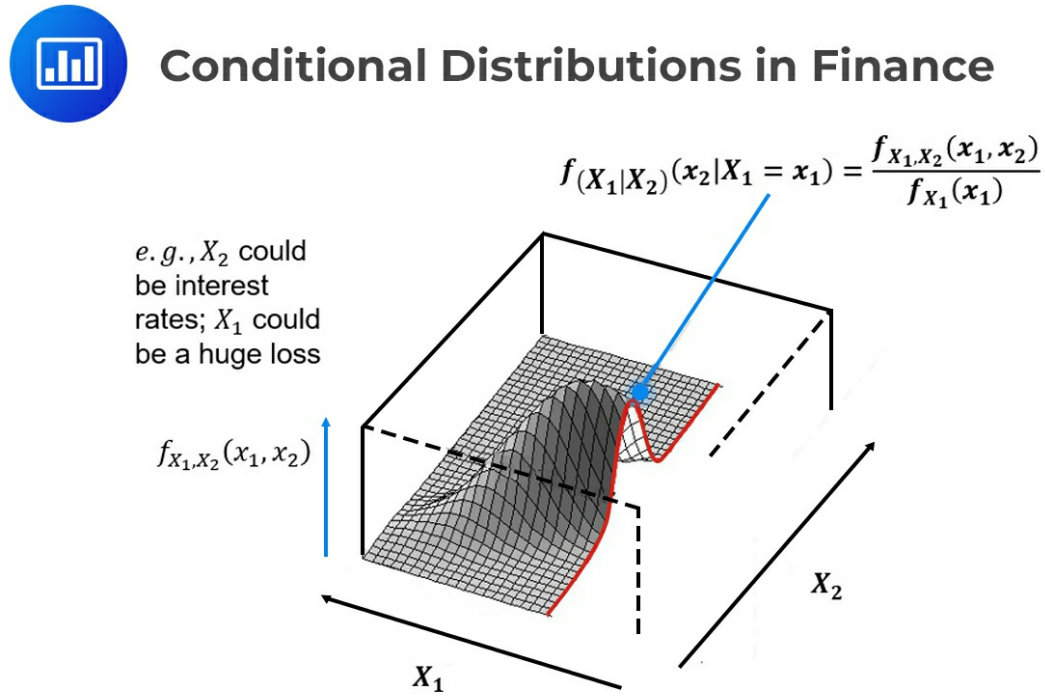
Conditional Distributions

The conditional distribution is analogously defined as that of discrete random variables. That is:

$$f_{(X_1|X_2)}(x_1|X_2 = x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

The conditional distributions are applied in the field of finance, such as risk management. For

instance, we may wish to compute the conditional distribution of interests rates, X_1 given that the investors X_2 experience a huge loss.



Independent, Identically Distributed (IID) Random Variables

A collection of **random variables** is independent and identically distributed (iid) if each **random variable** has the same probability distribution as the others and all are mutually independent.

Example:

Consider successive throws of a **fair** coin:

- The coin has **no memory**, so all the throws are **"independent"**.
- **The probability of head vs. tail in every throw is 50:50;** so the coin is equally likely and stays fair; the distribution from which every throw is drawn is normal and stays the same, and thus each outcome is **"identically distributed"**

iid variables are mostly applied in time series analysis.

Mean and Variance of iid Variables

Consider the iid variables generated by a normal distribution. They are typically defined as:

$$x_i^{iid} \sim N(\mu, \sigma^2)$$

The expected mean of these particular iid is given by:

$$E\left(\sum_i^n X_i\right) = \sum_i^n E(X_i) = \sum_i^n \mu = n\mu$$

Where $E(X_i) = \mu$

The result above is valid since the variables are independent and have similar moments. Maintaining this line of thought, the variance of iid random variables is given by:

$$\begin{aligned} Var\left(\sum_i^n X_i\right) &= \sum_i^n Var\left(X_i\right) + 2\sum_{j=1}^n \sum_{k=j+1}^n Cov(X_j, X_k) \\ &= \sum_i^n \sigma^2 + 2\sum_{j=1}^n \sum_{k=j+1}^n 0 = \sum_i^n \sigma^2 = n\sigma^2 \end{aligned}$$

The independence property is important because there's a difference between the variance of the sum of **multiple** random variables and the variance of a multiple of a single random variable. If X_1 and X_2 are iid with variance σ^2, then,

$$\begin{aligned} &Var(X_1 + X_2) = Var(X_1) + Var(X_2) = \sigma^2 + \sigma^2 = 2\sigma^2 \\ &Var(X_1 + X_2) \neq Var(2X_1) \end{aligned}$$

In the case of a multiple of a single variable, X_1, with variance σ^2,

$$Var(2X_1) = 4Var(X_1) = 4 \times \sigma^2 = 4\sigma^2$$

Practice Question

A company is reviewing fire damage claims under a comprehensive business insurance policy. Let X be the portion of a claim representing damage to inventory and let Y be the portion of the same application representing damage to the rest of the property. The joint density function of X and Y is:

$$f(x,y) = \begin{cases} 6[1-(x+y)], & x>0, y>0\ x+y<1 \\ 0, & \text{elsewhere} \end{cases}$$

What is the probability that the portion of a claim representing damage to the rest of the property is less than 0.3?

A. 0.657

B.0.450

C. 0.415

D. 0.752

The correct answer is A.

First, we should find the marginal PMF of Y:

$$f_Y(y) = \int_0^{1-y} 6[1-(x+y)]\,\partial x = [6(x - \frac{x^2}{2} - xy]_0^{1-y}$$

Substitute the limits as usual to get:

$$6\left[(1-y) - \frac{(1-y)^2}{2} - y(1-y)\right]$$

At this we can **factor out** $(1-y)$ and solve what remains in the square bracket:

$$6(1-y)[1-\frac{(1-y)}{2}-y]=6(1-y)[\frac{1-y}{2}]$$

Of course you can cancel 2 with 6 at this point:

$$6(1-y)[\frac{1-y}{2}]=3(1-y)[1-y]=3(1-2y+y^2)=3-6y+3y^2$$

So,

$$f_Y(y)=3-6y+3y^2, 0<y<1$$

We need $P(Y<0.3)$, So,

$$P(Y<0.3)=\int_0^{0.3}(3-6y+3y^2)dy=0.9-0.27+0.027=0.657$$

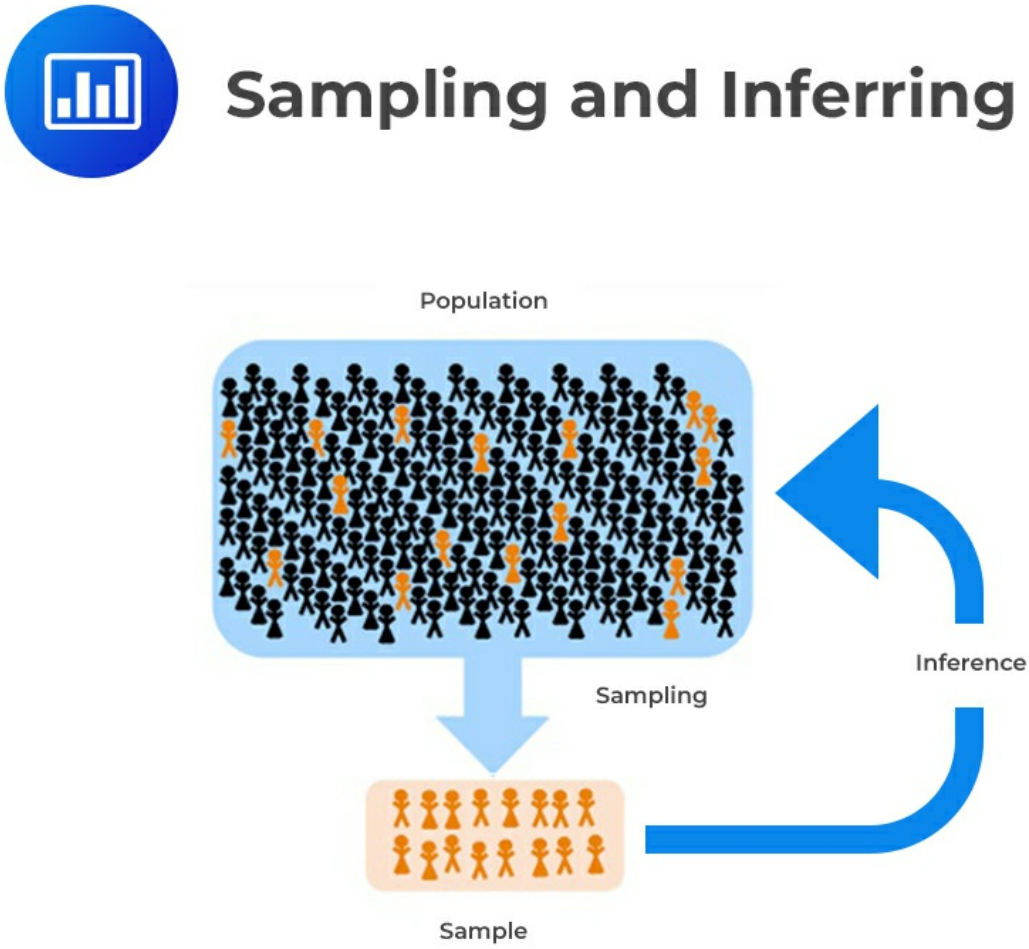
Reading 16: Sample Moments

After completing this reading, you should be able to:

- Estimate the mean, variance, and standard deviation using sample data.
- Explain the difference between a population moment and a sample moment.
- Distinguish between an estimator and an estimate.
- Describe the bias of an estimator and explain what the bias measures.
- Explain what is meant by the statement that the mean estimator is BLUE.
- Describe the consistency of an estimator and explain the usefulness of this concept.
- Explain how the Law of Large Numbers (LLN) and Central Limit Theorem (CLT) apply to the sample mean.
- Estimate and interpret the skewness and kurtosis of a random variable.
- Use sample data to estimate quantiles, including the median.
- Estimate the mean of two random variables and apply the CLT.
- Estimate the covariance and correlation between two random variables.
- Explain how coskewness and cokurtosis are related to skewness and kurtosis.

Sample Moments

Recall that moments are defined as the expected values that briefly describe the features of a distribution. Sample moments are those that are utilized to approximate the unknown population moments. Sample moments are calculated from the sample data.



Such moments include mean, variance, skewness, and kurtosis. We shall discuss each moment in detail.

Estimation of the Mean

The population mean, denoted by μ is estimated from the sample mean ($\bar{X}$). The estimated mean is denoted by $\hat{\mu}$ and defined by:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Where X_i is a random variable assumed to be independent and identically distributed so $E(X_i) = \mu$ and

n is the number of observations.

Note that the mean estimator is a function of random variables, and thus it is a random variable. Consequently, we can examine its properties as a random variable (its mean and variance)

For instance, the expectation of the mean estimator $\hat{\mu}$ is the population mean μ. This can be seen as follows:

$$E(\hat{\mu}) = E(\bar{X}) = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \times n\mu = \mu$$

The above result is true since we have assumed that X_i's are iid. The mean estimator is an unbiased estimator of the population mean.

The bias of an estimate is defined as:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Where $\hat{\theta}$ is the true estimator of the population value θ. So, in the case of the population mean:

$$Bias(\hat{\mu}) = E(\hat{\mu}) - \mu = \mu - \mu = 0$$

Since the value of the mean estimator is 0, it is an unbiased estimator of the population mean.

Using conventional features of a random variable, the variance of the mean estimator is calculated as:

$$Var(\hat{\mu}) = Var\,(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}[\sum_{i=1}^{n} Var(X_i) + Covariances]$$

But we are assuming that X_i's are iid, and thus they are uncorrelated, implying that their covariance is equal to 0. Consequently, taking $Var(X_i) = \sigma^2$, the above formula changes to:

$$Var(\hat{\mu}) = Var\,(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}[\sum_{i=1}^{n} Var(X_i)] = \frac{1}{n^2}[\sum_{i=1}^{n} \sigma^2] = \frac{1}{n}^2 \times n\sigma^2 = \frac{\sigma^2}{n}$$

Thus

$$Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

Looking at the last formula, the variance of the mean estimator depends on the data variance (σ^2) and the sample mean n. Consequently, the variance of the mean estimator decreases as the number of the observations (sample size) is increased. This implies that the larger the sample size, the closer the estimated mean to the population mean.

Example: Calculating the Sample Mean

An experiment was done to find out the number of hours that candidates spend preparing for the FRM part 1 exam. It was discovered that for a sample of **10 students**, the following times were spent:

$$318, 304, 317, 305, 309, 307, 316, 309, 315, 327$$

What is the sample mean?

Solution

We know that:

$$\bar{X} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\Rightarrow \bar{X} = \frac{318 + 304 + 317 + 305 + 309 + 307 + 316 + 309 + 315 + 327}{10}$$

$$= 312.7$$

Desirable Properties of the Sample Mean Estimator

- The mean estimator is averagely equal to the population mean.
- As the sample size (the number of the observation) increases, the sample mean tends to the population mean.
- The sample mean can be assumed to be distributed normally (normal distribution)

Estimation of Variance and Standard Deviation

The sample estimator of variance is defined as:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

Note that we are still assuming that X_i's are iid. As compared to the mean estimator, the sample estimator of variance is biased. It can be proved that:

$$\text{Bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{\sigma^2}{n}$$

This implies that the bias decreases as the number of observations are increased. Intuitively, the source of the bias is the variance of the mean estimator ($\frac{\sigma^2}{n}$). Since the bias is known, we construct an unbiased estimator of variance as:

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

It can be shown that $E(s^2) = \sigma^2$ and thus s^2 is an unbiased variance estimator. Maintaining this line of thought, it might seem s^2 is a better estimator of variance than $\hat{\sigma}^2$ but this is not necessarily true since the variance of $\hat{\sigma}^2$ is less than that of s^2. However, financial analysis involves large data sets, and thus either of these values can be used. However, when the number of observations is more than 30 ($n \geq 30$), $\hat{\sigma}^2$ is preferred conventionally.

The sample standard deviation is the square root of the sample variance. That is:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

or

$$s = \sqrt{s^2}$$

Note that the square root is a nonlinear function, and thus, the standard deviation estimators are biased but diminish as the sample size increases.

Example: Calculating the Sample Variance Estimator (Unbiased)

Using the example as in calculating the sample mean, what is the sample variance?

Solution

The sample estimator of variance is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

To make it easier, we will calculate we make the following table:

X_i	$(X_i - \hat{\mu})^2$
318	$(318-312.7)^2 = 28.09$
304	75.69
317	18.49
305	59.29
309	13.69
307	32.49
316	10.89
309	13.69
315	5.29
327	204.49
Total	434.01

So, the variance is given to be:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = \frac{434.01}{10-1} = 48.22$$

Reasons, why Mean and Standard Deviations are Used

I. The mean and the variance are almost adequate to describe data.
II. They give a clue on the range of the values that can be observed.
III. The units of the mean and the standard deviation are the same as those of the data, and thus they can be easily compared.

Skewness

As we saw in chapter two, the skewness is a cubed standardized central moment given by:

$$\text{skew}(X) = \frac{E([X - E(X)]^3}{\sigma^3} = E[(\frac{X - \mu}{\sigma})^3]$$

Note that $\frac{X-\mu}{\sigma}$ is a standardized X with a mean of 0 and variance of 1.

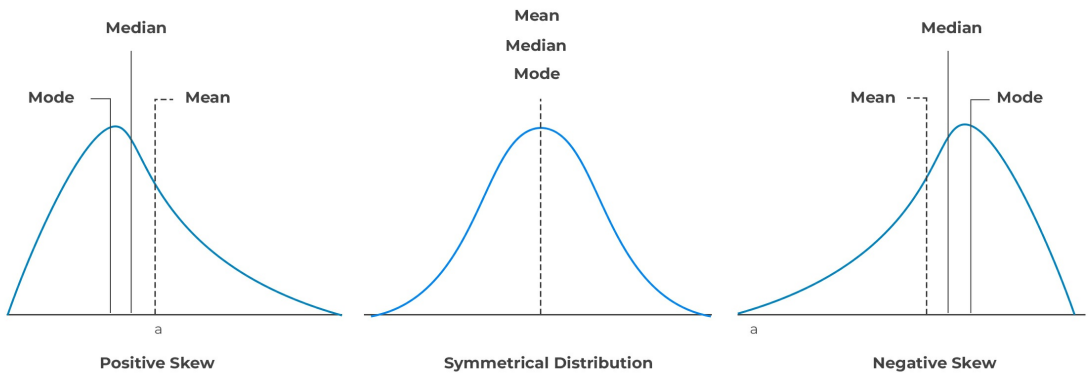
This can also be written as:

$$\text{skew}(X) = \frac{E([X - E(X)]^3}{E[(X - E(X))^2]^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3}$$

Where μ_3 is the third central moment, and σ is the standard deviation

The skewness measures the asymmetry of the distribution (since the third power depends on the sign of the difference). When the value of the asymmetry is negative, there is a high probability of observing the large magnitude of negative value than positive values (tail is on the left side of the distribution). Conversely, if the skewness is positive, there is a high probability of observing the large magnitude of positive values than negative values (tail is on the right side of the distribution).

Asymmetry of a Distribution



The estimators of the skewness utilize the principle of expectation and is denoted by:

$$\frac{\hat{\mu}^3}{\hat{\sigma}^3}$$

We can estimate $\hat{\mu}^3$ as:

$$\hat{\mu}^3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^3$$

Example: Calculating the Skewness

The following are the data on the financial analysis of a sales company's income over the last 100 months:

$n = 100, \sum_{i=1}^{n}(x_i - \hat{\mu})^2 = 674,759.90.$ and $\sum_{i=1}^{n}(x_i - \hat{\mu})^3 = -12,456.784$

Calculate that Skewness.

Solution

The skewness is given by:

$$\frac{\hat{\mu}^3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^3}{[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2]^{\frac{3}{2}}} = \frac{\frac{1}{100}(-12,456.784)}{[\frac{1}{100} \times 674,759.90]^{\frac{3}{2}}} = -0.000225$$

Kurtosis

The Kurtosis is defined as the fourth standardized moment given by:

$$\text{Kurt}(X) = \frac{E([X - E(X)])^4}{\sigma^4} = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right]$$

The above can be written as:

$$\text{Kurt}(X) = \frac{E([X - E(X)]^4}{E[(X - E(X))^2]^2} = \frac{\mu_4}{\sigma^4}$$

The description of kurtosis is analogous to that of the Skewness, only that the fourth power of the Kurtosis implies that it measures the absolute deviation of random variables. The reference value of a normally distributed random variable is 3. A random variable with Kurtosis exceeding 3 is termed to be **heavily or fat-tailed.**

The estimators of the skewness utilize the principle of expectation and is denoted by:

$$\frac{\hat{\mu}^4}{\hat{\sigma}^4}$$

We can estimate $\hat{\mu}^4$ (fourth central moment) as:

$$\hat{\mu}^4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^4$$

The BLUE Mean Estimator

We say that the mean estimator is the Best Linear Unbiased Estimator (BLUE) of the population mean when the data used are iid. That is,

I. The variance of the mean has the lowest variance of any Linear Unbiased Estimator (LUE).
II. It is the unbiased estimator of the population mean (as shown earlier)
III. It is a linear function of the data used.

The linear estimators are a function of the mean and can be defined as:

$$\hat{\mu} = \sum_{i=1}^{n}\omega_i X_i$$

Where ω_i is independent of X_i . In the case of the sample mean estimator, $\omega_i = \frac{1}{n}$. Recall that we had shown the unbiases of the sample mean estimator.

BLUE puts an estimator as the best by having the smallest variance among all linear and unbiased estimators. However, there are other superior estimators, such as Maximum Likelihood Estimators (MLE).

The Behavior of Mean in Large Sample Sizes

Recall that the mean estimator is unbiased, and its variance takes a simple form. Moreover, if the data used are iid and normally distributed, then the estimator is also normally distributed. However, it poses a great difficulty in defining the exact distribution of the mean in a finite number of observations.

To overcome this, we use the behavior of the mean in large sample sizes (that is as $n \to \infty$) to approximate the distribution of the mean infinite sample sizes. We shall explain the behavior of the mean estimator using the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT).

The Law of Large Numbers (LLN)

The law of large numbers (Kolmogorov Strong Law of Large Numbers) for iid states that if X_i's is a sequence of random variables, with $E(X_i) \equiv \mu$ then:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s} \mu$$

Put in words, the sample mean estimator $\hat{\mu}_n$ converges almost surely ($\xrightarrow{a.s}$) to population mean (μ).

An estimator is said to be consistent if LLN applies to it. Consistency requires that an estimator is:

I. Unbiased and that the bias should decrease as n increases.
II. The variance decreases as the number of observations n increases. That is:$Var(\hat{\mu}_n) \to 0$.

Moreover, under LLN, the sample variance is consistent. That is, LLN implies that $\hat{\sigma} \xrightarrow{2a.s} \sigma^2$

However, consistency is not easy to study because it tends to 0 as $n \to \infty$.

The Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) states that if $X_1, X_2, \ldots, X_n$ is a sequence of iid random variables with a finite mean μ and a finite non-zero variance σ^2, then the distribution of $\frac{\hat{\mu}-\mu}{\frac{\sigma}{\sqrt{n}}}$ tends to a standard

normal distribution as $n \to \infty$.

Put simply,

$$\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0, 1)$$

Note that $\hat{\mu} = \bar{X}$ = Sample Mean

Note that CLT extends LLN and provides a way of approximating the distribution of the sample mean estimator. CLT seems to be appropriate since it does not require the distribution of random variables used.

Since CLT is asymptotic, we can also use the unstandardized forms so that:

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

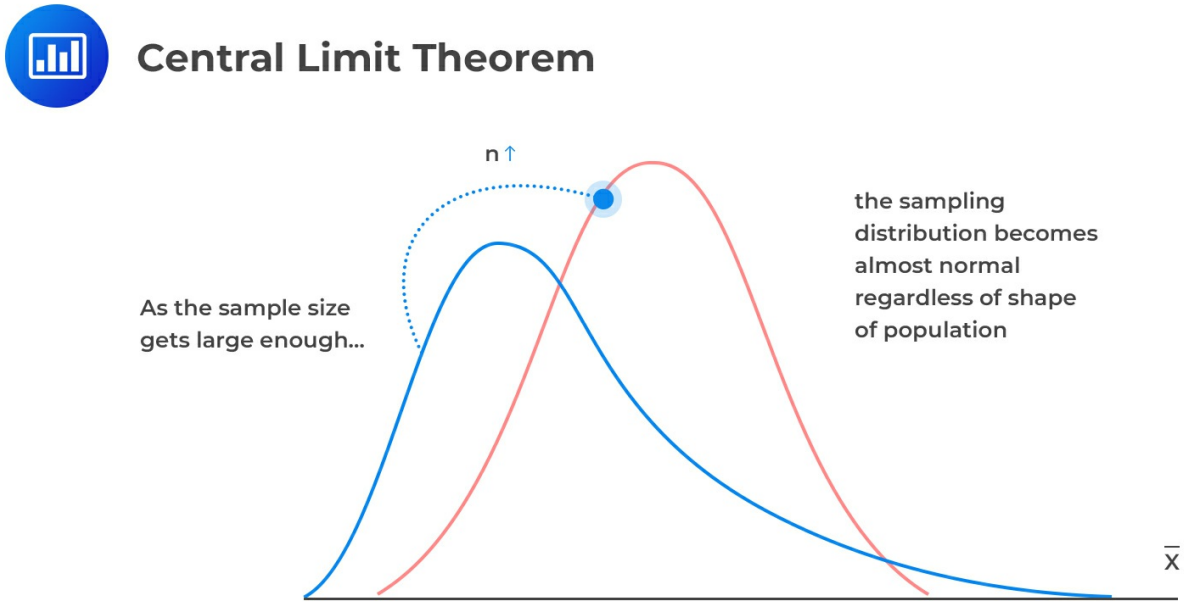
Note that we can go back to standard normal variable Z as:

$$Z = \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Which is actually the result we have initially.

The main question is, how large is n?

The value of n solely depends on the shape of the population (distribution of X_i's), i.e., the skewness. However, CLT is appropriate when $n \geq 30$



Example: Applying CLT

A sales expert believes that the number of sales per day for a particular company has a mean of 40 and a standard deviation of 12. He surveyed for over 50 working days. What is the probability that the sample mean of sales for this company is less than 35?

Solution

Using the information given in the question,

$\mu = 40, \sigma = 12$ and n=50

By central limit theorem,

$$\hat{\mu} \sim N(\mu, \frac{\sigma^2}{n})$$

We need

$$P(\hat{\mu} < 35) = P\left[Z < \frac{35 - 40}{\frac{12}{\sqrt{50}}}\right] = P(\hat{\mu} < -2.946)$$
$$= P(\hat{\mu} < -2.946) = 1 - P(\hat{\mu} < 2.946) = 0.00161$$

Estimation of Median and Other Quantiles

Median

Median is a central tendency measure of distribution, also called the 50% quantile, which divides the distribution in half (50% of observations lie on either side of the median value).

When the sample size is odd, the value in position (n + 1)/2 of the sorted list is used to estimate the median:

$$Med(x) = x_{\frac{n}{2}+1}$$

If the number of the observations is even, the median is estimated as the average of the two central points of the sorted list. That is:

$$Med(x) = \frac{1}{2}[x_{\frac{n}{2}} + x_{\frac{n}{2}+1}]$$

Example: Calculating the Median

The ages of experienced financial analysts in a country are:

$$56, 51, 43, 34, 25, 50.$$

What is the median age of the analysts?

Solution

We need to arrange the data in ascending order:

$$25, 34, 43, 50, 51, 56$$

The sample size is 6 (even), so the median is given by:

$$\text{Med(Age)} = \frac{1}{2}[x_{\frac{6}{2}} + x_{\frac{6}{2}+1}] = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(43 + 50) = 46.5$$

Properties of the Median

- It may not be an actual observation in the data set.
- It is not affected by extreme values because the median is a positional measure.
- It is used when the exact midpoint of the score distribution is desired, or when there are many outliers (extreme observations).

Other Quartiles

For other quantiles such as 25% and 75% quantiles, we estimate analogously as the median. For instance, a θ-quantile is determined using the nθ, which is a value in the sorted list. If nθ is not an integer, we will have to take the average below or above the value nθ.

So, in our example above, the 25% quantile (θ=0.25) is 6×0.25=1.5. This implies that we need to find the average value of the 1st and 2nd values:

$$\hat{q}_{25} = \frac{1}{2}(25 + 34) = 29.5$$

The Interquartile Range

The interquartile range (IQR) is defined as the difference between the 75% and 25% quartiles. That is:

$$(\hat{I}QR) = \hat{q}_{75} - \hat{q}_{25}$$

IQR is a measure of dispersion and thus can be used as an alternative to the standard deviation

If we use the example above, the 75% quantile is 6×0.75=4.5. So, we need to average the 4^{th} and 5^{th}

values:

$$\hat{q}_{75} = \frac{1}{2}(50 + 51) = 50.5$$

So that the IQR is:

$$(\hat{I}QR) = 50.5 - 29.5 = 21$$

Desirable Properties of Quantiles

I. The units of the quantiles are the same as those of the data used hence they are easy to interpret.
II. They are robust to outliers of the data. The median and the IQR are unaffected by the outliers.

The Multivariate Moments

We can extend the definition of moments from the univariate to multivariate random variables. The mean is unaffected by this because it is just the combination of the means of the two univariate sample means.

However, if we extend the variance, we would need to estimate the covariance between each pair plus the variance of each data set used. Moreover, we can also define Kurtosis and Skewness analogously to univariate random variables.

Covariance

In covariance, we focus on the relationship between the deviations of some two variables rather than the difference from the mean of one variable.

Recall that the covariance of two variables X and Y is given by:

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])]E[(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

The sample covariance estimator is analogous to this result. The sample covariance estimator is given by:

$$\hat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)$$

Where

$\hat{\mu}_X$-the sample mean of X

$\hat{\mu}_Y$- the sample mean of Y

The sample covariance estimator is biased towards zero, but we can remove the estimator by using n-1 instead of just n.

Correlation

Correlation measures the strength of the linear relationship between the two random variables, and it is always between -1 and 1. That is $-1 < \text{Corr}(X_1, X_1) < 1$.

Correlation is a standardized form of the covariance. It is approximated by dividing the sample covariance by the product of the sample standard deviation estimator of each random variable. It is defined as:

$$\rho_{XY} = \frac{\hat{\sigma}_{XY}}{\sqrt{\hat{\sigma}_X^2}\sqrt{\hat{\sigma}_Y^2}} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X\hat{\sigma}_Y}$$

Sample Mean of Two Variables

We estimate the mean of two random variables the same way we estimate that of a single variable. That is:

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n}(x_i)$$

And

$$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n}(y_i)$$

Assuming both of the random variables are iid, we can apply CLT in each estimator. However, if we consider the joint behavior (as a bivariate statistic), CLT stacks the two mean estimators into a 2x1 matrix:

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{bmatrix}$$

Which is normally distributed as long the random variable Z=[X, Y] is iid. The CLT on this vector depends on the covariance matrix:

$$\begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$$

Note that in a covariance matrix, one diagonal displays the variance of random variable series, and the other is covariances between the pair of the random variables. So, the CLT for bivariate iid data is given by:

$$\sqrt{n}\begin{bmatrix} \hat{\mu}_x - \mu_x \\ \hat{\mu}_y - \mu_y \end{bmatrix} \rightarrow N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right)$$

If we scale the difference between the vector of means, then the vector of means is normally distributed. That is:

$$\begin{bmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{bmatrix} \rightarrow N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \frac{\sigma_X^2}{n} & \frac{\sigma_{XY}}{n} \\ \frac{\sigma_{XY}}{n} & \frac{\sigma_Y^2}{n} \end{bmatrix}\right)$$

Example: Applying Bivariate CLT

The annualized estimates of the means, variances, covariance, and correlation for monthly return of

stock trade (T) and the government's bonds (G) for 350 months are as shown below:

Moment	$\hat{\mu}_T$	σ_T^2	$\hat{\mu}_G$	σ_G^2	σ_{TG}	ρ_{TG}
	11.9	335.6	6.80	26.7	14.0	0.1434

We need to compare the volatility, interpret the correlation coefficient, and apply bivariate CLT.

Solution

Looking at the output, it is evident that the return from the stock trade is more volatile than the government bond return since it has a higher variance. The correlation between the two forms of return is positive but very small.

If we apply bivariate CLT, then:

$$\sqrt{n}\begin{bmatrix}\hat{\mu}_x - \mu_x \\ \hat{\mu}_y - \mu_y\end{bmatrix} \rightarrow N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}335.6 & 14.0\\14.0 & 26.7\end{bmatrix}\right)$$

But the mean estimators have a limiting distribution (which is assumed to be normally distributed). So,

$$\begin{bmatrix}\hat{\mu}_x \\ \hat{\mu}_y\end{bmatrix} \rightarrow N\left(\begin{bmatrix}\mu_x\\\mu_y\end{bmatrix}, \begin{bmatrix}0.9589 & 0.04\\0.04 & 0.07629\end{bmatrix}\right)$$

Note the new covariance matrix is equivalent to the previous covariance divided by the sample size n=350.

In bivariate CLT, the correlation in the data is the correlation between the sample means and should be equal to the correlation between the data series.

Coskewness and Cokurtosis

Coskewness and Cokurtosis are an extension of the univariate skewness and kurtosis.

Coskewness

The two coskewness measures are defined as:

$$\text{Skew}(X,X,Y) = \frac{E[(X-E[X])^2(Y-E[Y])]}{\sigma_X^2\sigma_Y}$$

$$\text{Skew}(X,Y,Y) = \frac{E[(X-E[X])(Y-E[Y])^2]}{\sigma_X\sigma_Y^2}$$

These measures both capture the likelihood of the data taking a large directional value whenever the other variable is large in magnitude. When there is no sensitivity to the direction of one variable to the magnitude of the other, the two coskewnesses are 0. For example, the coskewness in a bivariate normal is always 0, even when the correlation is different from 0. Note that the univariate skewness estimators are s(X,X,X) and s{Y,Y,Y).

So how do we estimate coskewness?

The coskewness is estimated by using the estimation analogy. That is, replacing the expectation operator by summation. For instance, the two coskewness is given by:

$$\text{Skew}(X,X,Y) = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu}_X)^2(y_i-\hat{\mu}_Y)}{\hat{\sigma}_X^2\hat{\sigma}_Y}$$

$$\text{Skew}(X,Y,Y) = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu}_X)(y_i-\hat{\mu}_Y)^2}{\hat{\sigma}_X\hat{\sigma}_Y^2}$$

Cokurtosis

There intuitively three configurations of the cokurtosis. They are:

$$\text{Kurt}(X,X,Y,Y) = \frac{E[(X-E[X])^2(Y-E[Y])^2]}{\sigma_X^2\sigma_Y^2}$$

$$\text{Kurt}(X, X, X, Y) = \frac{E[(X - E[X])^3(Y - E[Y])]}{\sigma_X^3 \sigma_Y}$$

$$\text{Kurt}(X, Y, Y, Y) = \frac{E[(X - E[X])(Y - E[Y])^3]}{\sigma_X \sigma_Y^3}$$

The reference value of a normally distributed random variable is 3. A random variable with Kurtosis exceeding 3 is termed to be **heavily or fat-tailed**. However, comparing the cokurtosis to that of the normal is not easy since the cokurtosis of the bivariate normal depends on the correlation.

When the value of the cokurtosis is 1, then the random variables are uncorrelated and increases as the correlation devices from 0.

Practice Question

A sample of 100 monthly profits gave out the following data:

$\sum_{i=1}^{100} x_i = 3,353$ and $\sum_{i=1}^{100} x_i^2 = 844,536$

What is the sample mean and standard deviation of the monthly profits?

A. Sample Mean=33.53, Standard deviation=85.99

B. Sample Mean=53.53, Standard deviation=85.55

C. Sample Mean=43.53, Standard deviation=89.99

D. Sample Mean=33.63, Standard deviation=65.99

Solution

The correct answer is **A**.

Recall that the sample mean is given by:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\Rightarrow \bar{X} = \frac{1}{100} \times 3353 = 33.53$$

The variance is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

Note that,

$$(X_i - \hat{\mu})^2 = X_i^2 - 2X_i\hat{\mu} + \hat{\mu}^2$$

So that

$$\sum_{i=1}^{n}(X_i-\hat{\mu})^2=\sum_{i=1}^{n}X_i^2-2X_i\hat{\mu}+\hat{\mu}^2=\sum_{i=1}^{n}X_i^2-2\hat{\mu}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\hat{\mu}^2$$

Note again that

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i\Rightarrow\sum_{i=1}^{n}X_i=n\hat{\mu}$$

So,

$$\sum_{i=1}^{n}X_i^2-2\hat{\mu}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\hat{\mu}^2=\sum_{i=1}^{n}X_i^2-2\hat{\mu}.n\hat{\mu}+n\hat{\mu}$$
$$=\sum_{i=1}^{n}X_i^2-n\hat{\mu}^2$$

Thus:

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2=\frac{1}{n-1}\{\sum_{i=1}^{n}X_i^2-n\hat{\mu}^2\}$$

So, in our case:

$$s^2=\frac{1}{n-1}\{\sum_{i=1}^{n}X_i^2-n\hat{\mu}^2\}=\frac{1}{99}(844,536-100\times33.53^2)=7395.0496$$

So that the standard deviation is given to be:

$$s=\sqrt{7395.0496}=85.99$$

Reading 17: Hypothesis Testing

After completing this reading, you should be able to:

- Construct an appropriate null hypothesis and alternative hypothesis and distinguish between the two.
- Construct and apply confidence intervals for one-sided and two-sided hypothesis tests, and interpret the results of hypothesis tests with a specific level of confidence.
- Differentiate between a one-sided and a two-sided test and identify when to use each test.
- Explain the difference between Type I and Type II errors and how these relate to the size and power of a test.
- Understand how a hypothesis test and a confidence interval are related.
- Explain what the p-value of a hypothesis test measures.
- Interpret the results of hypothesis tests with a specific level of confidence.
- Identify the steps to test a hypothesis about the difference between two population means.
- Explain the problem of multiple testing and how it can bias results.

Hypothesis testing is defined as a process of determining whether a hypothesis is in line with the sample data. Hypothesis testing tries to test whether the observed data of the hypothesis is true. Hypothesis testing starts by stating the null hypothesis and the alternative hypothesis. The null hypothesis is an assumption of the population parameter. On the other hand, the alternative hypothesis states the parameter values (critical values) at which the null hypothesis is rejected. The critical values are determined by the distribution of the test statistic (when the null hypothesis is true) and the size of the test (which gives the size at which we reject the null hypothesis).

Components of the Hypothesis Testing

The elements of the test hypothesis include:

I. The null hypothesis.
II. The alternative hypothesis.
III. The test statistic.
IV. The size of the hypothesis test and errors
V. The critical value.
VI. The decision rule.

The Null hypothesis

As stated earlier, the first stage of the hypothesis test is the statement of the null hypothesis. The null hypothesis is the statement concerning the population parameter values. It brings out the notion that "there is nothing about the data."

The **null hypothesis**, denoted as H_0, represents the current state of knowledge about the population parameter that's the subject of the test. In other words, it represents the "status quo." For example, the U.S Food and Drug Administration may walk into a cooking oil manufacturing plant intending to confirm that each 1 kg oil package has, say, 0.15% cholesterol and not more. The inspectors will formulate a hypothesis like:

H_0: Each 1 kg package has 0.15% cholesterol.

A test would then be carried out to confirm or reject the null hypothesis.

Other typical statements of H_0 include:

$$H_0 : \mu = \mu_0$$

$$H_0 : \mu \leq \mu_0$$

Where:

μ = true population mean and,

μ_0 = the hypothesized population mean.

The Alternative Hypothesis

The **alternative hypothesis**, denoted H_1, is a **contradiction** of the null hypothesis. The null hypothesis determines the values of the population parameter at which the null hypothesis is rejected. Thus, rejecting the H_0 makes H_1 valid. We accept the alternative hypothesis when the "status quo" is discredited and found to be untrue.

Using our FDA example above, the alternative hypothesis would be:

H_1: Each 1 kg package does not have 0.15% cholesterol.

The typical statements of H_0 include:

$$H_0 : \mu \neq \mu_0$$

$$H_0 : \mu > \mu_0$$

Where:

μ = true population mean and,

μ_0 = the hypothesized population mean.

Note that we have stated the alternative hypothesis, which contradicted the above statement of the null hypothesis.

The Test Statistic

A test statistic is a standardized value computed from sample information when testing hypotheses. It compares the given data with what we would expect under the null hypothesis. Thus, it is a major determinant when deciding whether to reject H_0, the null hypothesis.

We use the test statistic to gauge the degree of agreement between sample data and the null hypothesis. Analysts use the following formula when calculating the test statistic.

$$\text{Test Statistic} = \frac{(\text{Sample Statistic} - \text{Hypothesized Value})}{(\text{Standard Error of the Sample Statistic})}$$

The test statistic is a random variable that changes from one sample to another. Test statistics assume a variety of distributions. We shall focus on normally distributed test statistics because it is used hypotheses concerning the means, regression coefficients, and other econometric models.

We shall consider the hypothesis test on the mean. Consider a null hypothesis $H_0 : \mu = \mu_0$. Assume that the data used is iid, and asymptotic normally distributed as:

$$\sqrt{n}(\hat{\mu} - \mu) \sim N(0, \sigma^2)$$

Where σ^2 is the variance of the sequence of the iid random variable used. The asymptotic distribution leads to the test statistic:

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \sim N(0, 1)$$

Note this is consistent with our initial definition of the test statistic.

The following table gives a brief outline of the various test statistics used regularly, based on the distribution that the data is assumed to follow:

Hypothesis Test	Test Statistic
Z-test	z-statistic
Chi-Square Test	Chi-Square statistic
t-test	t-statistic
ANOVA	F-statistic

We can subdivide the set of values that can be taken by the test statistic into two regions: One is called the non-rejection region, which is consistent with H_0 and the rejection region (critical region), which is inconsistent with H_0. If the test statistic has a value found within the critical region, we reject H_0.

Just like with any other statistic, the distribution of the test statistic must be specified entirely under H_0 when H_0 is true.

The Size of the Hypothesis Test and the Type I and Type II Errors

While using sample statistics to draw conclusions about the parameters of the population as a whole, there is always the possibility that the sample collected does not accurately represent the population. Consequently, statistical tests carried out using such sample data may yield incorrect results that may lead to erroneous rejection (or lack thereof) of the null hypothesis. We have two types of error:

Type I Error

Type I error occurs when we reject a true null hypothesis. For example, a type I error would manifest in the form of rejecting $H_0 = 0$ when it is actually zero.

Type II Error

Type II error occurs when we fail to reject a false null hypothesis. In such a scenario, the test provides insufficient evidence to reject the null hypothesis when it's false.

The level of significance denoted by α represents the probability of making a type I error, i.e., rejecting the null hypothesis when, in fact, it's true. α is the direct opposite of β, which is taken to be the probability of making a type II error within the bounds of statistical testing. The ideal but practically impossible statistical test would be one that **simultaneously** minimizes α and β. We use α to determine critical values that subdivide the distribution into the rejection and the non-rejection regions.

The Critical Value and the Decision Rule

The decision to reject or not to reject the null hypothesis is based on the distribution assumed by the test statistic. This means if the variable involved follows a normal distribution, we use the level of significance (α) of the test to come up with critical values that lie along with the standard normal distribution.

The decision rule is a result of combining the critical value (denoted by C_α), the alternative hypothesis, and the test statistic (T). The decision rule is to whether to reject the null hypothesis in favor of the alternative hypothesis or fail to reject the null hypothesis.

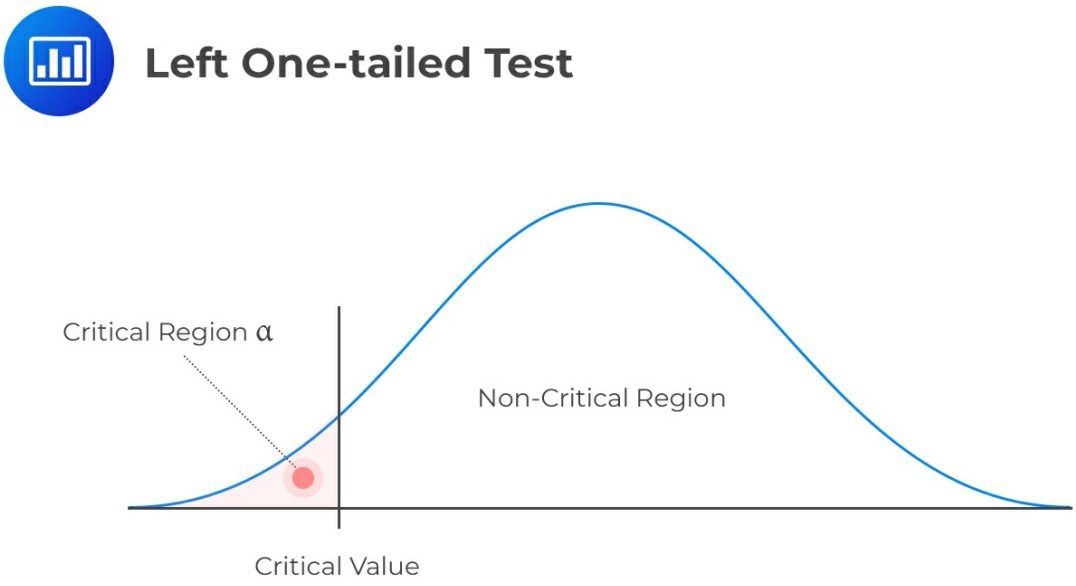
For the t-test, the decision rule is dependent on the alternative hypothesis. When testing the two-side alternative, the decision is to reject the null hypothesis if $|T| > C_\alpha$. That is, reject the null hypothesis if the absolute value of the test statistic is greater than the critical value. When testing on the one-sided, the decision rule, reject the null hypothesis if $T < C_\alpha$ when using a one-sided lower alternative and if $T > C_\alpha$ when using a one-sided upper alternative. When a null hypothesis is rejected at an α significance level, we say that the result is significant at α significance level.

Note that prior to decision making, one must decide whether the test should be one-tailed or two-tailed. The following is a brief summary of the decision rules under different scenarios:

Left One-tailed Test

H_1: parameter < X

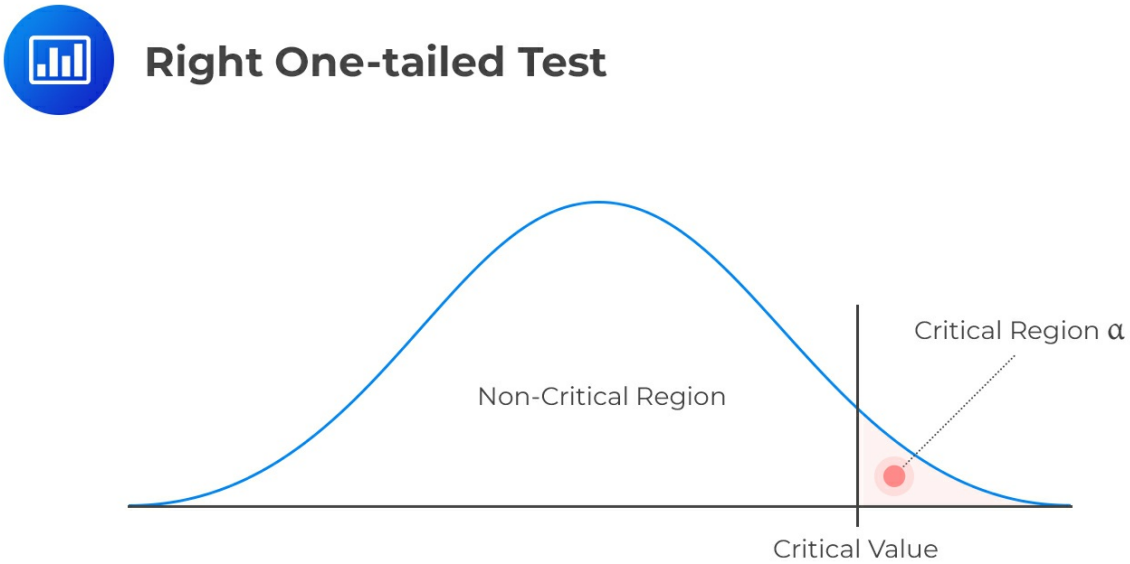
Decision rule: Reject H_0 if the test statistic is less than the critical value. Otherwise, **do not reject** H_0.



Right One-tailed Test

H_1: parameter > X

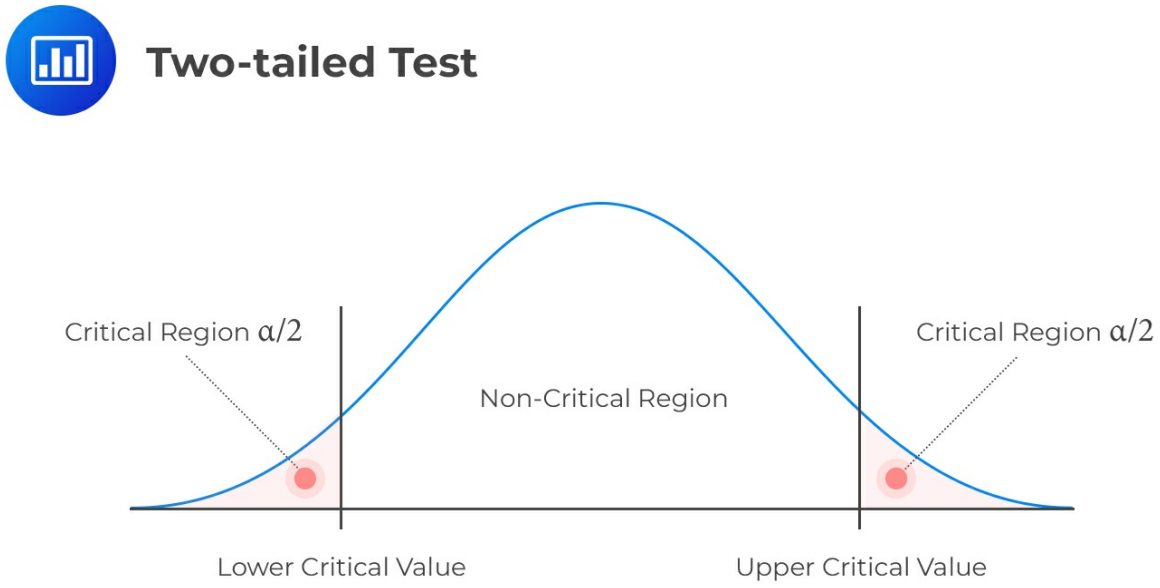
Decision rule: Reject H_0 if the test statistic is greater than the critical value. Otherwise, **do not reject** H_0.



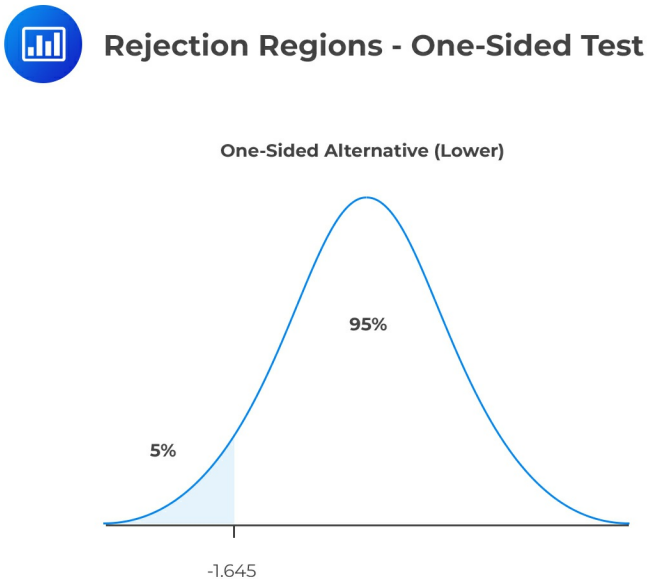
Two-tailed Test

H_1: parameter $\neq$ X (not equal to X)

Decision rule: Reject H_0 if the test statistic is greater than the upper critical value or less than the lower critical value.



Consider, α=5%. Consider a one-sided test. The rejection regions are shown below:



The first graph represents the rejection region when the alternative is one-sided lower. For instance, the hypothesis is stated as:

$\mathbf{H_0: \mu < \mu_0}$ **vs.** $\mathbf{H_1: \mu > \mu_0}$.

The second graph represents the rejection region when the alternative is a one-sided upper. The null hypothesis, in this case, is stated as:

$\mathbf{H_0: \mu > \mu_0}$ **vs.** $\mathbf{H_1: \mu < \mu_0}$.

Example: Hypothesis Test on the Mean

Consider the returns from a portfolio $X = (x_1, x_2, \dots, x_n)$ from 1980 through 2020. The approximated mean of the returns is 7.50%, with a standard deviation of 17%. We wish to determine whether the expected value of the return is different from 0 at a 5% significance level.

Solution

We start by stating the two-sided hypothesis test:

$\mathbf{H_0: \mu = 0}$ **vs.** $\mathbf{H_1: \mu \neq 0}$

The test statistic is:

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \sim N(0,1)$$

In this case, we have,

n=40

$\hat{\mu}$=0.075

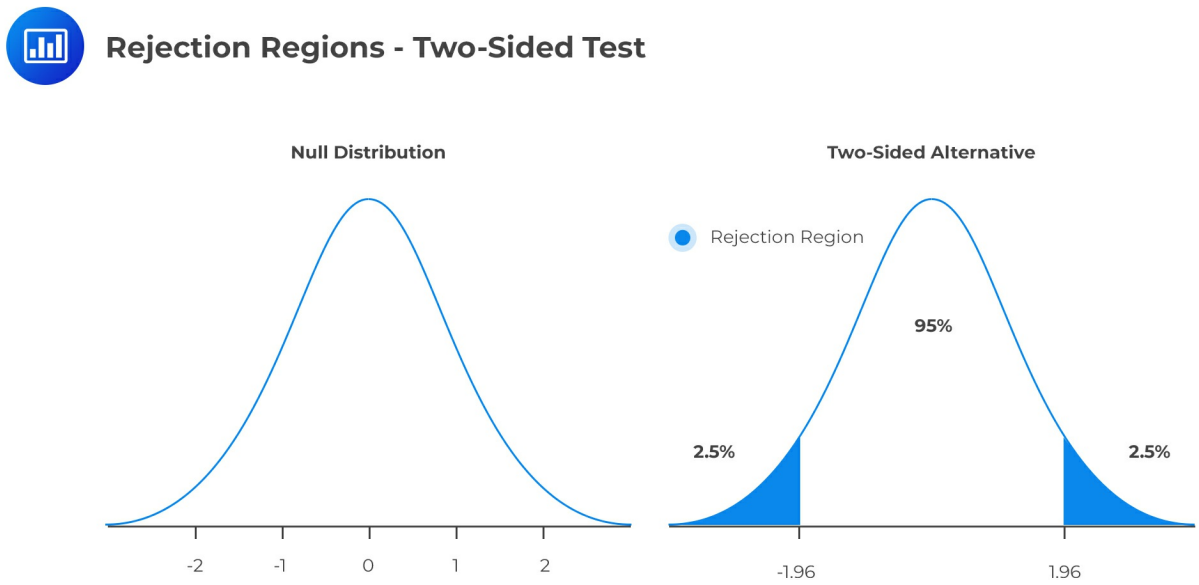
μ_0=0

$\hat{\sigma}^2$=0.17^2

So,

$$T = \frac{0.075 - 0}{\sqrt{\frac{0.17^2}{40}}} \approx 2.79$$

At the significance level, $\alpha = 5\%$,the critical value is ± 1.96. Since this is a two-sided test, the rejection regions are $(-\infty, -1.96)$ and $(1.96, \infty)$ as shown in the diagram below:



Since the test statistic (2.79) is higher than the critical value, then we reject the null hypothesis in favor of the alternative hypothesis.

The example above is an example of a Z-test (which is mostly emphasized in this chapter and immediately follows from the central limit theorem (CLT)). However, we can use the Student's t-distribution if the random variables are iid and normally distributed and that the sample size is small (n<30).

In Student's t-distribution, we used the unbiased estimator of variance. That is:

$$s^2 = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

Therefore the test statistic for $H_0 = \mu_0$ is given by:

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$$

The Type II Error and the Test Power

The power of a test is the direct opposite of the level of significance. While the level of relevance gives us the probability of rejecting the null hypothesis when it's, in fact, true, the power of a test gives the probability of correctly discrediting and rejecting the null hypothesis when it is false. In other words, it gives the likelihood of rejecting H_0 when, indeed, it's false. Denoting the probability of type II error by β, the power test is given by:

$$\text{Power of a Test} = 1 - \beta$$

The power test measures the likelihood that the false null hypothesis is rejected. It is influenced by the sample size, the length between the hypothesized parameter and the true value, and the size of the test.

Confidence Intervals

A confidence interval can be defined as the range of parameters at which the true parameter can be found at a confidence level. For instance, a 95% confidence interval constitutes the set of parameter values where the null hypothesis cannot be rejected when using a 5% test size. Therefore, a 1-α confidence interval contains the values that cannot be disregarded at a test size of α.

It is important to note that the confidence interval depends on the alternative hypothesis statement in the test. Let us start with the two-sided test alternatives.

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

Then the $1 - \alpha$ confidence interval is given by:

$$[\hat{\mu} - C_\alpha \times \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + C_\alpha \times \frac{\hat{\sigma}}{\sqrt{n}}]$$

C_α is the critical value at α test size.

Example: Calculating Two-Sided Alternative Confidence Intervals

Consider the returns from a portfolio $X = (x_1, x_2, \dots, x_n)$ from 1980 through 2020. The approximated mean of the returns is 7.50%, with a standard deviation of 17%. Calculate the 95% confidence interval for the portfolio return.

The $1 - \alpha$ confidence interval is given by:

$$\begin{aligned}
&[\hat{\mu} - C_\alpha \times \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + C_\alpha \times \frac{\hat{\sigma}}{\sqrt{n}}] \\
&= [0.0750 - 1.96 \times \frac{0.17}{\sqrt{40}}, 0.0750 + 1.96 \times \frac{0.17}{\sqrt{40}}] \\
&= [0.02232, 0.1277]
\end{aligned}$$

Thus, the confidence intervals imply any value of the null between 2.23% and 12.77% cannot be rejected against the alternative.

One-Sided Alternative

For the one-sided alternative, the confidence interval is given by either:

$$(-\infty, \hat{\mu} + C_{\alpha} \times \frac{\hat{\sigma}}{\sqrt{n}})$$

for the lower alternative

or,

$$(\hat{\mu} + C_{\alpha} \times \frac{\hat{\sigma}}{\sqrt{n}}, \infty)$$

for the upper alternative.

Example: Calculating the One-Sided Alternative Confidence Interval

Assume that we were conducting the following one-sided test:

$H_0 : \mu \leq 0$

$H_1 : \mu > 0$

The 95% confidence interval for the portfolio return is:

$$\begin{aligned} &= (-\infty, \hat{\mu} + C_{\alpha} \times \frac{\hat{\sigma}}{\sqrt{n}}) \\ &= (-\infty, 0.0750 + 1.645 \times \frac{0.17}{\sqrt{40}}) \\ &= (-\infty, 0.1192) \end{aligned}$$

On the other hand, if the hypothesis test was:

$H_0 : \mu > 0$

$H_1 : \mu \leq 0$

The 95% confidence interval would be:

$$= (-\infty, \hat{\mu} + C_{\alpha} \times \frac{\hat{\sigma}}{\sqrt{n}})$$

$$= (-\infty, 0.0750 + 1.645 \times \frac{0.17}{\sqrt{40}}) = (-\infty, 0.1192)$$

Note that the critical value decrease from 1.96 to 1.645 due to a change in the direction of the change.

The p-Value

When carrying out a statistical test with a fixed value of the significance level (α), we merely compare the observed test statistic with some critical value. For example, we might "reject H_0 using a 5% test" or "reject H_0 at 1% significance level". The problem with this 'classical' approach is that it does not give us the details about the **strength of the evidence** against the null hypothesis.

Determination of the *p-value* gives statisticians a more informative approach to hypothesis testing. The p-value is the lowest level at which we can reject H_0. This means that the strength of the evidence against H_0 increases as the *p-value* becomes smaller. The test-statistic depends on the alternative.

The p-Value for One-Tailed Test Alternative

For one-tailed tests, the *p-value* is given by the probability that lies below the calculated test statistic for left-tailed tests. Similarly, the likelihood that lies above the test statistic in right-tailed tests gives the *p-value.*

Denoting the test statistic by T, the p-value for $H_1 : \mu > 0$ is given by:

$$P(Z > |T|) = 1 - P(Z \leq |T|) = 1 - \Phi(|T|)$$

Conversely, for $H_1 : \mu \leq 0$ the p-value is given by:

$$P(Z \leq |T|) = \Phi(|T|)$$

Where z is a standard normal random variable, the absolute value of T ($|T|$) ensures that the right tail is measured whether T is negative or positive.

The p-Value for Two-Tailed Test Alternative

If the test is two-tailed, this value is given by the sum of the probabilities in the two tails. We start by determining the probability lying below the negative value of the test statistic. Then, we add this to the probability lying above the positive value of the test statistic. That is the p-value for the two-tailed hypothesis test is given by:

$$2[1 - \Phi[|T|]$$

Example 1: p-Value for One-Sided Alternative

Let θ represent the probability of obtaining a head when a coin is tossed. Suppose we toss the coin 200 times, and heads come up in 85 of the trials. Test the following hypothesis at 5% level of significance.

H_0: $\theta = 0.5$

H_1: $\theta < 0.5$

Solution

First, not that repeatedly tossing a coin follows a binomial distribution.

Our p-value will be given by P(X < 85) where X `binomial(200,0.5) with mean 100(np=200*0.5), assuming H_0 is true.

$$P\left[z < \frac{85.5 - 100}{\sqrt{50}}\right] = P(Z < -2.05)$$
$$= 1 - 0.97982 = 0.02018$$

Recall that for a binomial distribution, the variance is given by:

$$np(1-p) = 200(0.5)(1-0.5) = 50$$

(We have applied the Central Limit Theorem by taking the binomial distribution as approx. normal)

Since the probability is less than 0.05, H_0 is extremely unlikely, and we actually have strong evidence against H_0 that favors H_1. Thus, clearly expressing this result, we could say:

"There is very strong evidence against the hypothesis that the coin is fair. We, therefore, conclude that the coin is biased against heads."

Remember, failure to reject H_0 does not mean it's true. It means there's insufficient evidence to justify rejecting H_0, given a certain level of significance.

Example 2: p-Value for Two-Sided Alternative

A CFA candidate conducts a statistical test about the mean value of a random variable X.

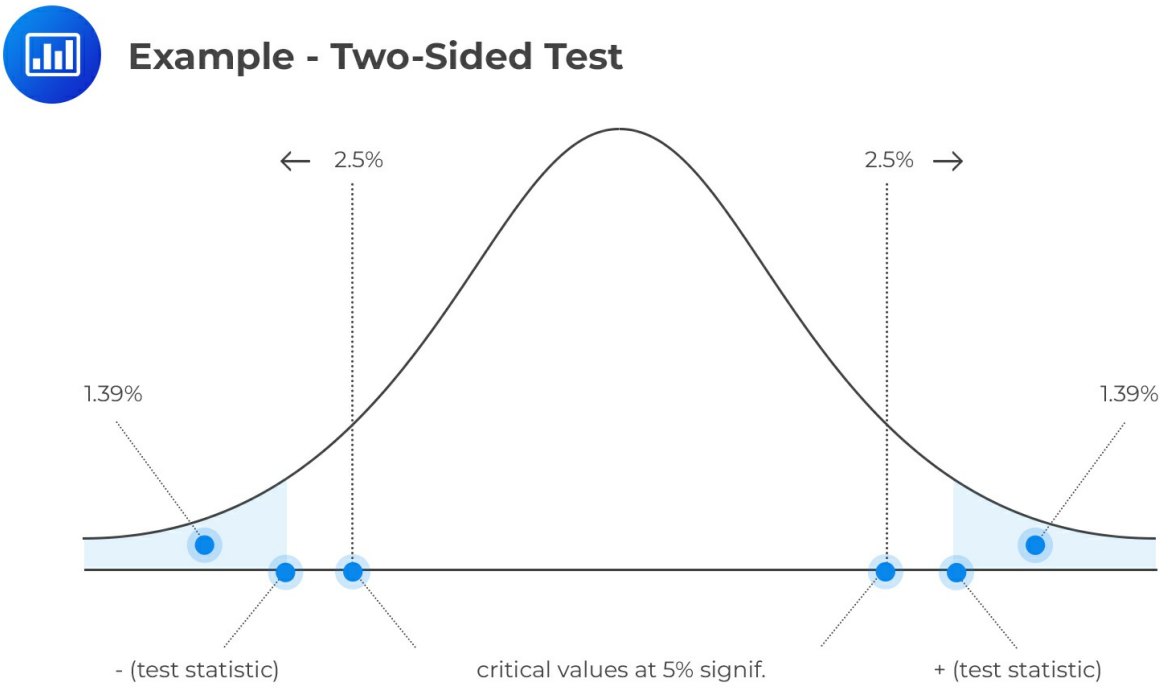
$H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$

She obtains a test statistic of 2.2. Given a 5% significance level, determine and interpret the *p-value*

Solution

$$\text{P-value} = 2P(Z > 2.2) = 2[1 - P(Z \leq 2.2)] = 1.39\% \times 2 = 2.78\%$$

(We have multiplied by two since this is a two-tailed test)



Interpretation

The p-value (2.78%) is less than the level of significance (5%). Therefore, we have sufficient evidence to reject H_0. In fact, the evidence is so strong that we would also reject H_0 at significance levels of 4% and 3%. However, at significance levels of 2% or 1%, we would not reject H_0 since the *p-value* surpasses these values.

Hypothesis about the Difference between Two Population Means.

It's common for analysts to be interested in establishing whether there exists a significant difference between the means of two different populations. For instance, they might want to know whether the average returns for two subsidiaries of a given company exhibit **significant** differences.

Now, consider a bivariate random variable:

$$W_i = [X_i, Y_i]$$

Assume that the components X_i and Y_i are both iid and are correlated. That is:

$Corr(X_i, Y_i) \neq 0$

Now, suppose that we want to test the hypothesis that:

$$H_0 : \mu_X = \mu_Y$$

$$H_1 : \mu_X \neq \mu_Y$$

In other words, we want to test whether the constituent random variables have equal means. Note that the hypothesis statement above can be written as:

$$H_0 : \mu_X - \mu_Y = 0$$

$$H_1 : \mu_X - \mu_Y \neq 0$$

To execute this test, consider the variable:

$$Z_i = X_i - Y_i$$

Therefore, considering the above random variable, if the null hypothesis is correct then,

$$E(Z_i) = E(X_i) - E(Y_i) = \mu_X - \mu_Y = 0$$

Intuitively, this can be considered as a standard hypothesis test of

H_0: $\mu_Z = 0$ vs. H_1: $\mu_Z \neq 0$.

The tests statistic is given by:

$$T = \frac{\hat{\mu}_z}{\sqrt{\frac{\hat{\sigma}_z^2}{n}}} \sim N(0, 1)$$

Note that the test statistic formula accounts for the correction between X_i and Y_i. It is easy to see that:

$$V(Z_i) = V(X_i) + V(Y_i) - 2COV(X_i, Y_i)$$

Which can be denoted as:

$$\hat{\sigma}_z^2 = \hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\sigma_{XY}$$

$$\hat{\mu}_z = \mu_X - \mu_Y$$

And thus the test statistic formula can be written as:

$$T = \frac{\mu_X - \mu_Y}{\sqrt{\frac{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\sigma_{XY}}{n}}}$$

This formula indicates that correlation plays a crucial role in determining the magnitude of the test statistic.

Another special case of the test-statistic is when X_i, and Y_i are iid and independent. The test statistic is given by:

$$T = \frac{\mu_X - \mu_Y}{\sqrt{\frac{\hat{\sigma}_X^2}{n_X} + \frac{\hat{\sigma}_Y^2}{n_Y}}}$$

Where n_X and n_Y are the sample sizes of X_i, and Y_i respectively.

Example: Hypothesis Test on Two Means

An investment analyst wants to test whether there is a significant difference between the means of the two portfolios at a 95% level. The first portfolio X consists of 30 government-issued bonds and has a mean of 10% and a standard deviation of 2%. The second portfolio Y consists of 30 private bonds with a mean of 14% and a standard deviation of 3%. The correlation between the two portfolios is 0.7. Calculate the null hypothesis and state whether the null hypothesis is rejected or otherwise.

Solution

The hypothesis statement is given by:

H_0: μ_X - μ_Y=0 vs. H_1: μ_X - $\mu_Y \neq 0$.

Note that this is a two-tailed test. At 95% level, the test size is α=5% and thus the critical value $C_\alpha = \pm 1.96$.

Recall that:

$$Cov(X, Y) = \sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y$$

Where ρ_XY is the correlation coefficient between X and Y.

Now the test statistic is given by:

$$T = \frac{\mu_X - \mu_Y}{\sqrt{\frac{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\sigma_{XY}}{n}}} = \frac{\mu_X - \mu_Y}{\sqrt{\frac{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}{n}}}$$

$$= \frac{0.10 - 0.14}{\sqrt{\frac{0.02^2 + 0.03^2 - 2\times 0.7\times 0.02\times 0.03}{30}}} = -10.215$$

The test statistic is far much less than -1.96. Therefore the null hypothesis is rejected at a 95% level.

The Problem of Multiple Testing

Multiple testing occurs when multiple multiple hypothesis tests are conducted on the same data set. The reuse of data results in spurious results and unreliable conclusions that do not hold up to scrutiny. The fundamental problem with multiple testing is that the test size (i.e., the probability that a true null is rejected) is only applicable for a single test. However, repeated testing creates test sizes that are much larger than the assumed size of alpha and therefore increases the probability of a Type I error.

Some control methods have been developed to combat multiple testing. These include Bonferroni correction, the False Discovery Rate (FDR), and Familywise Error Rate (FWER).

Practice Question

An experiment was done to find out the number of hours that candidates spend preparing for the FRM part 1 exam. It was discovered that for a sample of **10 students**, the following times were spent:

318, 304, 317, 305, 309, 307, 316, 309, 315, 327

If the sample mean and standard deviation are 312.7 and 7.2, respectively, calculate a symmetrical 95% confidence interval for the mean time a candidate spends preparing for the exam using the t-table.

q	0.95	0.975	0.99	0.995	0.999	0.9995
n=1	6.314	12.706	31.821	63.657	318.309	636.619
2	2.920	4.303	6.965	9.925	22.327	31.599
3	2.353	3.182	4.541	5.841	10.215	12.924
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.894	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318

A. [307.5, 317.9]

B. [307.6, 317.8]

C. [307.9, 317.5]

D. [307.3, 318.2]

The correct answer is **A.**

Population variance is unknown; we must use the **t-score**.

To find the value of $t_{1-\frac{\alpha}{2}}$, we use the t-table with (10 - 1 =) 9 degrees of freedom and the

(1 - 0.025 =) 0.975 which gives us 2.262.

So the confidence interval is given by:

$$\bar{X} \pm t_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} = 312.7 \pm 2.262 \times \frac{7.2}{\sqrt{10}}$$
$$= [307.5, 317.9]$$

Reading 18: Linear Regression

After completing this reading, you should be able to:

- Describe the models that can be estimated using linear regression and differentiate them from those which cannot.
- Interpret the results of an OLS regression with a single explanatory variable.
- Describe the key assumptions of OLS parameter estimation.
- Characterize the properties of OLS estimators and their sampling distributions.
- Construct, apply, and interpret hypothesis tests and confidence intervals for a single regression coefficient in a regression.
- Explain the steps needed to perform a hypothesis test in linear regression.
- Describe the relationship between a t-statistic, its p-value, and a confidence interval.

Linear regression is a statistical tool for modeling the relationship between two random variables. This chapter will concentrate on the linear regression model (regression model with one explanatory variable).

The Linear Regression Model

As stated earlier, linear regression determines the relationship between the dependent variable Y and the independent (explanatory) variable X. The linear regression with a single explanatory variable is given by:

$$Y = \beta_0 + \beta X + \epsilon$$

Where:

β_0=constant intercept (the value of Y when X=0)

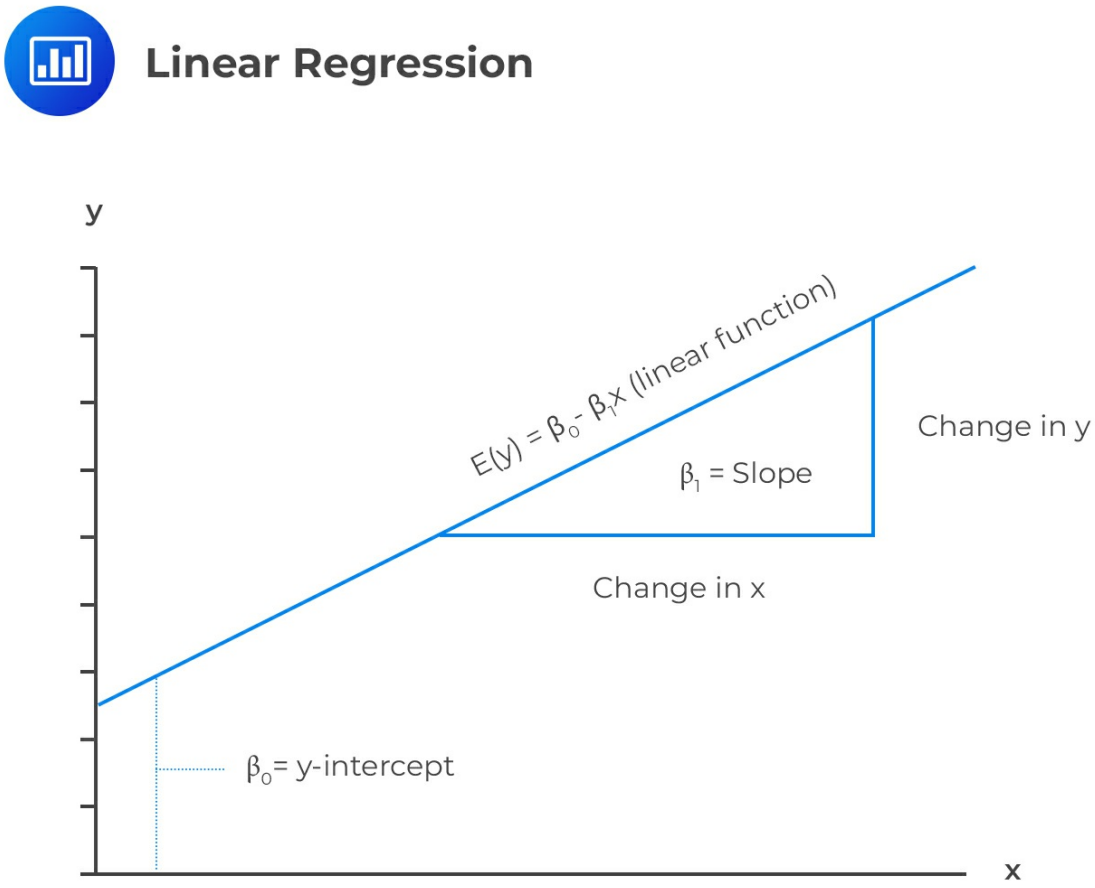
β=the Slope which measures the sensitivity of Y to variation in X.

ϵ=error(sometimes referred to as shock). It represents the portion of Y that cannot be explained by X.

The assumption is that the expectation of the error is 0. That is, $E(\epsilon) = 0$ and thus,

$$E[Y] = E[\beta_0] + \beta E[X] + E[\epsilon]$$

$$\Rightarrow E[Y] = \beta_0 + \beta E[X]$$



Note that β_0 is the value of Y when $X = 0$. However, there are cases when the explanatory variable is not equal to 0. In this case, β_0 is interpreted as the value that ensures that the $\bar{Y}$ in the regression line $\bar{Y} = \hat{\beta}_0 + \hat{\beta}\bar{X}$ where $\bar{Y}$ and $\bar{X}$ are the mean of y_i and x_i random variables.

The Linearity of a Regression

The independent variable can be continuous, discrete or even functions. Above the diversity of the explanatory variables, they must satisfy the following conditions:

1. The relationship between the dependent variable Y and the explanatory variables $(X_1, X_2, \dots, X_n)$ must be linear.
2. The error term must be additive except where the variance of the error term depends on the explanatory variables.
3. The independent (explanatory variables) must be observables. This ensures that a linear regression with missing data is not developed.

A good example of a violation of the linearity principle is:

$$Y = \beta_0 + \beta X^k + \epsilon$$

This model cannot be estimated using linear regression due to the presence of the unknown parameter *k,* which violates the first restriction (it is non-linear regression function). This kind of nonlinearity can be corrected through transformation.

Transformations

When a linear regression model does not satisfy the linearity conditions stated above, we can reverse the violation of the restrictions by transforming the model. Consider the model:

$$Y = \beta_0 X^{\beta} \epsilon$$

Where ϵ is the positive error term (shock). Clearly, this model violates the condition of the restriction since X is raised to an unknown parameter β, and the error term is not additive. However, we can make this model linear by taking natural logarithm on both sides of the equation so that:

$$\ln(Y) = (\beta_0 X^{\beta} \epsilon)$$

$$\ln(Y) = \ln\beta_0 + \beta \ln X + \ln\epsilon$$

The last equation can be written as:

$$Y = \hat{\beta}_0 + \beta\hat{X}^k + \hat{\epsilon}$$

Clearly, this equation satisfies the three linearity conditions. It is worth noting that when we are interpreting the parameters of the transformed model, we measure the change of the transformed independent variable X on the transformed variable Y.

For instance, $\ln(Y) = \ln\beta_0 + \beta\ln X + \ln\epsilon$ implies that β represents the change in $\ln Y$ corresponding to a unit change in $\ln X$.

The Use of the Dummy Variables

There are cases where the explanatory variables are binary numbers (0 and 1) representing the occurrences of an event. These binary numbers are called dummies. For instance,

Assuming D_i is a variable such that:

$$D_i = \begin{cases} 1 & \text{The student-teacher ratio in ith school} < 20 \\ 0 & \text{The student-teacher ratio in ith school} \geq 20 \end{cases}$$

The following is the population regression model whose regressor D_i:

$$Y_i = \beta_0 + \beta D_i + \epsilon_i, \forall i = 0, \dots, n$$

β is the coefficient on D_i.

The equation will change to the one written below under the condition that $D_i = 0$:

$$Y_i = \beta_0 + \epsilon_i$$

When $D_i = 1$:

$$Y_i = \beta_0 + \beta + \epsilon_i$$

This implies that when $D_i = 1, E(Y_i|D_i = 1) = \beta_0 + \beta_1$. The test scores will have a population mean value of $\beta_0 + \beta_1$ when the ratio of students to teachers is low. The conditional expectations of Y_i when $D_i = 1$ and when $D_i = 0$ will have a difference of β_1 between them written as:

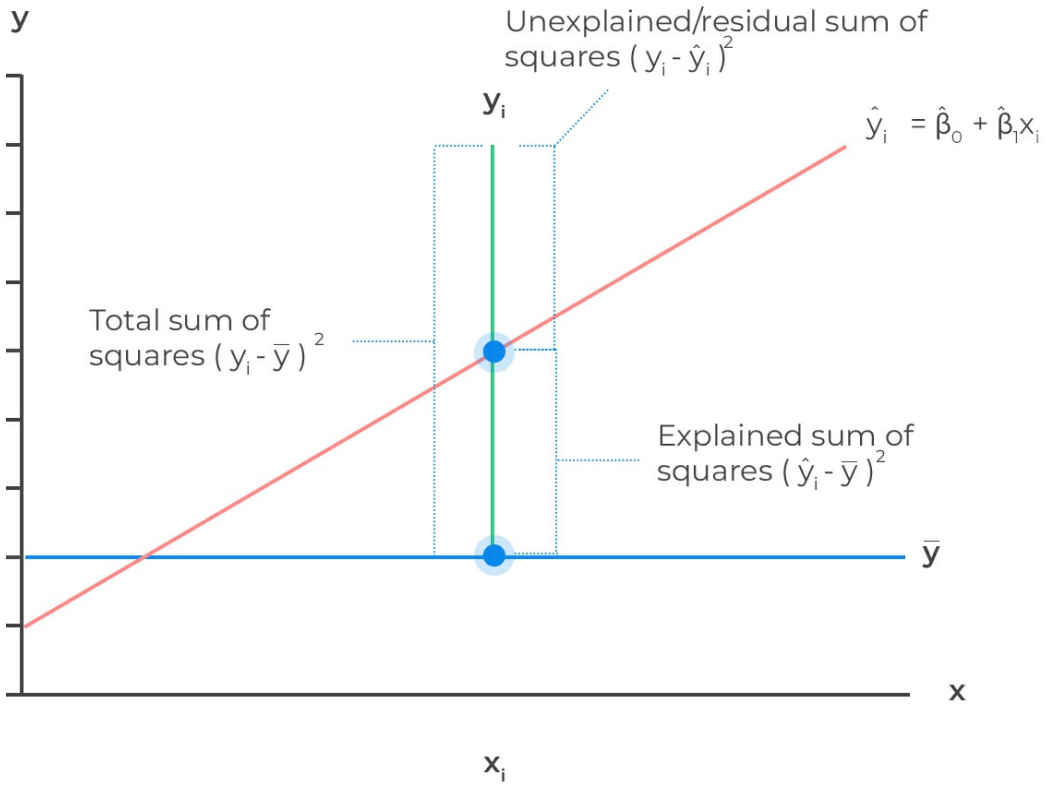
$$(\beta_0 + \beta) - \beta_0 = \beta$$

This makes β to be the difference between population means.

The Ordinary Least Squares

The Ordinary Least Squares (OLS) is a method of estimating the linear regression parameters by minimizing the sum of squared deviations. The regression coefficients chosen by the OLS estimators are such that the observed data and the regression line are as close as possible.

Ordinary Least Squares



Consider a regression equation:

$$Y = \beta_0 + \beta X + \epsilon$$

Where each of X and Y consists of n observations each $(X = x_1, x_2, \dots, n)$ and $(Y = y_1, y_2, \dots, y_n)$. Assume that each of x_i and y_i are linearly related, then the parameters can be estimated using the OLS. The estimators minimize the residual sum of squares such that:

$$\sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta} x_i)^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Where the $\hat{\beta}_0$ and $\hat{\beta}$ are parameter estimators (intercept and the slope respectively) which minimizes the squared deviations between the line $\hat{\beta_0} + \hat{\beta} x_i$ and y_i so that:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

Where $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

After the estimation of the parameters, the estimated regression line is given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}x_i$$

And the linear regression residual error term is given by:

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}x_i$$

The variance of the error term is approximated as:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

It can also be shown that:

$$s^2 = \frac{n}{n-2}\hat{\sigma}_Y^2(1 - \hat{\rho}_{XY}^2)$$

Note that n-2 implies that two parameters are estimated and that s^2 is an unbiased estimator of σ^2. Moreover, it is assumed that the mean of the residuals is zero and uncorrelated with the explanatory variables X_i.

Now, consider the formula:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

If we multiply both the numerator and the denominator by $\frac{1}{n}$, we have:

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - X)(y_i - Y)}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

Note that the numerator is the covariance between X and Y, and the denominator is the variance of X. So that we can write:

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2} = \frac{\hat{\sigma}_{XY}}{\sigma_X^2}$$

Also recall that:

$$\text{Corr}(X, Y) = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\Rightarrow \sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y$$

So,

$$\hat{\beta} = \frac{\rho_{XY}\sigma_X\sigma_Y}{\hat{\sigma}_X^2}$$

$$\therefore \hat{\beta} = \frac{\hat{\rho}_{XY}\hat{\sigma}_Y}{\hat{\sigma}_X}$$

Example: Estimating the Linear Regression Parameters

An investment analyst wants to explain the return from the portfolio (Y) using the prevailing interest rates (X) over the past 30 years. The mean interest rate is 7%, and the return from the portfolio is 14%. The covariance matrix is given by:

$$\begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\sigma}_{XY} \\ \hat{\sigma}_{XY} & \hat{\sigma}_X^2 \end{bmatrix} = \begin{bmatrix} 1600 & 500 \\ 500 & 338 \end{bmatrix}$$

Assume that the analyst wants to estimate the linear regression equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i$$

Estimate the parameters and, thus, the model equation.

Solution

Now,

$$\hat{\beta} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{500}{338} = 1.4793$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X} = 0.14 - 1.4793 \times 0.07 = 0.0364$$

So, the estimated equation is given by:

$$\hat{Y}_i = 0.0364 + 1.4793X_i$$

Assumptions of OLS

The OLS estimators assume the following:

1. The conditional distribution of the error term given the independent variables X_i is 0. More precisely $E(\epsilon_i|X_i) = 0$. This also implies that the independent variables and the error term are uncorrelated and that $E(\epsilon_i) = 0$.
2. Both the dependent and independent variables are i.i.d. This assumption concerns the drawing of the sample. According to this assumption, $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d in case a simple random sampling is applied when drawing observations from a single large population. Despite the i.i.d assumption being a reasonable assumption for many data collection schemes, all sampling schemes do not produce i.i.d observations on (X_i, Y_i).
3. Large outliers are unlikely. In this assumption, observations whose values of X_i and/or Y_i fall far outside the usual range of the data, are unlikely. These observations are known as significant outliers. Results of OLS regression can be misleading due to large outliers.
4. The variance of the independent variable is strictly nonnegative. That is, $\sigma_X^2 > 0$. This is essential in estimating the regression parameters.

5. The variance of the error term is independent of the explanatory variables and that $V(\epsilon_i|X) = \sigma^2 < \infty$ and that the variance of all the error terms (shocks) is equal. This assumption is termed as the homoskedasticity assumption.

The OLS estimators imply that the parameter estimators are unbiased estimators. That is, $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$. This is actually true for large sample sizes or rather as the sample sizes increases.

Lastly, the assumptions ensure that that the estimated parameters are normally distributed. The asymptotic distribution of the slope is given by:

$$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, \frac{\sigma^2}{\sigma_X^2})$$

Where σ^2 is the variance of the error term and σ_X^2 is the variance of X. It is easy to see that the variance of $\hat{\beta}$ increases as σ^2 increases.

For the intercept, the asymptotic distribution is defined as:

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) \sim N(0, \frac{\sigma^2(\mu_X^2 - \sigma_X^2)}{\sigma_X^2})$$

According to the central limit theorem (CLT), $\hat{\beta}$ can be treated as the standard random variable with the mean as the true value β and the variance $\frac{\sigma^2}{n\sigma_X^2}$. That is:

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{n\sigma_X^2})$$

However, we cannot use this value in hypothesis testing. We need to use the variance estimators such that:

$$\sigma^2 = s^2$$

So, recall that for a large sample size:

$$\hat{\sigma}_X = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$\Rightarrow n\hat{\sigma}_X = \sum_{i=1}^{n} (x_i - \bar{X})^2$$

Therefore, the variance of the parameter β can be written as:

$$\hat{\sigma}_\beta^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2} = \frac{s^2}{n\hat{\sigma}_X^2}$$

The standard error estimate of the β denoted as SEE_β is equivalent to the square root of its variance, so:

$$SEE_\beta = \sqrt{\frac{s^2}{n\hat{\sigma}_X^2}} = \frac{s}{\sqrt{n}\hat{\sigma}_X}$$

Analogously, the variance of the intercept:

$$\hat{\sigma}_{\beta_0}^2 = \frac{s^2 (\hat{\mu}_X^2 + \hat{\sigma}_X^2)}{n\hat{\sigma}_X^2}$$

Hypothesis Testing on the Linear Regression Parameters

When the OLS assumptions are met, the parameters are assumed to be normally distributed when large samples are used. Therefore, we can run a hypothesis tests on the parameters just like the random variable.

A hypothesis is a statistical procedure where an analyst tests an assumption on the population parameters. For instance, we may want to test the significance of a **single** regression coefficient in a simple linear regression. Most of the hypothesis tests are t-tests.

Whenever a statistical test is being performed, the following procedure is generally considered ideal:

1. Statement of both the null and the alternative hypothesis;
2. Select the appropriate test statistic, i.e., what's being tested, e.g., the population means, the difference between sample means, or variance;

3. Specify the level of significance;
4. Clearly, state the decision rule to guide you in choosing whether to reject or not to reject the null hypothesis;
5. Calculate the sample statistic, and finally
6. Make a decision based on the sample results.

For instance, assume we are testing the null hypothesis that:

$$H_0 : \beta = \beta_{H_0} \text{vs.} H_1 : \beta \neq \beta_{H_0}$$

Where β_{H_0} is the hypothesized slope parameter.

Then the test statistic will be:

$$T = \frac{\hat{\beta} - \beta_{H_0}}{SEE_{\beta}}$$

This statistic possesses asymptotic normal distribution, which is then compared to a critical value C_t . The null hypothesis is rejected if:

$$|T| > C_t$$

For instance, if we assume a 5% significance level in this case, then the critical value is 1.96.

We can also evaluate the p-values. For one-tailed tests, the p-value is given by the probability that lies below the calculated test statistic for left-tailed tests. Similarly, the likelihood that lies above the test statistic in right-tailed tests gives the p-value.

Denoting the test statistic by T, the p-value for $H_1 : \hat{\beta} > 0$ is given by:

$$P(Z > |T|) = 1 - P(Z \leq |T|) = 1 - \Phi(|T|)$$

Conversely, for $H_1 : \hat{\beta} \leq 0$ the p-value is given by:

$$P(Z \leq |T|) = \Phi(|T|)$$

Where z is a standard normal random variable, the absolute value of T (|T|) ensures that the right tail is measured whether T is negative or positive.

If the test is two-tailed, this value is given by the sum of the probabilities in the two tails. We start by determining the probability lying below the negative value of the test statistic. Then, we add this to the probability lying above the positive value of the test statistic. That is the p-value for the two-tailed hypothesis test is given by:

$$2[1 - \Phi(|T|)]$$

We can also construct confidence intervals (discussed in detail in the previous chapter). Recall that a confidence interval can be defined as the range of parameters at which the true parameter can be found at a confidence level. For instance, a 95% confidence interval constitutes that the set of parameter values where the null hypothesis cannot be rejected when using a 5% test size.

For instance, if we are performing the two-tailed hypothesis tests, then the confidence interval is given by:

$$[\hat{\beta} - C_t \times SEE_{\beta}, \hat{\beta} + C_t \times SEE_{\beta}]$$

Example: Hypothesis Test on the Linear Regression Parameters

An investment analyst wants to explain the return from the portfolio (Y) using the prevailing interest rates (X) over the past 30 years. The mean interest rate is 7%, and the return from the portfolio is 14%. The covariance matrix is given by:

$$\begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\sigma}_{XY} \\ \hat{\sigma}_{XY} & \hat{\sigma}_X^2 \end{bmatrix} = \begin{bmatrix} 1600 & 500 \\ 500 & 338 \end{bmatrix}$$

Assume that the analyst wants to estimate the linear regression equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i$$

Test whether the slope coefficient is equal to zero and construct a 95% confidence interval for the

slope of the coefficient.

Solution

We start by stating the hypothesis:

$$T = \frac{\hat{\beta} - \beta_{H_0}}{SEE_{\beta}}$$

We had calculated the slope from the matrix as:

$$\hat{\beta} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = \frac{500}{338} = 1.4793$$

Now, recall that:

$$SEE_{\hat{\beta}} = \frac{s}{\sqrt{n}\hat{\sigma}_X}$$

But

$$s^2 = \frac{n}{n-2}\hat{\sigma}_Y^2 (1 - \hat{\rho}_{XY})$$

So, in this case:

$$s^2 = \frac{30}{30-2} \times 1600 \left(1 - \frac{500}{\sqrt{338}\sqrt{1600}}\right) = 548.7251$$

(Note that for $\hat{\rho}_{XY}$ we have used the relationship $\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$.)

Therefore,

$$s = \sqrt{s^2} = \sqrt{548.7251} = 23.4249$$

So,

$$SEE_{\hat{\beta}} = \frac{s}{\sqrt{n}\hat{\sigma}_X} = \frac{23.4249}{\sqrt{30}\sqrt{338}} = 0.23263$$

Therefore the t-statistic is given by:

$$T = \frac{\hat{\beta} - \beta_{H_0}}{SEE_{\beta}} = \frac{1.4793}{0.23263} = 6.3590$$

For the two-tailed test, the critical value is 1.96, and since the t-statistic here is greater than the significant value, then we reject the null hypothesis.

For the 95% CI, we know it is given by:

$$\hat{\beta} - C_t \times SEE_{\beta}, \hat{\beta} + C_t \times SEE_{\beta}$$

$$= [1.4793 - 1.96 \times 0.23263, 1.4793 + 1.96 \times 0.23263]$$

$$= [1.0233, 1.9353]$$

Practice Question 1

Assume that you have carried out a regression analysis (to determine whether the slope is different from 0) and found out that the slope $\hat{\beta} = 1.156$. Moreover, you have constructed a 95% confidence interval of [0.550, 1.762]. What is the likely value of your test statistic?

A. 4.356

B. 3.7387

C. 0.7845

D. 0.6545

Solution

The Correct answer is B

This is a two-tailed test since we're asked to determine if the slope is different from zero. We know that:

$$[\hat{\beta} - C_t \times SEE_\beta, \hat{\beta} + C_t \times SEE_\beta]$$

Which in this case is [0.550, 1.762].

We need to find the value of SEE_β. That is:

$$1.156 - 1.96 \times SEE_\beta = 0.550 \Rightarrow SEE_\beta = \frac{1.156 - 0.550}{1.96} = 0.3092$$

And we know that:

$$T = \frac{\hat{\beta} - \beta_{H_0}}{SEE_\beta} = \frac{1.156 - 0}{0.3092} = 3.7387$$

Practice Question 2

A trader develops a simple linear regression model to predict the price of a stock. The estimated slope coefficient for the regression is 0.60, the standard error is equal to 0.25, and the sample has 30 observations. Determine if the estimated slope coefficient is significantly different than zero at a 5% level of significance by correctly stating the decision rule.

A. Accept H_1; The slope coefficient is statistically significant.

B. Reject H_0; The slope coefficient is statistically significant.

C. Reject H_0; The slope coefficient is not statistically significant.

D. Accept H_1; The slope coefficient is not statistically significant.

Solution

The correct answer is **B**.

Step 1: State the hypothesis

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Step 2: Compute the test statistic

$$\frac{\beta_1 - \beta_{H0}}{S_{\beta 1}} = \frac{0.60 - 0}{0.25} = 2.4$$

Step 3: Find the critical value, t_c

From the t table, we can find $t_{0.025,28}$ is 2.048

Step 4: State the decision rule

Reject H_0; The slope coefficient is statistically significant since $2.048 < 2.4$.

Reading 19: Regression with Multiple Explanatory Variables

After completing this reading, you should be able to:

- Distinguish between the relative assumptions of single and multiple regression.
- Interpret regression coefficients in multiple regression.
- Interpret goodness of fit measures for single and multiple regressions, including R^2 and adjusted R^2.
- Construct, apply, and interpret joint hypothesis tests and confidence intervals for multiple coefficients in regression.

Unlike linear regression, **multiple regression** simultaneously considers the influence of multiple explanatory variables on a response variable Y. In other words, it permits us to evaluate the effect of more than one independent variable on a given dependent variable.

The form of the multiple regression model (equation) is given by:

$$Y_i = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_k X_k i + \varepsilon_i \quad \forall i = 1, 2, \ldots n$$

Intuitively, the multiple regression model has *k* slope coefficients and *k+1* regression coefficients. Normally, statistical software (such as Excel and R) are used to estimate the multiple regression model.

Interpreting the Multiple Regression Coefficients

The slope coefficients β_k computes the level of variation of the dependent variable Y when the independent variable X_j changes by one unit while holding other independent variables constant. The interpretation of the multiple regression coefficients is quite different compared to linear regression with one independent variable. The effect of one variable is explored while keeping other independent variables constant.

For instance, a linear regression model with one independent variable could be estimated as $\hat{Y} = 0.6 + 0.85X_1$. In this case, the slope coefficient is 0.85, which implies that a 1 unit increase in X_1 results in 0.85 units increases independent variable Y.

Now, assume that we had the second independent variable to the regression so that the regression equation is $\hat{Y} = 0.6 + 0.85X_1 + 0.65X_2$. A unit increase in X_1 will not result in a 0.85 unit increase in Y unless X_1 and X_2 are uncorrelated. Therefore, we will interpret 0.85 as one unit of X_1 leads to 0.85 units increase in the dependent variable Y, while keeping X_2 constant.

OLS Estimators for the Multiple Regression Parameters

Although the multiples regression parameters can be estimated, it is challenging since it involves a huge amount of algebra and the use of matrices. However, we build a foundation of understanding using the multiple regression model with two explanatory variables.

Consider the following multiples regression equation.

$$Y_i = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \varepsilon_i$$

The OLS estimator of $\boldsymbol{\beta}_1$ is estimated as follows:

The first step is to regress X_1 and X_2 and to get the residual of X_{1i} given by:

$$\epsilon_{X_{1_i}} = X_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 X_{2i}$$

Where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are the OLS estimators of X_{2i}.

The next step is to regress Y on X_2 to get the residuals of $Y_{i,}$ which is intuitively given by:

$$\epsilon_{Y_i} = Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 X_{2i}$$

Where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are the OLS estimators of X_{2i}. The final step is to regress the residual of X_1 and Y ($\epsilon_{X_{1_i}}$ and ϵ_{Y_i}) to get:

$$\epsilon_{Y_i} = \hat{\beta}_1 \epsilon_{X_{1_i}} + \epsilon_i$$

Note that we do not have a constant, the expected values of ϵ_{Y_i} and ϵ_{X_i} are both 0. Moreover, the main purpose of the first and the second regression is to exclude the effect of X_2 from both Y and X_1 by dividing the variable into the fittest value which is correlated with X_2, and the residual error which is uncorrelated with X_2 and thus the two-residual obtained is uncorrelated with X_2 by intuition. The last step of the regression gives the regression between the components of Y and X_1, which is uncorrelated with X_2.

The OLS estimator for β_2 can be approximated analogously as that of β_1 by exchanging X_2 for X_1 in the process above. By repeating this process, we can estimate a k-parameter model such as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \forall i = 1, 2, \ldots n$$

Most of the time, this is done using a statistical package such as Excel and R.

Assumptions of the Multiple Regression Model

Suppose that we have n observations of the dependent variable (Y) and the independent variables ($X_1, X_2, \ldots, X_k$), we need to estimate the equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \quad \forall i = 1, 2, \ldots n$$

For us to make a valid inference from the above equation, we need to make classical normal multiple linear regression model assumptions as follows:

1. The relationship between the dependent variable, Y, and the independent variables, $X_1, X_2, \ldots, X_k$, is linear.
2. The independent variables ($X_1, X_2, \ldots, X_k$) are iid. Moreover, there is no definite linear relationship that exists between two or more of the independent variables, $X_1, X_2, \ldots, X_k$.
3. The expectation of value of the error term, conditioned on the independent variables, is 0: $E(\epsilon | X_1, X_2, \ldots, X_k) = 0$

4. The variance of the error term is equal for all observations. That is, $E(\epsilon_i^2) = \sigma_\epsilon^2, i = 1, 2, \dots, n$ (homoskedasticity assumption). The assumption enables us to estimate the distribution of the regression coefficients.
5. The error term ϵ is uncorrelated in all observations. Mathematically put, $E(\epsilon_i \epsilon_j) = 0 \quad \forall i \neq j$
6. The error term ϵ is normally distributed. This allows us to test the hypothesis about regression analysis.
7. There are no outliers so that $E(X_{ji}^4) < \infty$ for all j=1,2….k

The assumptions are almost the same as those of linear regression with one independent variable, only that the second assumption is tailored to ensure no linear relationships between the independent variables (multicollinearity).

Measures of Goodness of Fit

The goodness of fit of a regression is a measure using the Coefficient of determination (R^2) and the adjusted coefficient of determination.

The Coefficient of Determination (R^2)

Recall that the standard error estimate gives a percentage at which we are certain of a forecast made by a regression model. However, it does not tell us how suitable is the independent variable in determining the dependent variable. The coefficient of variation corrects this shortcoming.

The coefficient of variation measures a proportion of the total change in the dependent variable explained by the independent variable. We can calculate the coefficient of variation in two ways:

1. Squaring the Correlation Coefficient between the Dependent and Independent Variables.

The coefficient of variation can be computed by squaring the correlation coefficient (r) between the dependent and independent variables. That is:

$$R^2 = r^2$$

Recall that:

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Where

$Cov(X, Y)$-covariance between two variables, X and Y

σ_X-standard deviation of X

σ_Y-standard deviation of Y

However, this method only accommodates regression with one independent variable.

Example: Calculating the Coefficient of Determination using Correlation Coefficient

The correlation coefficient between the money supply growth rate (dependent, Y) and inflation rates (independent, X) is 0.7565. The standard deviation of the dependent (explained) variable is 0.050, and that of the independent variable is 0.02. Regression analysis for the ten years was conducted on this variable. We need to calculate the coefficient of determination.

Solution

We know that:

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{0.0007565}{0.05 \times 0.02} = 0.7565$$

So, the coefficient of determination is given by:

$$r^2 = 0.7565^2 = 0.5723 = 57.23$$

So, in regression, the money supply growth rate explains roughly 57.23% of the variation in the inflation rate over the ten years.

2. Method for Regression Model with One or More Independent

Variables

If the regression analysis is known, then our best estimate for any observation for the dependent variable would be the mean, . Alternatively, instead of using the mean as an estimate of Y_i, we can predict an estimate using the regression equation. The resulting solution will be denoted as:

$$Y_i = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_k X_k i + \varepsilon_i = \hat{Y}_i + \hat{\epsilon}_i$$

So that:

$$Y_i = \hat{Y}_i + \hat{\epsilon}_i$$

Now if we subtract the mean of the dependent variable in the above equation and square and sum on both sides so that:

$$\begin{aligned}\sum_{i=1}^{n} (Y_i - \bar{Y})^2 &= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y} + \hat{\epsilon}_i)^2 \\ &= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n} \hat{\epsilon}_i (\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n} \hat{\epsilon}_i^2\end{aligned}$$

Note that:

$$2\sum_{i=1}^{n} \hat{\epsilon}_i (\hat{Y}_i - \bar{Y}) = 0$$

Since the sample correlation between $\hat{Y}_i$ and $\hat{\epsilon}_i$ is 0. The expression, therefore, reduces to,

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

But

$$\hat{\epsilon}_i^2 = (Y_i - \hat{Y})^2$$

So, that

$$\sum_{i=2}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

Therefore,

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

If the regression analysis is useful for predicting Y_i using the regression equation, then the error should be smaller than predicting Y_i using the mean.

Now let:

Explained Sum of Squares (ESS)$=\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

Residual Sum of Squares (RSS) $=\sum_{i=1}^{n} (Y_i - \hat{Y})^2$

Total Sum of Squares (TSS)$=\sum_{i=1}^{n} (Y_i - \bar{Y})^2$

Then:

$$TSS = ESS + RSS$$

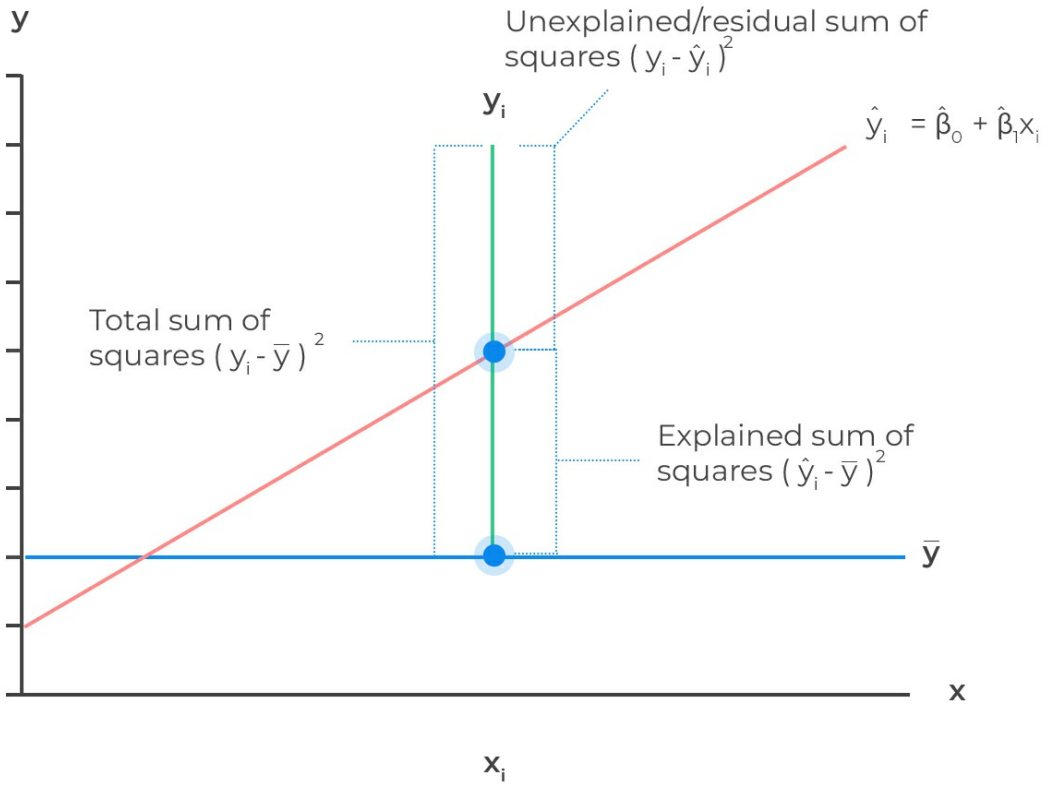
If we divide both sides by TSS, we get:

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$
$$\Rightarrow \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Now, recall than the coefficient of determination is the fraction of the overall change that is reflected in the regression. Denoted by R^2, coefficient of determination is given by:

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Ordinary Least Squares



If a model does not explain any of the observed data, then it has an R^2 of 0 . On the other hand, if the model perfectly describes the data, then it has an R^2 of 1. Other values are in the range of 0 and 1 and are always positive. For instance, in the above example, the R^2 is approximately 1 and thus, the money supply growth rate perfectly explains the level of inflation rates in the countries.

Limitations of R^2

1. As the number of explanatory variables increases, the value of R^2 always increases even if the new variable is almost completely irrelevant to the dependent variable. For instance, if a regression model with one explanatory variable is modified to have two explanatory variables, the new R^2 is greater or equal to that of a single explanatory model. In the case where $\beta = 0$, adding a variable will not increase R^2. In that case, the RSS will remain the same and so does R^2.

2. The Coefficient of Determination R^2 cannot be compared in different dependent variables. For instance, we cannot compare the R^2 for Y_i and $\ln Y_i$.
3. There is no standard value of R^2 that is considered good because its values depend on the nature of the data involved.

Considering the first limitation, we now discuss the adjusted R^2.

The Adjusted R^2

Denoted by $\bar{R}^2$, the adjusted-R^2 measures the goodness of fit, which does not automatically increase if an independent variable is added to the model; that is, it is adjusted for the degrees of freedom. Note that $\bar{R}^2$ is produced by statistical software. The relationship between the R^2 and $\bar{R}^2$ is given by:

$$\bar{R}^2 = 1 - \frac{(\frac{RSS}{n-k-1})}{(\frac{TSS}{n-1})}$$

$$= 1 - (\frac{n-1}{n-k-1})(1 - R^2)$$

Where

n=number of observations

k=number of the independent variables (Slope coefficients)

The adjusted R-squared can increase, but that happens only if the new variable improves the model more than would be expected by chance. If the added variable improves the model by less than expected by chance, then the adjusted R-squared decreases.

When $k \geq 1$, then $R^2 > \bar{R}^2$ since adding an extra new independent variable results in a decrease in $\bar{R}^2$ if that addition causes a small increase in R^2. This explains the fact that $\bar{R}^2$ can be a negative though R^2 is always nonnegative.

A point to note is that when we decide to use $\bar{R}^2$ to compare the regression models, the dependent variable is defined the same way and that the sample size is the same as that of R^2.

The following are the factors to watch out for when guarding against applying the R^2 or the $\bar{R}^2$:

- An added variable doesn't have to be statistically significant just because the R^2 or the $\bar{R}^2$ has increased.
- It is not always true that the regressors are a true cause of the dependent variable, just because there is a high R^2 or $\bar{R}^2$.
- It is not necessary that there is no omitted variable bias just because we have a high R^2 or $\bar{R}^2$.
- It is not necessarily true that we have the most appropriate set of regressors just because we have a high R^2 or $\bar{R}^2$
- It is not necessarily true that we have an inappropriate set of regressors just because we have a low R^2 or $\bar{R}^2$.

$\bar{R}^2$ does not automatically indicate that regression is well specified due to its inclusion of a right set of variables since a high $\bar{R}^2$ could reflect other uncertainties in the data in the analysis. Moreover, $\bar{R}^2$ can be negative if the regression model produces an extremely poor fit.

Joint Hypothesis Test on Multiple Regression Parameters

Previously, we had conducted hypothesis tests on individual regression coefficients using the t-test. We need to perform a joint hypothesis test on the multiple regression coefficients using the F-test based on the F-statistic.

In multiple regression, we cannot test the null hypothesis that all the slope coefficients are equal to 0 using the t-test. This is because an individual test on the coefficient does not accommodate the effect of interactions among the independent variables (multicollinearity).

F-test (test of regression's generalized significance) determines whether the slope coefficients in

multiple linear regression are all equal to 0. That is, the null hypothesis is stated as $H_0 : \beta_1 = \beta_2 = \ldots = \beta_K = 0$ against the alternative hypothesis that at least one slope coefficient is not equal to 0.

To accurately compute the test statistic for the null hypothesis that the slope is equal to 0, we need to identify the following:

I. The Sum of Squared Residuals (SSR) given by:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

This is also called the residual sum of squares.

II.Explained Sum of Squares (ESS) given by:

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2$$

III. The total number of observations (n).

III. The number of parameters to be estimated. For example, in a regression analysis with one independent variable, there are two parameters: the slope and the intercept coefficients.

Using the above four requirements, we can determine the F-statistic. The F-statistic measures how effective the regression equation explains the changes in the dependent variable. The F-statistic is denoted by $F_{(\text{Number of slope parameters, n-(number of parameters))}}$. *For* instance, the F-statistic for multiple regression with two slope coefficients (and one intercept coefficient) is denoted as $F_{2,\ n-3}$. The value $n-3$ represents the degrees of freedom for the F-statistic.

The F-statistic is the ratio of the average regression sum of squares to the average amount of squared errors. The average regression sum of squares is the regression sum of squares divided by the number of slope parameters (**k**) estimated. The average sum of squared errors is the sum of squared errors divided by the number of observations (n) less a total number of parameters estimated ($(n - (k + 1))$). Mathematically:

$$F = \frac{\text{Average regression sum of squares}}{\text{The average sum of squared errors}}$$

$$= \frac{\frac{\text{Explained sum of squares}}{\text{ESS Slope parameters estimated}}}{\frac{\text{Sum of squared residuals (SSR)}}{n - \text{number of parameters estimated}}}$$

In this case, we are dealing with a multiple linear regression model with **k** independent variable whose F-statistic is given by:

$$F = \frac{\left(\frac{ESS}{k}\right)}{\left(\frac{SSR}{n-(k+1)}\right)}$$

In regression analysis output (ANOVA part), MSR and MSE are displayed as the first and the second quantities under the MSS (mean sum of the squares) column, respectively. If the overall regression's significance is high, then the ratio will be large.

If the independent variables do not explain any of the variations in the dependent variable, each predicted independent variable $\hat{Y}_i$) possess the mean value of the dependent variable (Y). Consequently, the regression sum of squares is 0 implying the F-statistic is 0.

So, how do we decide F-test? We reject the null hypothesis at α significance level if the computed F-statistic is greater than the upper α critical value of the F-distribution with the provided numerator and denominator degrees of freedom (F-test is always a one-tailed test).

Example: Conducting F-test

An analyst runs a regression of monthly value-stock returns on four independent variables over 48 months.

The total sum of squares for the regression is 360, and the sum of squared errors is 120.

Test the null hypothesis at a 5% significance level (95% confidence) that all the four independent variables are equal to zero.

Solution

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_4 = 0$$

Versus

$$H_1 : \beta_j \neq 0 (\text{at least one j is not equal to zero, j=1,2… k})$$

ESS = TSS – SSR = 360 – 120 = 240

The calculated test statistic:

$$F = \frac{(\frac{ESS}{k})}{(\frac{SSR}{n-(k+1)})} = \frac{\frac{240}{4}}{\frac{120}{43}} = 21.5$$

$F^{3}_{4,43}$ is approximately 2.59 at a 5% significance level.

Decision: Reject H_0.

Conclusion: **at least one** of the 4 independent variables is significantly different than zero.

Example: Calculating F-statistic and Conducting the F-test

An investment analyst wants to determine whether the natural log of the ratio of bid-offer spread to the price of a stock can be explained by the natural log of the number of market participants and the amount of market capitalization. He assumes a 5% significance level. The following is the result of the regression analysis.

	Coefficient	Standard Error	t-Statistic
Intercept	1.6959	0.2375	7.0206
Number of market participants	−1.6168	0.0708	−22.8361
Amount of Capitalization	−0.4709	0.0205	−22.9707

ANOVA	df	SS	MSS	F	Significance F
Regression	2	3,730.1534	1,865.0767	2,217.95	0.00
Residual	2,797	2,351.9973	0.8409		
Total	2,799	5,801.2051			

Residual standard error	0.9180
Multiple R-squared	0.6418
Observations	2,800

We are concerned with the ANOVA (Analysis of variance) results. We need to conduct F-test to determine the significance of regression analysis.

Solution

So, the hypothesis is stated as:

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = 0$$

vs

$$H_1 : \text{At least } 1\hat{\beta}_j \neq 0, \forall j = 1,2$$

There are two slope coefficients, k=2 (coefficients on the natural log of the number of market participants and the amount of market capitalization), which is degrees of freedom for the numerator of the F-statistic formula. For the denominator, the degrees of freedom are n- (k + 1) =2800-3= 2,797.

The sum of the squared errors is 2,351.9973, while the regression sum of squares is 3,730.1534. Therefore, the F-statistic is:

$$F_{2,2797} = \frac{(\frac{ESS}{k})}{(\frac{SSR}{n-(k+1)})} = \frac{\frac{3730.1534}{2}}{\frac{2351.9973}{2797}} = 2217.9530$$

Since we are working at a 5% (0.05) significance level, we look at the F-distribution table on the second column which displays the F-distributions with degrees of freedom in the numerator of the F-statistic formula as seen below:

F Distribution: Critical Values of F (5% significance level)

	1	2	3	4	5	6	7	8	9	10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84

As seen from the table, the critical value of the F-test for the null hypothesis to be rejected is between 3.00 and 3.07. The actual F-statistic is 2217.95, which is far higher than the F-test critical value, and thus we reject the null hypothesis that all the slope coefficients are equal to 0.

Calculating the Confidence Interval for the Regression Coefficient

Confidence interval (CI) is a closed interval in which the actual parameter is believed to lie with some degree of confidence. Confidence intervals are used to perform hypothesis tests. For instance, we may want to ascertain stock valuation using the capital asset pricing model (CAPM). In this case, we may wish to hypothesize that the beta possesses the market's systematic risk or averaged beta.

The same analogy used in the regression analysis with one explanatory variable is also used in a multiple regression model using the t-test.

Example: Calculating the Confidence Interval (CI)

An economist tests the hypothesis that interest rates and inflation can explain GDP growth in a country. Using some 73 observations, the analyst formulates the following regression equation:

$$\text{GDP growth} = \hat{b}_0 + \hat{b}_1(\text{Interest}) + \hat{b}_2(\text{Inflation})$$

The regression estimates are as follows:

	Coefficient	Standard Error
Intercept	0.04	0.6%
Interest rates	0.25	6%
Inflation	0.20	4%

What is the 95% confidence interval for the coefficient on the inflation rate?

A. 0.12024 to 0.27976

B. 0.13024 to 0.37976

C. 0.12324 to 0.23976

D.0.11324 to 0.13976

Solution

The correct answer is A

From the regression analysis, $\hat{\beta} = 0.20$ and the estimated standard error, $s_{\hat{\beta}} = 0.04$. The number of

degrees of freedom is 73-3=70. So, the t-critical value at the 0.05 significance level is $= t_{\frac{0.05}{2},73-2-1} = t_{0.025,70} = 1.994$. Therefore, the 95% confidence level for the stock return is:

$$\hat{\beta} \pm t_c s_{\hat{\beta}} = 0.2 \pm 1.994 \times 0.04 = [0.12024, 0.27976]$$

Practice Questions

Question 1

An analyst runs a regression of monthly value-stock returns on four independent variables over 48 months. The total sum of squares for the regression is 360 and the sum of squared errors is 120. Calculate the R^2.

A. 42.1%

B. 50%

C. 33.3%

D. 66.7%

The correct answer is **D**.

$$R^2 = \frac{ESS}{TSS} = \frac{360 - 120}{360} = 66.7$$

Question 2

Refer to the previous problem and calculate the adjusted R^2.

A. 27.1%

B. 63.6%

C. 72.9%

D. 36.4%

The correct answer is **B**.

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{n-1}{n-k-1} \times (1 - R^2) \\ &= 1 - \frac{48-1}{48-4-1} \times (1 - 0.667) \\ &= 63.6\%\end{aligned}$$

Question 3

Refer to the previous problem. The analyst now adds four more independent variables to the regression and the new R^2 increases to **69%**. What is the new adjusted R^2 and which model would the analyst prefer?

A. The analyst would prefer the model with four variables because its adjusted R^2 is higher.

B. The analyst would prefer the model with four variables because its adjusted R^2 is lower.

C. The analyst would prefer the model with eight variables because its adjusted R^2 is higher.

D. The analyst would prefer the model with eight variables because its adjusted R^2 is lower.

The correct answer is **A**.

$$\text{New } R^2 = 69\%$$

$$\text{New adjusted } R^2 = 1 - \frac{48-1}{48-8-1} \times (1 - 0.69) = 62.6\%$$

The analyst would prefer the first model because it has a higher adjusted R^2 and the model has four independent variables as opposed to eight.

Question 4

An economist tests the hypothesis that GDP growth in a certain country can be explained by interest rates and inflation.

Using some 30 observations, the analyst formulates the following regression equation:

$$\text{GDP growth} = \hat{\beta}_0 + \hat{\beta}_1 \text{Interest} + \hat{\beta}_2 \text{Inflation}$$

Regression estimates are as follows:

	Coefficient	Standard Error
Intercept	0.10	0.5%
Interest Rates	0.20	0.05
Inflation	0.15	0.03

Is the coefficient for interest rates significant at 5%?

A. Since the test statistic < t-critical, we accept H_0; the interest rate coefficient is ***not*** significant at the 5% level.

B. Since the test statistic > t-critical, we reject H_0; the interest rate coefficient is ***not*** significant at the 5% level.

C. Since the test statistic > t-critical, we reject H_0; the interest rate coefficient is significant at the 5% level.

D. Since the test statistic < t-critical, we accept H_1; the interest rate coefficient is significant at the 5% level.

The correct answer is **C**.

We have GDP growth = 0.10 + 0.20(Int) + 0.15(Inf)

Hypothesis:

$$H_0 : \hat{\beta}_1 = 0 \quad \text{vs} \quad H_1 : \hat{\beta}_1 \neq 0$$

The test statistic is:

$$t = \left(\frac{0.20 - 0}{0.05}\right) = 4$$

The critical value is $t_{(\alpha/2,\ n-k-1)} = t_{0.025,27} = 2.052$ (which can be found on the t-table).

df/p	0.40	0.25	0.10	0.05	0.025	0.01
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202

Decision: Since test statistic > t-critical, we reject H_0.

Conclusion: The interest rate coefficient is significant at the 5% level.

Reading 20: Regression Diagnostics

After completing this reading, you should be able to:

- Explain how to test whether regression is affected by heteroskedasticity.
- Describe approaches to using heteroskedastic data.
- Characterize multicollinearity and its consequences; distinguish between multicollinearity and perfect collinearity.
- Describe the consequences of excluding a relevant explanatory variable from a model and contrast those with the consequences of including an irrelevant regressor.
- Explain two model selection procedures and how these relate to the bias-variance tradeoff.
- Describe the various methods of visualizing residuals and their relative strengths.
- Describe methods for identifying outliers and their impact.
- Determine the conditions under which OLS is the best linear unbiased estimator.

Regression Model Specifications

Model specification is a process of determining which independent variables should be included in or excluded from a regression model.

That is, an ideal regression model should consist of all the variables that explain the dependent variables and remove those that do not.

Model specification includes the residual diagnostics and the statistical tests on the assumptions of OLS estimators. Basically, the choice of variables to be included in a model depends on the bias-variance tradeoff. For instance, large models that include the relevant number of variables are likely to have unbiased coefficients. On the other side, smaller models lead to accurate estimates of the impact of removing some variables.

The conventional specification makes sure that the functional form of the model is adequate, the

parameters are constant, and the homoscedasticity assumption is met.

The Omitted Variables

An omitted variable is one with a non-zero coefficient, but they are excluded in the regression model.

Effects of Omitting Variables

I. The remaining variables sustain the impact of the excluded variables in terms of the common variation. Thus, they do not consistently approximate the change in the independent variable on the dependent variable while keeping all other things constant.
II. The magnitude of the estimated residuals is larger than the true value. This is true since the estimated residuals have the true value and the effect of the omitted value that cannot be reflected in the included variables.

Illustration of the Omitted Variables

Suppose that the regression model is stated as:

$$Y_i = \alpha + \beta_i X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

If we omit X_2 from the estimated model, then the model is given by:

$$Y_i = \alpha + \beta_i X_{1i} + \epsilon_i$$

Now, in large samples sizes, the OLS estimator $\hat{\beta}_1$ converges to:

$$\beta_1 + \beta_2 \delta$$

Where:

$$\delta = \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

δ is the population slope coefficient in a regression of X_2 on X_1.

It is clear that the bias – due to the omitted variable – depends on the population coefficient of the excluded variable β_2 and the relational strength of the X_2 and X_1, represented by δ.

When the correlation between X_1 and X_2 is high, X_1 can explain a significant proportion of variation in X_2 and hence the bias is high. On the other hand, if the independent variables are uncorrelated, that is $\delta = 0$ then $\hat{\beta}_1$ is a consistent estimator of β_1.

Conclusively, the omitted variable leads to biasness of the coefficient on the variables that are correlated with the omitted variables.

Inclusion of Extraneous Variables

An extraneous variable is one that is unnecessarily included in the model, whose actual coefficient is 0 and is consistently estimated to be 0 in large samples. If we include these variables is costly.

Illustration of Effect of Inclusion of Extraneous Random Variables

Recall that the adjusted R^2 is given by:

$$\bar{R}^2 = 1 - \xi \frac{RSS}{TSS}$$

Where:

$$\xi = \frac{(n-1)}{(n-k-1)}$$

Looking at the formula above, adding more variables increase the value of k which in turn increases the value of ξ and hence reducing the value of $\bar{R}^2$. However, if the model is large, then RSS is smaller which reduces the effect of ξ and produces larger $\bar{R}^2$.

Contrastingly, this is not always the case when the true coefficient is equal to 0 because, in this

case, RSS remains constant as ξ increases leading to a smaller $\bar{R}^2$ and a large standard error.

Lastly, if the correlation between X_1 and X_2 increases, the standard error value rises.

The Bias-Variance Tradeoff

The bias-variance tradeoff amounts to choosing between the including irrelevant variables and excluding relevant variables. Bigger models tend to have low bias level because it includes more relevant variables. However, they are less accurate in approximating the regression parameters due to the possibility of involving extraneous variables.

Moreover, regression models with fewer independent variables are characterized by low estimation error but more prone to biased parameter estimates.

Methods of Choosing a Model from a Set of Independent Variables

1. **General-to-Specific Model Selection**

   In the general-to-specific method, we start with a large general model that incorporates all the relevant variables. Then, the reduction of the general model starts. We use hypothesis tests to establish if there are any statistically insignificant coefficients in the estimated model. When such coefficients are found, the variable with the coefficient with the smallest t-statistic is removed. The model is then re-estimated using the remaining set of independent variables. Once more, hypothesis tests are carried out to establish if statistically insignificant coefficients are present. These two steps (remove and re-estimate) are repeated until all coefficients that are statistically insignificant have been removed.

2. **m-fold Cross-Validation**

   The m-fold cross-validation model-selection method aims at choosing the model that's best at fitting observations not used to estimate parameters.

How is this method executed?

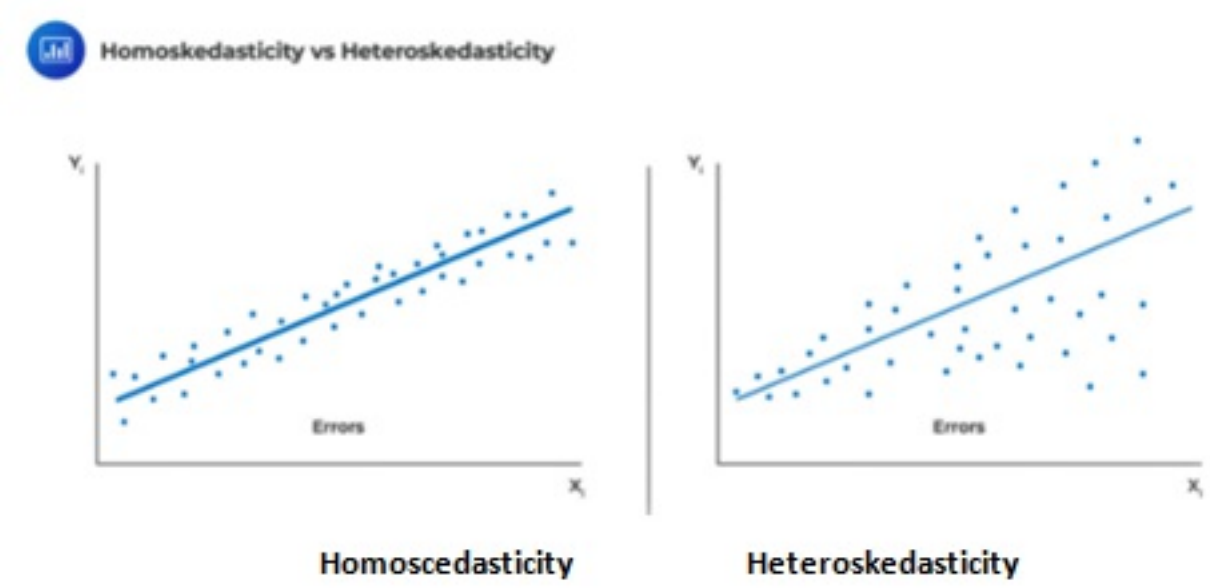
As a first step, the number of models has to be decided, and this is determined in part by the number of explanatory variables. When this number is small, the researcher can consider all the possible combinations. With 10 variables, for example, 1,024 (=) distinct models can be constructed.

The cross-validation process proceeds as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into *m* groups.
3. Estimate parameters using m-1 of the groups; these groups make up what we call the **training block**. The excluded group is referred to as the **validation block.**
4. Use the estimated parameters and the data in the excluded block (validation block) to compute residual values. These residuals are referred to as out-of-sample residuals since they are arrived at using data not included in the sample used to come up with the parameter estimates.
5. Repeat parameter estimation and residual computation a total of *m* times; each group has to serve as the validation block and used to compute residuals.
6. Compute the sum of squared errors using the residuals estimated from the out-of-sample data.
7. Select the model with the smallest out-of-sample sum of squared residuals.

Heteroskedasticity

Recall that homoskedasticity is one of the critical assumptions in the determination of the distribution of the OLS estimator. That is, the variance of ϵ_i is constant and that it does not vary with any of the independent variables; formally stated as $Var(\epsilon_i \left| X_{1i}, X_{2i}, \dots, X_{ki}) = \delta^2$. **Heteroskedasticity** is a **systematic** pattern in the residuals where the variances of the residuals are **not constant**.



Test for Heteroskedasticity

Halbert White proposed a simple test, with the following two-step procedures:

I. Approximate the model and calculate the residuals, ϵ_i
II. Regress the **squared** residuals on:
    1. A constant
    2. All explanatory variables
    3. The cross product of all the independent variables, including the product of each variable with itself.

Consider an original model with two independent variables:

$$Y_i = \alpha + \beta_i X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

The first step is to calculate the residuals by utilizing the OLS parameter estimators:

$$\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$$

Now, we need to regress the squared residuals on a constant X_1, X_2, X_1^2, X_2^2 and $X_1 X_2$

$$\hat{\epsilon}_i^2 = Y_0 + Y_1 X_{1i} + Y_2 X_{2i} + Y_3 X_{1i}^2 + Y_4 X_{2i}^2 + Y_5 X_{1i}^2 X_{2i}^2$$

If the data is homoscedastic, then $\hat{\epsilon}_i^2$ must not be explained by any of the variables and the null hypothesis is: $H_0 : Y_1 = \cdots = Y_5 = 0$

The test statistic is calculated as nR^2 where R^2 is calculated in the second regression and that the test statistic has a $\chi^2_{\frac{k(k+3)}{2}}$ (chi-distribution), where k is the number of explanatory variables in the first-step model.

For instance, if the number of the explanatory variables is two, k=2, then the test statistic has a distribution of χ_5.

Modeling Heteroskedastic Data

The three common methods of handling data with heteroskedastic shocks include:

1. **Ignoring the heteroskedasticity when approximating the parameters and then utilizing the White covariance estimator in hypothesis tests.**

   However simple, this method leads to less accurate model parameter estimates compared to other methods that address the heteroskedasticity.

2. **Transformation of data.**

   For instance, positive data can be log-transformed to try and remove heteroskedasticity and give a better view of data. Another transformation can be in the form of dividing the dependent variable by another positive variable.

3. **Use of weighted least squares (WLS).**

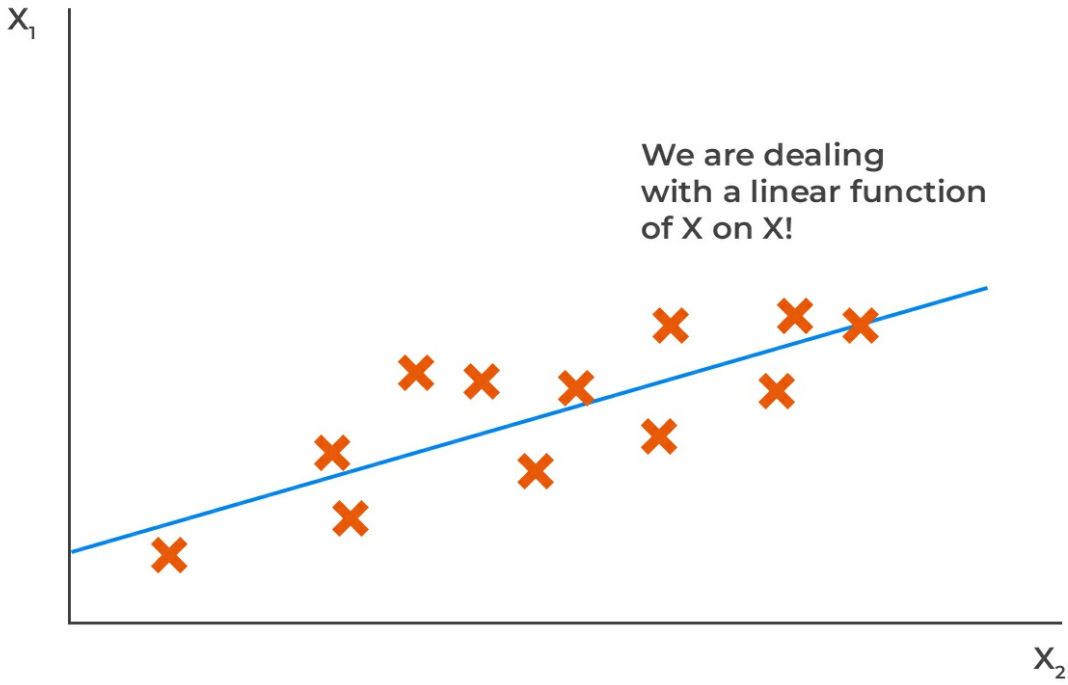
   This is a complicated method that applies weights to the data before approximating the parameters. That is if we know that $Var(\epsilon_i) = w_i^2 \sigma^2$ where w_i is known then we can transform the data by dividing by w_i to remove the heteroskedasticity from the errors. In other words, the WLS regresses $\frac{Y_i}{w_i}$ on $\frac{X_i}{w_i}$ such as:

$$\frac{Y_i}{w_i} = \alpha\frac{1}{w_i} + \beta\frac{X_i}{w_i} + \frac{\epsilon_i}{w_i}$$
$$\bar{Y}_i = \alpha\bar{C}_i + \beta\bar{X}_i + \bar{\epsilon}_i$$

Note that the parameters of the model above are estimated using OLS on the transformed data. That is, the weighted version of Y_i which is $\bar{Y}_i$ on two weighted explanatory variables $\bar{C}_i = \frac{1}{w_i}$ and $\bar{X}_i = \frac{X_i}{w_i}$. Note that the WLS model does not clearly include the intercept α, but the interpretation is still the same, that is, the intercept.

Multicollinearity

Multicollinearity occurs when others can significantly explain one or more independent variables. For instance, in the case of two independent variables, there is evidence of multicollinearity if the R^2 is very high if one variable is regressed on the other.



In contrast with multicollinearity, perfect correlation is where one of the variables is perfectly correlated to others such that the R^2 of regression of X_j on the remaining independent variable is precisely 1.

Conventionally, when R^2 is above 90% leads to problems in medium sample sizes such as that of 100. Multicollinearity does not pose an issue in parameter approximation, but rather, it brings some difficulties in modeling the data.

When multicollinearity is present, some of the coefficients in a regression model are jointly statistically significant (F-statistic is substantial), but the individual t-statistic is very small (less than 1.96) since the regression analysis assumes the collective effect of the variables rather than the individual effect of the variables.

Addressing Multicollinearity

There are two ways of dealing with multicollinearity:

I. Ignoring multicollinearity altogether since it technically not a problem.
II. Identification of the multicollinear variables and excluding them from the model. Identification of multicollinear variables using the variance inflation factor which compares the variance of the regression coefficients on independent variable X_j in two models: one that incorporates only X_j and one that omits k independent variables:

$$X_{ji} = Y_0 + Y_1 X_{1i} + \cdots + Y_{j-1} X_{j-1i} + Y_{j+1} X_{j+1i} + \cdots + Y_k X_{ki} + \eta_i$$

The variance inflation factor (VIF) for the variable X_j is given by:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where R_j^2 originates from regressing X_j on the other variable in the model. When the value of the VIF is above 10, then it is considered too much and the variable should be excluded from the model.

Residual Plots

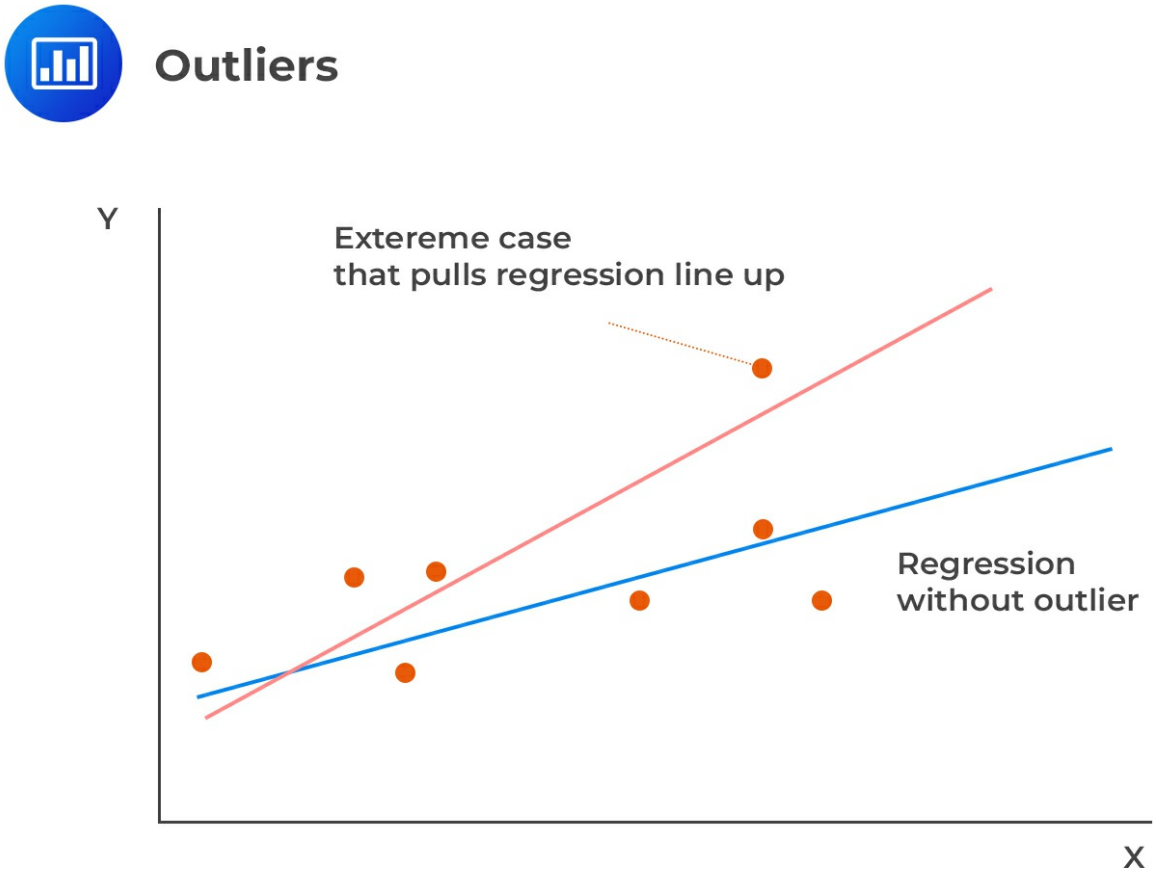
Residual plots are utilized to identify the deficiencies in a model specification. When the residual plots are not systematically related to any of the included independent (explanatory variables) and relatively small (within ±4s, where s, is the standard shock deviation of the model) in magnitude, then the model is ideally good.

Residual plot is a graph of $\hat{\epsilon}_i$ (vertical axis) against the independent variables x_i. Alternatively, we could use the standardized residuals $\frac{\hat{\epsilon}_i}{s}$ which makes sure that the deviation is apparent.

Outliers

Outliers are values that, if removed from the sample, produce large changes in the estimated

coefficients. They can also be viewed as data points that **deviate significantly** from the normal objects as if they were **generated by a different mechanism.**



Cook's distance helps us measure the impact of dropping a single observation j on a regression (and the line of best fit).

The Cook's distance is given by:

$$D_j = \frac{\sum_{i=1}^{n} (\bar{Y}_i^{(-j)} - \hat{Y}_i)^2}{ks^2}$$

Where:

$\bar{Y}_i^{(-j)}$=fitted value of $\bar{Y}_i$ when the observed value j is excluded, and the model is approximated using n-1 observations.

k=number of coefficients in the regression model

s^2=estimated error variance from the model using all observations

When a variable is an inline (does not affect the coefficient estimates when excluded), the value of its Cook's distance (D_j) is small. On the other hand, D_j is higher than 1 if it is an outlier.

Example: Calculating Cook's Distance

Consider the following data sets:

Observation	Y	X
1	3.67	1.85
2	1.88	0.65
3	1.35	−0.63
4	0.34	1.24
5	−0.89	−2.45
6	1.95	0.76
7	2.98	0.85
8	1.65	0.28
9	1.47	0.75
10	1.58	−0.43
11	0.66	1.14
12	0.05	−1.79
13	1.67	1.49
14	−0.14	−0.64
15	9.05	1.87

If you look at the data sets above, it is easy to see that observation 15 is quite more significant than the rest of the observations, and there is a possibility to be an outlier. However, we need to ascertain this.

We begin by fitting the whole dataset ($\bar{Y}_i$) and then the 14 observations which remain after excluding the dataset that we believe is an outlier.

If we fit the whole dataset, we get the following regression equation:

$$\bar{Y}_i = 1.4465 + 1.1281X_i$$

And if we exclude the observation that we believe it is an outlier we get:

$$\bar{Y}_i^{(-j)} = 1.1516 + 0.6828X_i$$

Now the fitted values are as shown below:

Observation	Y	X	$\bar{Y}_i$	$\bar{Y}_i^{(-j)}$	$(\bar{Y}_i^{(-j)} - \bar{Y}_i)^2$
1	3.67	1.85	3.533	2.4148	1.2504
2	1.88	0.65	2.179	1.5954	0.3406
3	1.35	0.63	0.7358	0.7214	0.0002
4	0.34	1.24	2.8453	1.9983	0.7174
5	0.89	2.45	−1.3174	−0.5213	0.6338
6	1.95	0.76	2.3039	1.6705	0.4012
7	2.98	0.85	2.4053	1.732	0.4533
8	1.65	0.28	1.7624	1.3428	0.1761
9	1.47	0.75	2.2926	1.6637	0.3955
10	1.58	0.43	0.9614	0.858	0.0107
11	0.66	1.14	2.7325	1.921	0.6585
12	0.05	1.79	−0.5728	−0.07061	0.2522
13	1.67	1.49	3.1274	2.169	0.9185
14	0.14	0.64	0.7245	0.7146	0.0001
15	9.05	1.87	3.556	2.4284	1.2715
				Sum	7.4800

If the $s^2 = 3.554$ the Cook's distance is given by:

$$D_j = \frac{\sum_{i=1}^{n} (\bar{Y}_i^{(-j)} - \hat{Y}_i)^2}{ks^2} = \frac{7.4800}{2 \times 3.554} = 1.0523$$

Since $D_j > 1$, then observation 15 can be considered as an outlier.

Strengths of Ordinary Least Squares (OLS)

OLS is the Best Linear Unbiased Estimator (BLUE) when some key assumptions are met, which implies that it can assume the smallest possible variance among any given estimator that is linear and unbiased:

- **Linearity**: the parameters being estimated using the OLS method must be themselves linear.

- **Random**: the data must have been randomly sampled from the population.
- **Non-Collinearity**: the regressors being calculated should not be perfectly correlated with each other.
- **Exogeneity**: the regressors aren't correlated with the error term.
- **Homoscedasticity**: the variance of the error term is constant

However, being a BLUE estimator comes with the following limitations:

I. A big proportion of the estimators are not linear such as maximum likelihood estimators (but biased).
II. BLUE property is heavily dependent on residuals being homoskedastic. In the case that the variances of residuals vary the independent variables, then it is possible to construct linear unbiased estimators (LUE) of the coefficients α and β using WLS but with extra assumptions.

When the residuals and iid and normally distributed with a mean of 0 and variance of σ^2, formally stated as $\epsilon_i \sim^{iid} N(0, \sigma^2)$ makes the upgrades BLUE to BUE (Best Unbiased Estimator) by virtue having the smallest variance among all linear and non-linear estimators. However, errors being normally distributed is not a requirement for accurate estimates of the model coefficients or a necessity for desirable properties of estimators.

Practice Question 1

Which of the following statements is/are correct?

I. Homoskedasticity means that the variance of the error terms is constant for all independent variables.

II. Heteroskedasticity means that the variance of error terms varies over the sample.

III. The presence of conditional heteroskedasticity reduces the standard error.

A. Only I

B. II and III

C. All statements are correct

D. None of the statements are correct

Solution

The correct answer is **C**.

All statements are correct

If the variance of the residuals is constant across all observations in the sample, the regression is said to be homoskedastic. When the opposite is true, the regression is said to exhibit heteroskedasticity, i.e., the variance of the residuals is not the same across all observations in the sample. The presence of conditional heteroskedasticity poses a significant problem: it introduces a bias into the estimators of the standard error of the regression coefficients. As such, it understates the standard error.

Practice Question 2

A financial analyst fails to include a variable which inherently has a non-zero coefficient in his regression analysis. Moreover, the ignored variable is highly correlated with the

remaining variables.

What is the most likely deficiency of the analyst's model?

A. Omitted variable bias.

B. Bias due to inclusion of extraneous variables.

C. Presence of heteroskedasticity.

D. None of the above.

Solution

The correct answer is **A**.

Ommitted variable bias occurs under two conditions:

I. A variable with a non-zero coefficient is omitted

II. A variable that is omitted is correlated with remaining (included) variables.

These conditions are met in the description of the analyst's model.

Option B is incorrect since an extraneous variable is one that is unnecessarily included in the model, whose true coefficient and consistently approximated value is 0 in large sample sizes.

Option C is incorrect because heteroskedasticity is a condition where the variance of the errors varies systematically with the independent variables of the model.

Reading 21: Stationary Time Series

After completing this reading, you should be able to:

- Describe the requirements for a series to be covariance stationary.
- Define the autocovariance function and the autocorrelation function.
- Define white noise; describe independent white noise and normal (Gaussian) white noise.
- Define and describe the properties of autoregressive (AR) processes.
- Define and describe the properties of moving average (MA) processes.
- Explain how a lag operator works.
- Explain mean reversion and calculate a mean-reverting level.
- Define and describe the properties of autoregressive moving average (ARMA) processes.
- Describe the application of AR, MA, and ARMA processes.
- Describe sample autocorrelation and partial autocorrelation.
- Describe the Box-Pierce Q-statistic and the Ljung-Box Q statistic.
- Explain how forecasts are generated from ARMA models.
- Describe the role of mean reversion in long-horizon forecasts.
- Explain how seasonality is modeled in a covariance-stationary ARMA.

Time series is a collection of observations on a variable's outcome in distinct periods — for example, monthly sales of a company for the past ten years. Time series are used to forecast the future of the time series. The time series are classified into the trend, seasonal, and cyclical components. A trend time-series changes its level over time, while a seasonal time series has predictable changes over a given time. Lastly, a cyclical time series, as its name suggests, reflects the cycles in a given data. We will concentrate on the cyclical data (especially linear stochastic processes).

A stochastic process is a set of variables. The stochastic process is mostly denoted by Y_t and by the subscript, the random variable is ordered time so that Y_s occurs first before Y_t if $s < t$.

A linear process has a general form of:

$$\begin{aligned} Y_t &= \alpha_t + \beta_0 \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + \dots \\ &= \alpha_t + \sum_{i=0}^{\infty} \beta_i \epsilon_{t-i} \end{aligned}$$

The linear process is linear on the shock, ϵ_t. α_t is a deterministic while β_i is a constant coefficient.

Covariance Stationary Time Series

The ordered set: $\{\dots, y_{-2}, y_{-1}, y_0, y_1, y_2, \dots\}$ is called the realization of a time series. Theoretically, it starts from the infinite past and proceeds to the infinite future. However, only a finite subset of realization can be used in practice, and is called a sample path.

A series is said to be covariance stationary if both its mean and covariance structure is stable over time.

More specifically, a time series is said to be covariance stationary if:

I. The mean does not change and thus constant over time. That is:

$$E(Y_t) = \mu \forall t$$

II. The variance does change over time, and it is constant. That is:

$$V(Y_t) = \gamma_0 < \infty \forall t$$

III. The autocovariance of the time series is finite and does not change over time, and it depends on the distance between two observations. That is:

$$Cov(Y_t, Y_{t-h}) = \gamma_h \forall t$$

The covariance stationarity is crucial so that the time series has a constant relationship across time

and that the parameters are easily interpreted since the parameters will be asymptotically normally distributed.

Autocovariance and Autocorrelation Functions

The Autocovariance Function

It can be quite challenging to quantify the stability of a covariance structure. We will, therefore, use the autocovariance function. The autocovariance is the covariance between the stochastic process at a different point in time (analogous to the covariance between two random variables). It is given by:

$$\gamma_{t,h} = E\left[(Y_t - E(Y_t))(Y_{t-h} - E(Y_{t-h}))\right]$$

And if the length $h = 0$ then:

$$\gamma_{t,h} = E\left[(Y_t - E(Y_t))^2\right]$$

Which is the variance of Y_t.

The autocovariance is a function of h so that:

$$\gamma_h = \gamma_{|h|}$$

This is asserting the fact that the autocovariance depends on the length h and not the time t. So that:

$$Cov(Y_t, Y_{t-h}) = Cov(Y_{t-h}, Y_t)$$

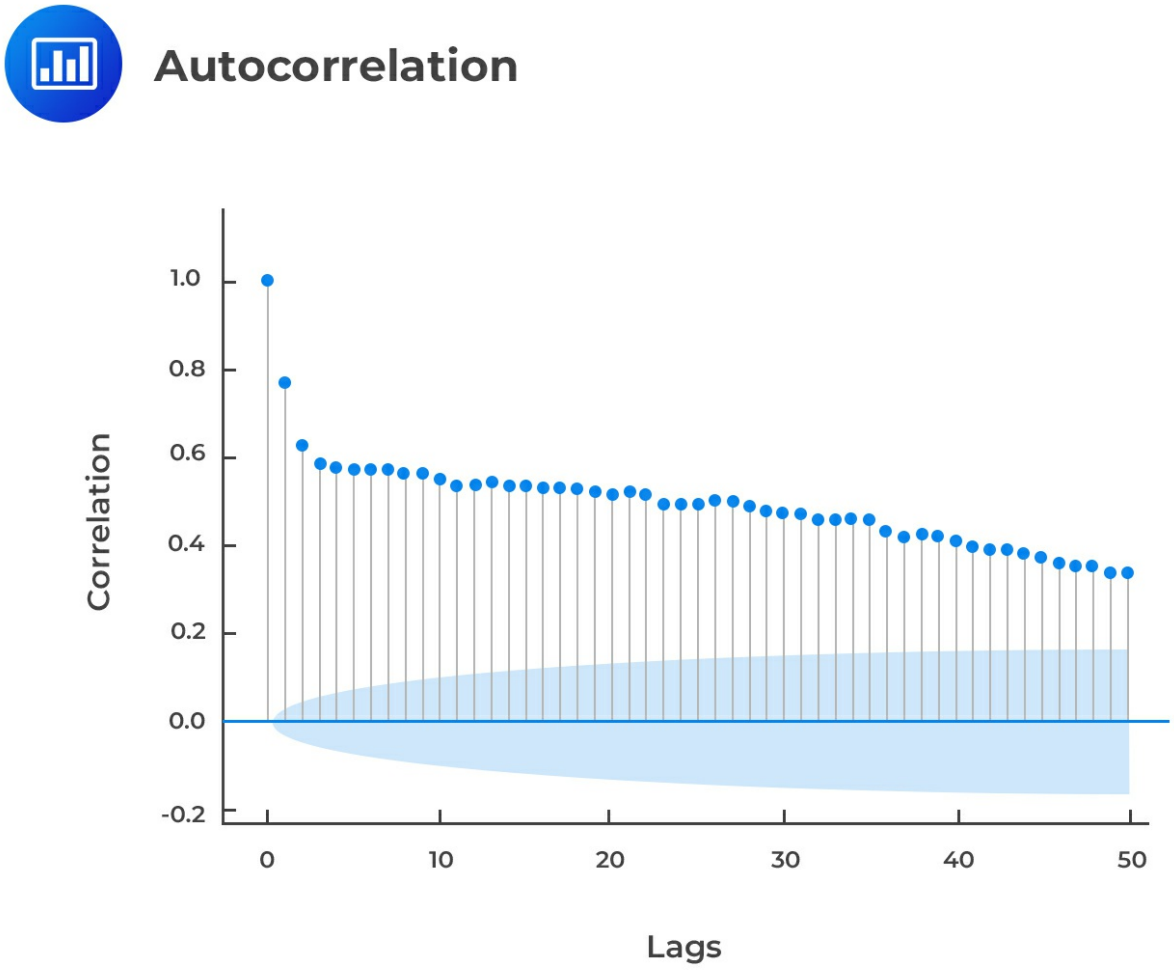
The Autcorrelation is defined as:

$$\rho(t) = \frac{Cov(Y_t, Y_{t_h})}{\sqrt{V(Y_t)}\sqrt{V(Y_{t-h})}} = \frac{\gamma_h}{\sqrt{\gamma_0\gamma_0}} = \frac{\gamma_h}{\gamma_0}$$

Similarly, for $h = 0$.

$$\rho(t) = \frac{\gamma_0}{\gamma_0} = 1$$

The autocorrelation ranges from -1 and 1 inclusively. The partial autocorrelation function is denoted as, p(h), and in a linear population regression of Y_t on $Y_{t-1}, \ldots, Y_{t-h}$, it is the coefficient of y_{t-h}. This regression is referred to as the autoregression. This is because the regression is on the lagged values of the variable.



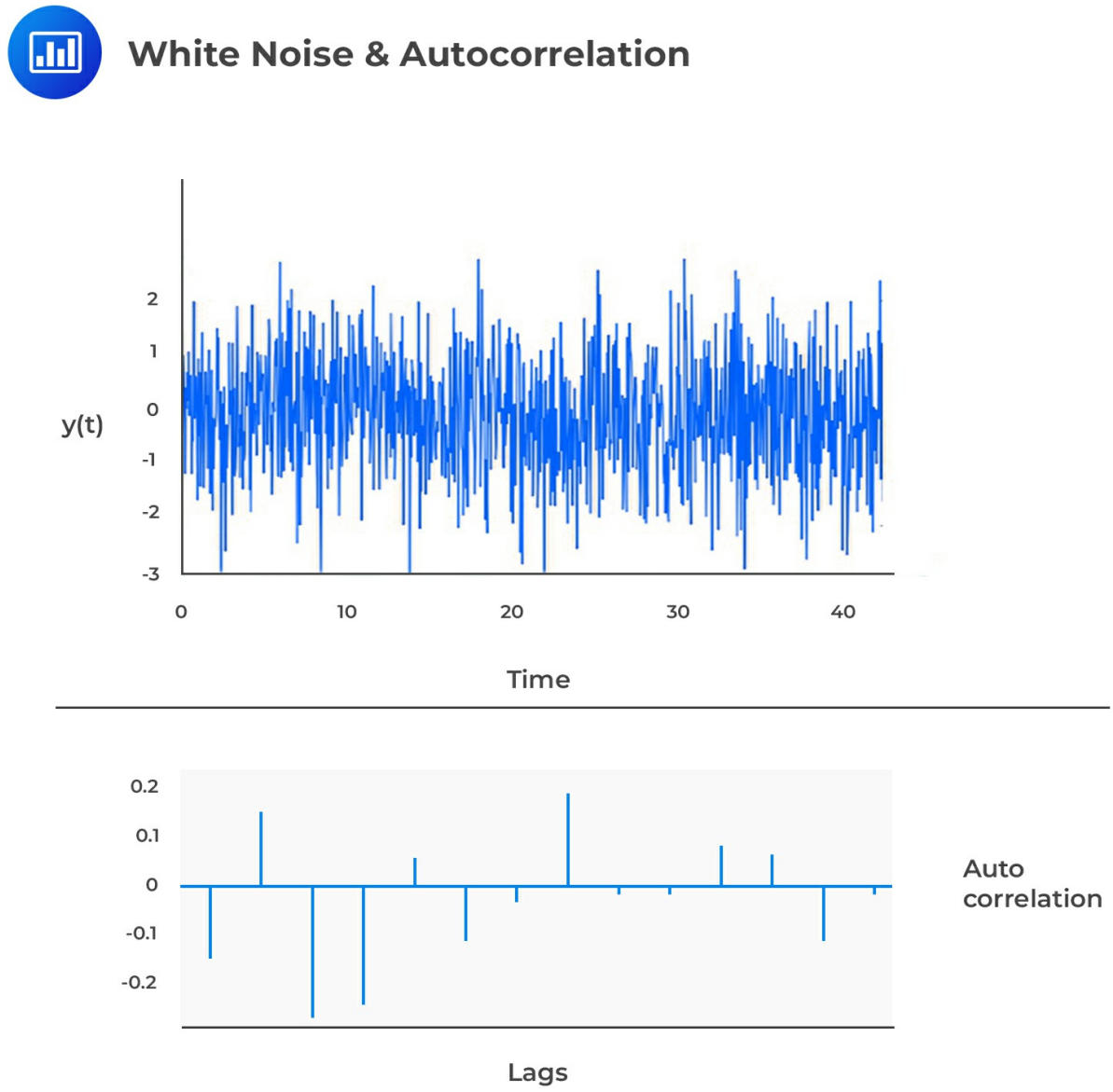
White Noise

Assume that:

$$y_t = \epsilon_t$$

$$\epsilon_t \sim (0, \sigma^2), \quad \forall \quad \sigma^2 < \infty$$

where ϵ_t is the shock and is uncorrelated over time. Therefore, ϵ_t and y_t are said to be serially uncorrelated.



This auto-correlation that has a zero mean and unchanging variance is referred to as the zero-mean white noise (or just white noise) and is written as:

$$\epsilon_t \sim WN\ (0, \sigma^2)$$

And:

$$y_t \sim WN(0, \sigma^2)$$

ϵ_t and y_t serially uncorrelated but not necessarily serially independent. If y possesses this property, (serially uncorrelated but not necessarily serially independent) then it is said to be an independent white noise.

Therefore, we write:

$$y_t \tilde{iid}(0, \sigma^2)$$

This is read as "y is independently and identically distributed with a mean 0 and constant variance. y is said to be serially independent if it is serially uncorrelated and it has a normal distribution. In this case, y is called the normal white noise or the Gaussian white noise.

Written as:

$$y_t \tilde{iid}N(0, \sigma^2)$$

To characterize the dynamic stochastic structure of $y_t \sim WN(0, \sigma^2)$, it follows that the unconditional mean and variance of y are:

$$E(y_t) = 0$$

And:

$$var(y_t) = \sigma^2$$

These two are constant since only displacement affects the autocovariances rather than time. All the autocovariances and autocorrelations are zero beyond displacement zero since white noise is uncorrelated over time.

The following is the autocovariance function for a white noise process:

$$\gamma(h) = \begin{cases} \sigma^2, & h = 0 \\ 0, & h \geq 0 \end{cases}$$

The following is the autocorrelation function for a white noise process:

$$\rho(h) = \begin{cases} 1, & h = 0 \\ 0, & h \geq 1 \end{cases}$$

Beyond displacement zero, all partial autocorrelations for a white noise process are zero. Thus, by construction white noise is serially uncorrelated. The following is the function of the partial autocorrelation for a white noise process:

$$p(h) = \begin{cases} 1, & h = 0 \\ 0, & h \geq 1 \end{cases}$$

Simple transformations of white noise are considered in the construction of processes with much richer dynamics. Then the white noise should be the 1-step-ahead forecast errors from good models.

The mean and variance of a process, conditional on its past, is another crucial characterization of dynamics with crucial implications for forecasting.

To compare the conditional and unconditional means and variances, consider the independence white noise: $y_t \tilde{iid}(0, \sigma^2)$. y has an unconditional mean and variance of zero and σ^2 respectively. Now, consider the transformational set:

$$\Omega_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$$

Or:

$$\Omega_{t-1} = \{\epsilon_{t-1}, \epsilon_{t-2}, \ldots\}$$

The conditional mean and variance do not necessarily have to be constant. The conditional mean for the independent white noise process is:

$$E(y_t|\Omega_{t-1}) = 0$$

The conditional variance is:

$$\text{var}(y_t|\Omega_{t-1}) = E((y_t - E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}) = \sigma^2$$

Independent white noise series have identical conditional and unconditional means and variances.

Wold's Theorem

Assuming that $\{y_t\}$ is any zero-mean covariance-stationary process. Then:

$$Y_t = \epsilon_t + \beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots = \sum_{i=0}^{\infty} \beta_i\epsilon_{t-i}$$

Where:

$$\epsilon_t \sim WN(0, \sigma^2)$$

Note that $\beta_0 = 1$ and $\sum_{i=0}^{\infty} \beta_i^2 < \infty$.

The accurate model for any stationary covariance series is the Wold's representation. Since ϵ_t corresponds to the 1-step-ahead forecast errors to be incurred should a particularly good forecast be applied, the ϵ_t's are the innovations.

Time-Series Models

The Autoregressive (AR) Models

AR models are time series models mostly used in finance and economics which links the stochastic process Y_t to the previous value Y_{t-1}. The first order AR model denoted by AR(1) is given by:

$$Y_t = \alpha + \beta Y_{t-1} + \epsilon_t$$

Where:

α = intercept

β = AR parameter

ϵ_t = the shock which is white noise ($\epsilon_t \sim WN(0, \sigma^2)$

SinceY_t is assumed to be covariance stationary, the mean,variance, and autocovariances are all constant. By the principle of covariance stationarity,

$$E(Y_t) = E(Y_{t-1}) = \mu$$

Therefore,

$$E(Y_t) = E(\alpha + \beta Y_{t-1} + \epsilon_t = \alpha + \beta E(Y_{t-1}) + E(\epsilon_t)$$

$$\Rightarrow \mu = \alpha + \beta\mu + 0$$

$$\therefore \mu = \frac{\alpha}{1-\beta}$$

And for the variance,

$$V(Y_t) = V(\alpha + \beta Y_{t-1} + \epsilon_t) = \beta^2 V(Y_{t-1}) + V(\epsilon_t) + Cov(Y_{t-1}, \epsilon_t)$$

$$\gamma_0 = \beta^2\gamma_0 + \sigma^2 + 0$$

$$\therefore \frac{\sigma^2}{1-\beta^2}$$

Note that $Cov(Y_{t-1}, \epsilon_t)$=0 since Y_{t-1} is uncorrelated with the shocks $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$

The Autocovariances for AR(1) process is calculated recursively. The first autocovariance for the AR(1) model is given by:

$$\begin{aligned} Cov(Y_t, Y_{t-1}) &= Cov(\alpha + \beta Y_{t-1} + \epsilon_t, Y_{t-1}) \\ &= \beta Cov(Y_t, Y_{t-1}) + Cov(Y_{t-1}, \epsilon_t) \\ &= \beta\gamma_0 \end{aligned}$$

The remaining autocovariance is recursively calculated as:

$$\begin{aligned} \text{Cov}(Y_t, Y_{t-h})) &= \text{Cov}(\alpha + \beta Y_{t-1} + \epsilon_t, Y_{t-h}) \\ &= \beta \text{Cov}(Y_{t-1}, Y_{t-h}) + \text{Cov}(Y_{t-h}, \epsilon_t) \\ &= \beta \gamma_{h-1} \end{aligned}$$

It should be easy to see that $\text{Cov}(Y_{t-h}, \epsilon_t) = 0$. Applying this recursion analogy:

$$\gamma_h = \beta^h \gamma_0$$

Therefore we can generalize the autocovariance as:

$$\gamma_h = \beta^{|h|} \gamma_0$$

Intuitively the autocorrelation function is given by:

$$\rho(\rho) = \frac{\beta^h \gamma_0}{\gamma_0} = \beta^{|h|}$$

The ACF tends to 0 when h increases and that -1<β<0. The Partial autocorrelation of an AR(1) model is given by:

$$\partial(h) = \begin{cases} \beta^{|h|}, h \in \{0, \pm\} \\ 0, h \geq 2 \end{cases}$$

The Lag Operator

The lag operator denoted by L is important for manipulating complex time-series models. As its name suggests, the lag operator moves the index of a particular observation one step back. That is:

$$LY_t = Y_{t-1}$$

Properties of the Lag Operator

(I). The lag operator moves the index of a time series one step back. That is:

$$LY_t = Y_{t-1}$$

(II). Consider the following mth-order lag operator polynomial Lm then:

$$L^m Y_t = y_{t-m}$$

For instance $L^2 Y_t = L(LY_t) = L(Y_{t-1}) = Y_{t-2}$

(III). The lag operator of a constant is just a constant.

For example $L\alpha = \alpha$

(IV). The p^{th} order lag operator is given by:

$$a(L) = 1 + a_1 L + a_2 L^2 + \ldots + a_p L^p$$

so that:

$$a(L)Y_t = Y_t + a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_p Y_{t-p}$$

(V). The lag operator has a multiplicative property. Consider two lag operators a(L) and b(L). Then:

$$\begin{aligned} a(L)b(L)Y_t &= (1 + a_1(L))\,(1 + b_1(L))\,Y_t \\ &= (1 + a_1(L))\,(Y_t + b_1 Y_{t-1}) \\ &= Y_t + b_1 Y_{t-1} + a_1 Y_{t-1} + a_1 b_1 Y_{t-2} \end{aligned}$$

Moreover, the lag operator has a commutative property so that:

$$a(L)b(L) = b(L)a(L)$$

IV. Under some restrictive conditions, the lag operator polynomial can be inverted so that: $a(L)a(L)^{-1} = 1$. When a(L) is a first-order lag operator polynomial given by $1 - a_1(L)$, is invertible if $|a_1| < 1$ so that its inverse is given by:

$$(1 - a_1(L))^{-1} = \sum_{i=1}^{\infty} a^i L^i$$

For an AR(1) model,

$$Y_t = \alpha + \beta Y_{t-1} + \epsilon_t$$

This can be expressed with the lag operator so that:

$$Y_t = \alpha + \beta(LY)_t + \epsilon_t$$

$$\Rightarrow (1 - \beta L)Y_t = \alpha + \epsilon_t$$

If $|\beta|<1$, then the lag polynomial above is invertible so that:

$$(1 - \beta L)^{-1}(1 - \beta L)Y_t = (1 - \beta L)^{-1}\alpha + (1 - \beta L)^{-1}\epsilon_t$$

$$\Rightarrow Y_t = \alpha \sum_{i=1}^{\infty} \beta^i + \sum_{j=1}^{\infty} \beta^j L^j \epsilon_t = \frac{\alpha}{1-\beta} + \sum_{i=1}^{\infty} \beta^i L^i \epsilon_{t-i}$$

The p^{th} Order Autoregressive Model (AR(p))

The AR(p) model is a generalization of the AR(1) model to include the p lags of Y_{t-1}. Thus, the AR(p) is given by:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \epsilon_t$$

If Y_t is covariance stationary, then the long-run mean is given by:

$$E(Y_t) = \frac{\alpha}{1 - \beta_1 - \beta_2 - \ldots \beta_p}$$

And the long-run variance is given by:

$$V(Y_t) = \gamma_0 = \frac{\sigma^2}{1 - \beta_1\rho_1 - \beta_2\rho_2 - \ldots \beta_p\rho_p}$$

From the formulas of the mean and variance of the AR(p) model, the covariance stationarity property is satisfied if:

$$\beta_1 + \beta_2 + \cdots + \beta_p < 1$$

Otherwise, the covariance stationarity will be violated.

The autocorrelations function of the AR(p) model bears the same structural model as AR(1) model; the ACF tends to 1 as the length between the two-time series increases and may oscillate. However, higher-order ARs may bear complex structures in their ACFs.

The Moving Average Models (MA)

The first-order moving average model denoted by MA(1) is given by:

$$Y_t = \mu + \theta\epsilon_{t-1} + \epsilon_t$$

Where $\epsilon_t \sim WN(0, \sigma^2)$.

Evidently, the process Y_t depends on the current shock ϵ_t and the previous shock ϵ_{t-1} where the coefficient θ measures the magnitude at which the previous shock affects the process. Note μ is the mean of the process since:

$$\begin{aligned} E(Y_t) &= E(\mu + \theta\epsilon_{t-1} + \epsilon_t) = E(\mu) + \theta E(\epsilon_{t-1}) + E(\epsilon_t) \\ &= \mu + 0 + 0 = \mu \end{aligned}$$

For $\theta > 0$, MA(1) is persistent because the consecutive values are positively correlated. On the other hand, if $\theta < 0$, the process mean reverts because the effect of the previous shock is reversed in the current period.

The MA(1) model is always a covariance stationary process. The mean is as shown above, while the variance of the MA(1) model is given by:

$$\begin{aligned} V(Y_t) &= V(\mu + \theta\epsilon_{t-1} + \epsilon_t) = V(\mu) + \theta^2 V(\epsilon_{t-1}) + V(\epsilon_t) \\ &= 0 + \theta^2 V(\epsilon_{t-1}) + V(\epsilon_t) = \theta^2\sigma^2 + \sigma^2 \\ \Rightarrow V(Y_t) &= \sigma^2(1 + \theta^2) \end{aligned}$$

The variance uses the intuition that the shock is white noise processes that are uncorrelated.

The MA(1) model has a non-zero autocorrelation function given by:

$$\rho(h) = \begin{cases} 1, h = 0 \\ \frac{\theta}{1+\theta^2}, h = 1 \\ 0, h \geq 2 \end{cases}$$

The partial autocorrelations (PACF) of the MA(1) model is a complex and non-zero at all lags.

From the MA(1), we can generalize the q^{th} order MA process. Denoted by MA(q), it is given by:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$$

The mean of the MA(q) process is still μ since all the shocks are white noise process (their expectations are 0). The autocovariance function of the MA(q) process is given by:

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{i=0}^{q-h} \theta_i \theta_{i+h}, 0 \leq h \leq q \\ 0, h > q \end{cases}$$

And θ_0=1

The value of θ can be determined by substituting the value taken by the autocorrelation function and solving the resulting quadratic equation. The partial autocorrelation of an MA(q) model is complex and non-zero at all lags.

Example: Moving Average Process.

Given an MA(2), $Y_t = 3.0 + 5\epsilon_{t-1} + 5.75\epsilon_{t-2} + \epsilon_t$ where $\epsilon_t \sim WN(0, \sigma^2)$. What is the mean of the process?

Solution

The MA(2) is given by:

$$Y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t$$

Where μ is the mean. So, the mean of the above process is 3.0

The Autoregressive Moving Average (ARMA) Models

The ARMA model is a combination of AR and MA processes. Consider a first-order ARMA model (ARMA(1,1)). It is given by:

$$Y_t = \alpha + \beta Y_{t-1} + \theta \epsilon_{t-1} + \epsilon_t$$

The mean of the ARMA(1,1) model is given by:

$$\mu = \frac{\alpha}{1-\beta}$$

And variance is given by

$$\gamma_0 = \frac{\sigma^2(1+2\beta\theta)}{1-\beta^2}$$

The autocovariance function is given by:

$$\gamma(h) = \begin{cases} \sigma^2 \frac{1+2\beta\theta+\theta^2}{1-\beta^2}, h = 0 \\ \sigma^2 \frac{\beta(1+\beta\theta)+\theta(1+\beta\theta)}{1-\beta^2}, h = 1 \\ \beta\gamma_{h-1}, h \geq 2 \end{cases}$$

The ACF form of the ARMA(1,1) decays as the length h increases and oscillate if $\beta < 0$, which is consistent with the AR model.

The PACF tends to 0 as the length h increase, which is consistent with the MA process. The decay of ARMA's ACF and PACF is slow, which distinguishes it from the pure AR and MA models.

From the variance formula of ARMA(1,1), it is easy to see that the process is covariance stationery if $|\beta| < 1$

ARMA(p,q) Model

As the name suggests, ARMA(p,q) is a combination of the AR(p) and MA(q) process. Its form is given by:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \theta \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

When expressed using lag polynomial, this expression reduces to:

$$\beta(L)Y_t = \alpha + \theta(L)\epsilon_t$$

Analogous to ARMA(1,1), ARMA(p,q) is covariance -stationary if the AR portion is covariance stationary. The autocovariance and ACFs of the ARMA process are complex that decay at a slow pace to 0 as the lag h increases and possibly oscillate.

Sample Autocorrelation

The sample autocorrelation is utilized in validating the ARMA models. The autocovariance estimator is given by:

$$\hat{\gamma}_h = \frac{1}{T-h}\sum_{i=h+i}^{T}(Y_i - \bar{Y})(Y_{i-h} - \bar{Y})$$

Where $\bar{Y}$ is the full sample mean.

The autocorrelation estimator is given by:

$$\hat{\rho}_h = \frac{\sum_{i=h+i}^{T}(Y_i - \bar{Y})(Y_{i-h} - \bar{Y})}{\sum_{i=1}^{T}(Y_i - \bar{Y})^2} = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}$$

The autocorrelation is such that $-1 \leq \hat{\rho} \leq 1$

Test for Autocorrelation

Test for autocorrelation is done using the graphical examination by plotting ACF and PACF of the residuals and check for any deficiencies such as inadequacy of the model to capture the dynamics of the data. However, graphical methods are unreliable.

The common tests used are Box-Pierce and Ljung-Box tests.

Box-Pierce and Ljung-Box Tests.

Box-Pierce and Ljung-Box test both tests the null hypothesis that:

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_h$$

Against the alternative that:

$$H_1 : \rho_j \neq 0 (\text{At least one is non-zero})$$

Both the test are chi-distributed (χ^2_h) random variables. If the test statistic is larger than the critical value, the null hypothesis is rejected.

Box-Pierce Test

The test statistic under the Box-Pierce is given by:

$$Q_{BP} = T \sum_{i=1}^{h} \hat{\rho}_i^2$$

That is, the test statistic is the sum of squared autocorrelation scaled by the sample size T, which is (χ^2_h) random variable if the null hypothesis is true.

Ljung-Box Test

Ljung-Box test is a revised version of Box-Pierce that is appropriate with small sample sizes. The test statistic is given by:

$$Q_{LP} = T\ (T + 2) \sum_{i=1}^{h} \left(\frac{1}{T - i}\right) \hat{\rho}_i^2$$

The Ljung-Box test statistic is also χ^2_h random variable.

Model Selection

The first step in model selection is the inspection of the sample autocorrelations and the PACFs.

This provides the initial signs of the correlation of the data and thus can be used to select the type of models to be used.

The next step is to measure the fit of the selected model. The most commonly used method of measuring the model's fit is Mean Squared Error (MSE) which is defined as:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_t^2$$

When the MSE is small, the model selected explains more of the time series. However, choosing a model with a small MSE implies that we need to increase the coefficient of variation R^2, which can lead to overfitting. To attend to this problem, other methods have been developed to measure the fit of the model. These methods involve adding an adjustment factor to MSE each time a parameter is added. These measures are termed as **the Information Criteria (IC).** There are two such ICs: Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC).

Akaike Information Criteria (AIC)

Akaike Information Criteria (AIC) is defined as:

$$AIC = T\ ln\hat{\sigma}^2 + 2k$$

Where T is the sample size, and k is the number of the parameter. The AIC model adds the adjustment of adding two more parameters.

Bayesian Information Criteria (BIC).

Bayesian Information Criteria (BIC) is defined as:

$$BIC = T\ ln\hat{\sigma}^2 + klnT$$

Where the variables are defined as in AIC; however, note that the adjustment factor in BIC increases with an increase in the sample size T. Hence, it is a consistent model selection criterion. Moreover, the BIC criterion does not select the model that is larger than that selected by AIC.

The Box-Jenkin Methodology

The Box-Jenkin methodology provides a criterion of selecting between models that are equivalent but with different parameter values. The equivalency of the models implies that their mean, ACF and PACF are equal.

The Box-Jenkin methodology postulates two principles of selecting the models. One of the principles is termed as **Parsimony**. Under this principle, given two equivalent models, choose a model with fewer parameters.

The last principle is **invertibility**, which states that when selecting an MA or ARMA, select the model such that the coefficient in MA is invertible.

Model Forecasting

Forecasting is the process of using current information to forecast the future. In time series forecasting, we can make a one-step forecast or any time horizon h.

The one-step forecast time series forecasts the conditions expectation $E(Y_{T+1}|\Omega_T)$. Ω_T is termed as the information set at time T which includes the entire history of Y (Y_T, $Y_{T-1...}$) and the shock history ($\epsilon_T, \epsilon_{T-1}...$). In practice, this forecast is shortened to $E_T(Y_{T+1})$ so that:

$$E_T(Y_{T+1}|\Omega_T) = E_T(Y_{T+1})$$

Principles of Forecasting.

There are three rules of forecasting:

I. The expectation of a variable is the realization of that variable. That is: $E_T(Y_T) = Y_T$. This applies to the residuals: $E_T(\epsilon_{T-1}) = \epsilon_{T-1}$

II. The value of the expectation of future shocks is always 0. That is,

$$E_T(\epsilon_{T+h}) = 0$$

III. The forecasts are done recursively, beginning with $E_T(Y_{T+1})$ and that the forecast of a given time horizon might depend on the forecast of the previous horizon.

Let us consider some examples.

For the AR(1) model, the one-step forecast is given by:

$$\begin{aligned} E_T(Y_{T+1}) &= E_T(\alpha + \beta Y_T + \epsilon_{T+1}) = \alpha + \beta E_T(Y_T) + 0 \\ &= \alpha + \beta Y_T \end{aligned}$$

Note that we are using the current values Y_T to predict Y_{T+1} and shock used is that of the future ϵ_{T+1}.

The two-step forecast is given by:

$$\begin{aligned} E_T(Y_{T+2}) &= E_T(\alpha + \beta Y_{T+1} + \epsilon_{T+2}) \\ &= \alpha + \beta E_T(Y_{T+1}) + E_T(\epsilon_{T+2}) \end{aligned}$$

But $E_T(\epsilon_{T+2}) = 0$ and $E_T(Y_{T+1}) = \alpha + \beta Y_T$

So that:

$$\begin{aligned} E_T(Y_{T+2}) &= \alpha + \beta E_T(\alpha + \beta Y_T) = \alpha + \beta(\alpha + \beta Y_T) \\ \Rightarrow E_T(Y_{T+2}) &= \alpha + \alpha\beta + \beta^2 Y_T \end{aligned}$$

Analogously, the forecast for time horizon h we have:

$$\begin{aligned} E_T(Y_{T+h}) &= \alpha + \alpha\beta + \alpha\beta^2 + \ldots + \alpha\beta^{h-1} + \beta^h Y_T \\ &= \sum_{i=1}^{h} \alpha\beta^i + \beta^h Y_T \end{aligned}$$

The Mean Reverting Level

When h is large, β^h must be very small by the intuition of covariance stationary of Y_t. Therefore, it can be shown that:

$$\lim_{h\to\infty}\sum_{i=0}^{h}\alpha\beta^{i}\beta^{h}Y_T=\frac{\alpha}{1-\beta}$$

The limit is actually the mean of the AR(1) model. The mean-reverting level implies Y_T does not affect the future value of Y. That is,

$$\lim_{h\to\infty}E_T\left(Y_{T+h}\right)=E(Y_t)$$

The same procedure is applied to MA and ARMA models.

The forecast error is the difference between the true future value and the forecasted value, that is,

$$\epsilon_{T+1}=Y_{T+1}-E_T\left(Y_{T+1}\right)$$

For longer time-horizon, the forecast is mostly functions of the model parameters.

Example: Model Forecasting

The ARMA(1,1) for modeling the default in premiums for an insurance company is given by

$$D_t=0.055+0.934D_{t-1}+\epsilon_t$$

Given that $D_T=1.50$, what is the first step forecast of the default?

Solution

We need:

$$\begin{aligned}&E_T\left(Y_{T+1}\right)=\alpha+\beta Y_T\\&\Rightarrow E_T\left(D_{T+1}\right)=0.055+0.934\times1.5=1.4560\end{aligned}$$

Seasonality of Time Series

Some time-series data are seasonal. For instance, the sales at the time of summer that may differ from that of winter. The time series with deterministic seasonality is termed as non-stationary, while those with stochastic seasonality are called stationary time series and hence modeled with AR or ARMA process.

A pure seasonal lag utilizes the lags at a seasonal frequency. For instance, assume that we are using the semi-annual data, then the pure seasonal AR(1) model of quarterly time seasonal time series is:

$$(1 - \beta L^4)Y_t = \alpha + \epsilon_t$$

So that:

$$Y_t = \alpha + \beta Y_{t-4} + \epsilon_t$$

A more efficient seasonality includes the short-term and seasonal lag components. The short-term components utilize the lags at the observation frequency.

Seasonality can also be introduced to AR, MA, or both models by multiplying the short run lag polynomial and by the seasonal lag polynomial. For instance, the seasonal ARMA is specified as:

$$ARMA(p, q) \times (p_s, q_s)_f$$

Where p and q are the orders of the short run-lag polynomials, and p_s and q_s are the seasonal lag polynomials. Practically, seasonal lag polynomials are restricted to one seasonal lag because the accuracy of the parameter approximations depends on the number of full seasonal cycles in the sample data.

Question 1

The following sample autocorrelation estimates are obtained using 300 data points:

Lag	1	2	3
Coefficient	0.25	−0.1	−0.05

Compute the value of the Box-Pierce Q-statistic.

A. 22.5

B. 22.74

C. 30

D. 30.1

The correct answer is **A**.

$$\begin{aligned} Q_{BP} &= T \sum_{h=1}^{m} \hat{\rho}^2(h) \\ &= 300(0.25^2 + (-0.1)^2 + (-0.05)^2) \\ &= 22.5 \end{aligned}$$

Question 2

The following sample autocorrelation estimates are obtained using 300 data points:

Lag	1	2	3
Coefficient	0.25	−0.1	−0.05

Compute the value of the Ljung-Box Q-statistic.

A. 30.1

B. 30

C. 22.5

D. 22.74

The correct answer is **D**.

$$\begin{aligned} Q_{LB} &= T\ (T+2)\sum_{h=1}^{m}\left(\frac{\hat{}1}{T-h}\right)\rho^2(h) \\ &= 300(302)\left(\frac{0.25^2}{299}+\frac{-0.1^2}{298}+\frac{-0.05^2}{297}\right) \\ &= 22.74 \end{aligned}$$

Note: Provided the sample size is large, the Box-Pierce and the Ljung-Box tests typically arrive at the same result.

Question 3

Assume the shock in a time series is approximated by Gaussian white noise. Yesterday's realization, y(t) was 0.015, and the lagged shock was -0.160. Today's shock is 0.170.

If the weight parameter theta, θ, is equal to 0.70 and the mean of the process is 0.5, determine today's realization under a first-order moving average, MA(1), process.

A. -4.205

B. 4.545

C. 0.558

D. 0.282

The correct answer is **C**.

Today's shock = ϵ_t; yesterday's shock = ϵ_{t-1}; today's realization = y_t; yesterday's realization = y_{t-1}.

The MA(1) is given by:

$$\begin{aligned} y_t &= \mu + \theta\epsilon_{t-1} + \epsilon_t \\ &= 0.5 + 0.170 + 0.7(-0.160) = 0.558 \\ &= 0.558 \end{aligned}$$

Reading 22: Nonstationary Time Series

After completing this reading, you should be able to:

- Describe linear and nonlinear time trends.
- Explain how to use regression analysis to model seasonality.
- Describe a random walk and a unit root.
- Explain the challenges of modeling time series containing unit-roots.
- Describe how to test if a time series contains a unit root.
- Explain how to construct an h-step-ahead point forecast for a time series with seasonality.
- Calculate the estimated trend value and form an interval forecast for a time series.

Recall that the stationary time series have means, variance, and autocovariance that are independent of time. Therefore any time series that violates this rule is termed as the non-stationary time series.

The nonstationary time series include time trends, random walks(also called unit-roots) and seasonalities. Time trends reflect the feature of the time series to grow over time.

Seasonalities occur due to change in the time series over different seasons such as each quarter. Seasonalities can be shifts of the mean (for example depending on the period of the year) and the mean cycle of the time series (this occurs when the shock of the current value depends on the shock of the same future period). Seasonalities can be modeled using the dummy variables or modeling it period after period changes (such as year after year) in an attempt to remove the seasonal change of the mean.

In a random walk, time series depends on each other and their respective shocks. We discuss each of the non-stationarities.

Time Trends.

The time trend deterministically shifts the mean of the time series. The time trend can be linear and non-linear (which includes log and quadratic time series).

Linear Time Trends

Linear trend models are those that the dependent variable changes at a constant rate with time. If the time series y_t has a linear trend, we can model the series by the following equation:

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t, t = 1, 2, \dots, T$$

Where

Y_t =the value of the time series at time t (trend value at time t)

β_0=the y-intercept term

β_1=the slope coefficient

t=time, the independent (explanatory) variable

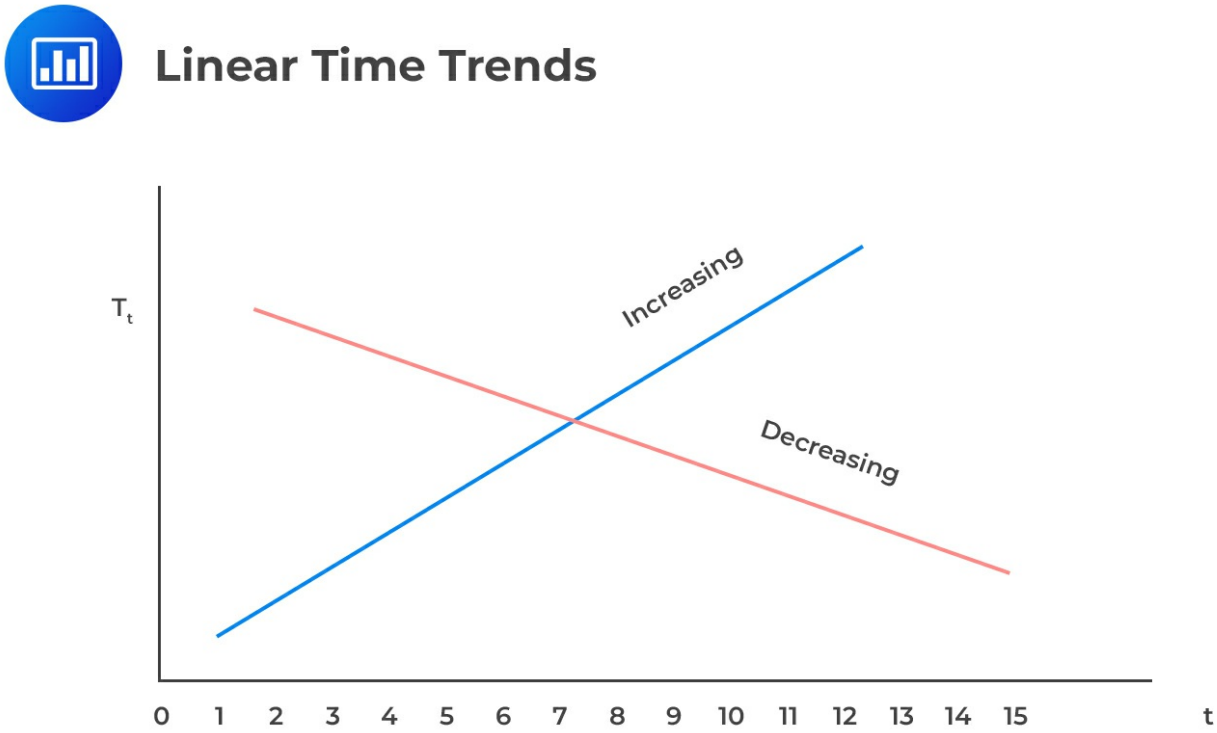
ϵ_t= a random error term (Shock) and is white noise ($\epsilon_t \sim WN(0, \sigma^2)$)

From the equation above, the $\beta_0 + \beta_1 t$ predicts y_t at any time t. The slope β_1 is described as the trend coefficient since it is the slope coefficient. We estimate both factors β_0 and β_1 using the ordinary least squares and denoted as: $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively.

The mean of the linear time series is:

$$E(Y_t) = \beta_0 + \beta_1 t$$

On a graph, a linear trend appears as a straight line angled diagonally up or down.



Estimation of the Trend Value Under Linear Trend Models

Using the estimated coefficients, we can predict the value of the dependent variable at any time (t=1, 2…, T). For instance, the trend value at time 2 is $\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 (2)$. We can also forecast the value of the time series outside the sample's period, that is, T+1. Therefore, the predicted value of Y_t at time T+1 is $\hat{Y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1(T + 1)$.

Example: Calculating the Trend Value

A linear trend is defined to be $Y_t = 17.5 + 0.65t$. What is the trend projection for time 10?

Solution

We substitute t=10, which is:

$$T = 17.5 + 0.65 \times 10 = 24$$

Disadvantages of Linear Time Series

In linear time series, the growth is a constant which might pose problems in economic and financial time series.

1. When the trend is positive, then the growth rate is expected to decrease over time.
2. If the slope coefficient is less than 0, the Y_t will tend toward negative values, a situation that would not be plausible in most financial time series, e.g., asset prices and quantiles.

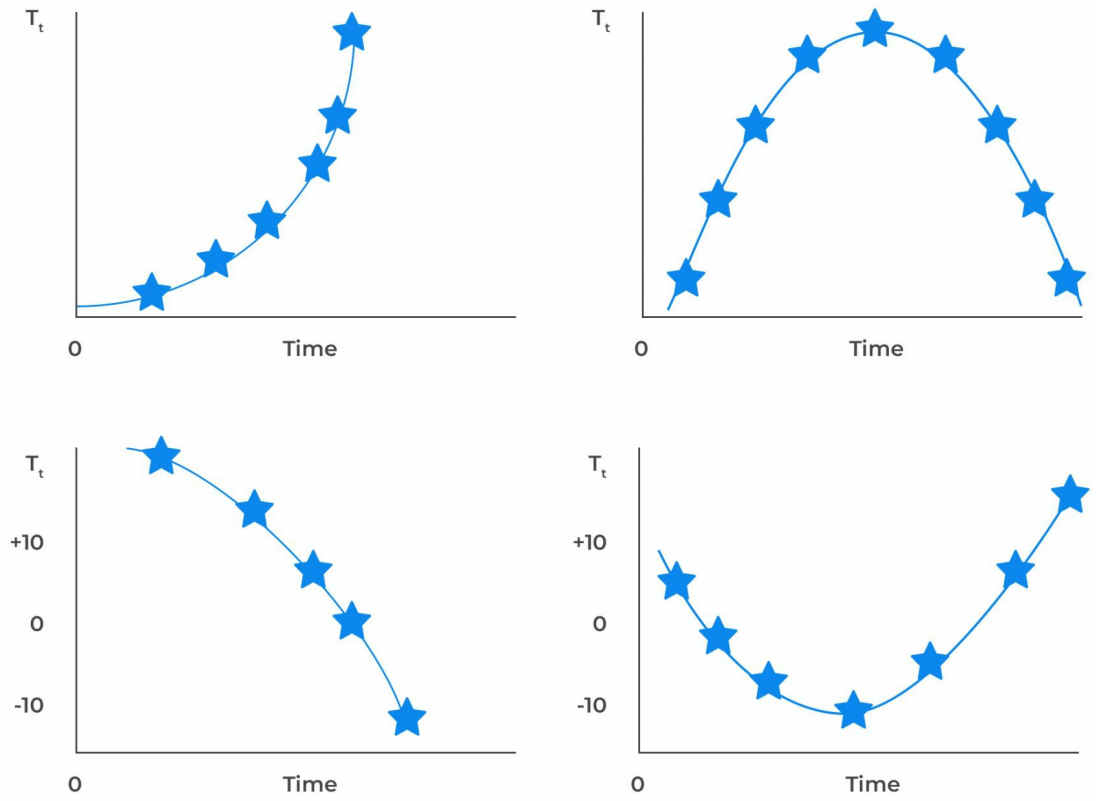
Considering these limitations, we discuss the log-linear time series, with a constant growth rate rather than just a constant rate.

Log-Linear Trend Models

Sometimes the linear trend models result in uncorrelated errors. For instance, the time series with exponential growth rates. The appropriate model for the time series with exponential growth is the Log-linear trend model.

Log-linear trends are those in which the variable changes at an increasing or decreasing rate rather than at a constant rate like in linear trends.

Log-linear Time Trends (Examples)



Assume that the time series is defined as:

$$Y_t = e^{\beta_0 + \beta_1 t}, t = 1, 2, \ldots, T$$

Which also can be written as (by taking the natural logarithms on both sides):

$$\ln Y_t = \beta_0 + \beta_1 t, \quad t = 1, 2, \ldots, T$$

By Exponential rate, we mean growth at a constant rate with continuous compounding. This can be seen as follows: Using the time series formula above, the value of the time series at time 1 and 2 are $y_1 = e^{\beta_0 + \beta_1 (1)}$ and $y_2 = e^{\beta_0 + \beta_1 (2)}$. The ratio $\frac{y_2}{y_1}$ is given by:

$$\frac{Y_2}{Y_1} = \frac{e^{\beta_0 + \beta_1 (2)}}{e^{\beta_0 + \beta_1 (1)}} = e^{\beta (1)}$$

Similarly, the value of the time-series at time t is $Y_t = e^{\beta_0+\beta_1 t}$, and at t+1, we have $Y_{t+1} = e^{\beta_0+\beta_1(t+1)}$. This implies that the ratio:

$$\frac{Y_{t+1}}{Y_t} = \frac{e^{\beta_0+\beta_1(t+1)}}{e^{\beta_0+\beta_1(t)}} = e^{\beta_1}$$

If we take the natural logarithm on both sides of the above equation we have:

$$\ln\left(\frac{Y_{t+1}}{Y_t}\right) = \ln Y_{t+1} - \ln Y_t = \beta_1$$

The log-linear model implies that:

$$E(\ln Y_{t+1} - \ln Y_t) = \beta_1$$

From the above results, proportional growth in time series over the two consecutive periods is equal. That is:

$$\frac{y_{t+1} - y_t}{y_t} = \frac{y_{t+1}}{y_t} - 1 = e^{\beta_1} - 1$$

Example: Calculating the Trend Value of a Log-Linear Trend Time Series

An investment analyst wants to fit the weekly sales (in millions) of his company by using the sales data from Jan 2016 to Feb 2018. The regression equation is defined as:

$$\ln Y_t = 5.1062 + 0.0443t, t = 1, 2, \dots, 100$$

What is the trend estimated value of the sales in the 80th week?

Solution

From the regression equation, $\hat{\beta}_0 = 5.1062$ and $\hat{\beta}_1 = 0.0443$. We know that, under log-linear trend models, the predicted trend value is given by:

$$Y_t = e^{\hat{\beta}_0 + \hat{\beta}_1 t}$$

$$\Rightarrow Y_{80} = e^{5.1062 + 0.0443 \times 80} = 5711.29 \text{ Million}$$

Quadratic Time Trend

A polynomial-time trend can be defined as:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_m t^m \epsilon_t, t = 1, 2, \ldots, T$$

Practically speaking, the polynomial-time trends are only limited to the linear (discussed above) and the quadratic (second degree) time trend. In a quadratic time trend, the parameter can be estimated using the OLS. The approximated parameter are asymptotically normally distributed and hence statistical inference using the t-statistics and the standard error happen only if the residuals ϵ_t are white noise.

The Log-Quadratic Time Trend

As the name suggests, this time trend is a mixture of the log-linear and quadratic time series. It is given by:

$$\ln Y_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

It can be shown that the growth rate of the log-quadratic time trend is $\beta_1 + 2\beta_2 t$. This can be seen as follows:

The value of the time-series at time t is $Y_t = e^{\beta_0 + \beta_1 t + \beta_2 t^2}$, and at t+1, we have

$Y_{t+1} = e^{\beta_0 + \beta_1 (t+1) + \beta_2 (t+1)^2}$. This implies that the ratio:

$$\frac{Y_{t+1}}{Y_t} = \frac{e^{\beta_0 + \beta_1 (t+1) + \beta_2 (t+1)^2}}{e^{\beta_0 + \beta_1 t + \beta_2 t^2}} = e^{\beta_1 + 2\beta_2 t}$$

If we take a natural log on the results, we get the desired result.

Example: Calculating the Growth Rate of Log-Quadratic Time Trend

The monthly real GDP of a country over 20 years can be modeled by the time series equation given by:

$$RG_T = 6.75 + 0.015t + 0.0000564t^2$$

What is the growth rate of the real GDP of this country at the end of 20 years?

Solution

This is the log-quadratic time trend whose growth rate is given by

$$\beta_1 + 2\beta_2 t$$

From the regression time-series equation given, we have $\hat{\beta}_1 = 0.015$ and $\hat{\beta}_2 = 0.0000564$ so that the growth rate is given by:

$$\beta_1 + 2\beta_2 t = 0.015 + 2 \times 0.0000564 \times 240 = 0.0421$$

Note that, since the data is modeled monthly, at the end of 20 years implies 240th month!

The coefficient of variation (R^2) for the time trend series is always high and will tend to 100% as the sample size increases. Therefore, the coefficient of variation is not an appropriate measure in trend series. Other alternatives such as residual diagnostics, can be useful.

Seasonality

Seasonality is a feature of a time series in which the data undergoes regular and predictable changes that recur every calendar year. For instance, gas consumption in the US rises during the winter and falls during the summer.

Seasonal effects are observed within a calendar year, e.g., spikes in sales over Christmas, while cyclical effects span time periods shorter or longer than one calendar year, e.g., spikes in sales due

to low unemployment rates.

Modeling Seasonal Time Series

Regression on seasonal dummies is an essential method of modeling seasonality. Assuming that there are s seasons in a year. Then the pure annual dummy model is:

$$\begin{aligned} Y_t &= \beta_0 + \gamma_1 D_{1t} \gamma_2 D_{2t} + \cdots + \gamma_{s-1} D_{s-1t} + \epsilon_t \\ &= \beta_0 + \sum_{j=1}^{s-1} \gamma_j D_{jt} + \epsilon_t \end{aligned}$$

D_{jt} is defined as:

$$D_{jt} = \begin{cases} 1, t \bmod s = j \\ 1, t \bmod s \neq j \end{cases}$$

γ_j measures the amount of difference of the mean at period j and s.

Note X mod Y is the remainder of the X/Y.For instance, 9 mod 4=1.

The mean of the first period of the seasonality is:

$$E[Y_1] = \beta_0 + \gamma_1$$

And the mean of period 2 is:

$$E[Y_2] = \beta_0 + \gamma_2$$

Since period s, all dummy variables are zero, then the mean of the seasonality at time s is:

$$E[Y_s] = \beta_0$$

The parameters of seasonality are estimated using the OLS estimators by regressing Y_t on constant and s-1 dummy variables.

Combination of Stationary and Non-Stationary Time Series

Time trends and seasonalities can be insufficient in explaining economic time series and since their residuals might not be white noise. In the case that the non-stationary time series appears to be stationary, but the residuals are not white noise, we can add stationary time series components (such as AR and MA) to reflect the components of the non-stationary time series.

Consider the following linear time trend.

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

If the residuals are not white noise but the time series appears to be stationary, we can include an AR term to make the model's residuals white noise:

$$Y_t = \beta_0 + \beta_1 t + \delta_1 Y_{t-1} + \epsilon_t$$

We can also add the seasonal component (if it exists):

$$Y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{s-1} \gamma_j D_{jt} + \delta_1 Y_{t-1} + \epsilon_t$$

Note that the AR component reflects the cyclicity of the time series, γ_j measures the shifts of the mean from the trend growth, i.e $\beta_1 t$. However, combinations of the time series do not always lead to a model with the required dynamics. For instance, the Ljung-Box statistics may suggest rejection of the null hypothesis.

Unit Roots and Random Walks

A random walk is a time series in which the value of the series in one period is equivalent to the value of the series in the previous period plus the unforeseeable random error. A random walk can be defined as follows:

Let

$$Y_t = Y_{t-1} + \epsilon_t$$

Intuitively,

$$Y_{t-1} = Y_{t-2} + \epsilon_{t-1}$$

If we substitute Y_{t-1} in the first equation, we get,

$$Y_t = (Y_{t-2} + \epsilon_{t-1}) + \epsilon_t$$

Continuing this process, it implies that a random walk is given by:

$$Y_t = Y_0 + \sum_{i=1}^{t} \epsilon_i$$

The random walk equation is a particular case of an AR(1) model with $\beta_0 = 0$ and $\beta_1 = 1$. Thus, we cannot utilize the regression techniques to estimate such AR(1). This is because a random walk does not have a finite mean-reverting level or finite variance. Recall that if Y_t has a mean-reverting level, then $Y_t = \beta_0 + \beta_1 Y_t$ and thus $\frac{\beta_0}{1-\beta_1}$. However, in a random walk, $\beta_0 = 0$ and $\beta_1 = 1$ so, $\frac{0}{1-1} = 0$.

The variance of a random walk is given by:

$$V(Y_t) = t\sigma^2$$

The implication of the infinite variance of a random walk is that we are unable to use standard regression analysis on a time series that appears to be a random walk.

Unit Roots

We have been discussing the random walks without a drift; that the current value is the best predictor of the time series in the next period.

A random walk with a drift is defined as a time-series where it increases or decreases by a constant amount in each period. It is mathematically described as:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

$\beta_0 \neq 1, \beta_1 = 1$

Or

$$Y_t = \beta_0 + Y_{t-1} + \epsilon_t$$

Where $\epsilon_t \sim WN(0, \sigma^2)$

Recall that $\beta_1 = 1$ implies undefined mean-reversion level and hence non-stationarity. Therefore, we are unable to use the AR model to analyze a time series unless we transform the time series by taking the first difference we get:

$$\Delta Y_t = Y_t - Y_{(t-1)}, y_t = \beta_0 + \epsilon_t, \forall \beta_0 \neq 0$$

Which is covariance stationary.

The unit root test involves the application of the random walk concepts to determine whether a time series is nonstationary by focusing on the slope coefficient in a random walk time series with a drift case of AR(1) model. This test is popularly known as the Dickey-Fuller Test

The Unit Root Problem

Consider an AR(1) model. If the time-series originates from an AR(1) model, then the time-series is covariance stationary if the absolute value of the lag coefficient β_1 is less than 1. That is, $|\beta_1| < 1$. Therefore, we could not depend on the statistical results if the lag coefficient is greater or equal to 1 ($|\beta_1| \geq 1$).

When the lag coefficient is precisely equal to 1, then the time series is said to have a unit root. In other words, the time-series is a random walk and hence not covariance stationary.

The unit root problem can also be expressed using the lag polynomial. Let

$\psi(L)$ be the full lag polynomial, which can be factorized into the unit root lag denoted by (1-L) and the remainder lag polynomial $\phi(L)$ which is the characteristic lag for stationary time series. Moreover, let $\theta(L)\epsilon_t$ be an MA. Thus, the unit root process can be described as:

$$\psi(L)Y_t = \theta(L)\epsilon_t$$

This can be factorized into:

$$(1 - L)\phi(L) = \theta(L)\epsilon_t$$

Example: Checking for Unit Roots using the Lag Polynomials

An AR(2) model is given by $Y_t = 1.7Y_{t-1} - 0.7Y_{t-2} + \epsilon_t$. Does the process contain a unit root?

Solution

If we rearrange the equation:

$$Y_t - 1.7Y_{t-1} + 0.7Y_{t-2} = \epsilon_t$$

Using the definition of a lag polynomial, we can write the above equation as:

$$(1 - 1.7L + 0.7L^2)Y_t = \epsilon_t$$

The right-hand side is a quadratic equation which can be factorized. So,

$$(1 - L)(1 - 0.7L)Y_t = \epsilon_t$$

Therefore, the process has a unit root due to the presence of a unit root lag operator (1-L).

Challenges of Modeling Time Series Containing Unit Roots

1. A unit root process does not have a mean-reverting level. Recall that the stationary time series does mean revert, that is, the long-run mean can be estimated.
2. In a time series with a unit root, spotting spurious relationships is a problem. A spurious correlation is where there is no important link between the time series but regression analysis produces significant parameter estimates.
3. The parameter estimators in ARMA time series with a unit root possess Dickey-Fuller (DF) distribution, which is asymmetric, dependent on the sample size, and that its critical value

depends on whether time trends have been incorporated. This characteristic makes it difficult to come up with sound statistical inference and model selection when fitting the models.

Transformation of Time Series with Unit Roots

If the time series seem to have unit roots, the best method is to model it using the first-differencing series as an autoregressive time series, which can be effectively analyzed using regression analysis.

Recall that the time series with a drift is a form of AR(1) model given by:

$$y_t = \beta_0 + Y_{t-1} + \epsilon_t,$$

Where $\epsilon_t \sim WN(0, \sigma^2)$

Clearly $\beta_1 = 1$ implies that the time series has an undefined mean-reversion level and hence non-stationary. Therefore, we are unable to use the AR model to analyze time series unless we transform the time series by taking the first difference to get:

$$Y_t = Y_t - Y_{(t-1)} \Rightarrow y_t = \beta_0 + \epsilon_t, \forall \beta_0 \neq 0$$

Where the $\epsilon_t \sim WN(0, \sigma^2)$ and thus covariance stationary.

Using the lag polynomials, let $\Delta Y_t = Y_t - Y_{(t-1)}$ where Y_t has a unit root (implying that $Y_t - Y_{(t-1)}$ does not have a unit root.), then:

$$\begin{aligned} (1 - L)\phi(L)Y_t &= \epsilon_t \\ \phi(L)[(1 - L)Y_t] &= \epsilon_t \\ \phi(L)[(Y_t - LY_t)] &= \epsilon_t \\ \phi(L)\Delta Y_t &= \epsilon_t \end{aligned}$$

Since the lag polynomial $\phi(L)$ is stationary series lag polynomial, the time series defined by ΔY_t must be stationary.

Unit Root Test

The unit root test is done using the Augmented Dickey-Fuller (ADF) test. The test involves OLS estimation of the parameters where the difference of the time series is regressed on the lagged level, appropriate deterministic terms, and the lagged difference.

The ADF regression is given by:

$$\Delta Y_t = \gamma Y_{t-1} + (\delta_0 + \delta_1 t) + (\lambda \Delta Y_{t-1} + \lambda_2 \Delta Y_{t-2} + \cdots + \lambda_p \Delta Y_{(t-p)})$$

Where:

γY_{t-1}=Lagged level

$\delta_0 + \delta_1 t$=deterministic terms

$\lambda \Delta Y_{t-1} + \lambda_2 \Delta Y_{t-2} + \cdots + \lambda_p \Delta Y_{(t-p)}$=Lagged differences.

The test statistic for the ADF test is that of $\hat{\gamma}$(estimate of γ).

To get the gist of this, assume that we are conducting an ADF test on a time series with lagged level only:

$$\Delta Y_t = \gamma Y_{t-1}$$

Intuitively, if the time series is a random walk, then:

$$Y_t = Y_{t-1} + \epsilon_t$$

If we subtract Y_{t-1} on both sides we get:

$$\begin{aligned} Y_t - Y_{t-1} &= Y_{t-1} - Y_{t-1} + \epsilon_t \\ \Rightarrow \Delta Y_t &= 0 \times Y_{t-1} + \epsilon_t \end{aligned}$$

Therefore, it implies that the time series is a random walk if γ=0. This leads us to the hypothesis statement of the ADF test:

$H_0 : \gamma = 0$ (The time series is a random walk)

$H_1 : \gamma < 0$ (the time series is a covariance stationary)

You should note this is a one-sided test, and thus, the null hypothesis is not rejected if $\gamma>0$. The positivity of γ corresponds to an AR time series stationary. For example, recall that the AR(1) model is given by:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

If we subtract Y_{t-1} from both sides of the AR(1) above we have:

$$Y_t - Y_{t-1} = \beta_0 + (\beta_1 - 1)Y_{t-1} + \epsilon_t$$

Now let $\gamma = (\beta_1 - 1)$. Therefore,

$$\Delta Y_t = \beta_0 + \gamma Y_{t-1} + \epsilon_t$$

Clearly, if $\beta_1 = 1$, then let $\gamma = 0$. Therefore, $\gamma = 0$ is the test for $\beta_1 = 1$. In other words, if there is a unit root in an AR(1) model (with the dependent variable being the difference between the time series and independent variable of the first lag) then, $\gamma = 0$, implying that the series has a unit root and is nonstationary.

Implementing an ADF test on a time series requires making two choices: which deterministic terms to include and the number of lags of the differenced data to use. The number of lags to include is simple to determine—it should be large enough to absorb any short-run dynamics in the difference ΔY_t

The appropriate method of selecting the lagged differences is the AIC (which selects a relatively larger model as compared to BIC). The length of the lag should be set depending on the length of the time series and the frequency of the sampling.

The Dickey-Fuller distributions are dependent on the choice of deterministic terms included. The deterministic terms can be excluded, and instead, use constant terms or trend deterministic terms. While keeping all other things equal, the addition of more deterministic terms reduces the chance of rejecting the null hypothesis when the time series does not have a unit root, and hence the power of the ADF test is reduced. Therefore, relevant deterministic terms should be included.

The recommended method of choosing appropriate deterministic terms is by including the

deterministic terms that are significant at 10% level. In case the deterministic trend term is not significant at 10%, it is then dropped and the constant deterministic term is used instead. If the trend is also insignificant, then it can be dropped and the test is rerun without the deterministic term. It is important to note that the majority of macroeconomic time series require the use of the constant.

In the case that the null of the ADF test cannot be rejected, the series should be differenced and the test is rerun to make sure that the time series is stationary. If this is repeated (double differenced) and the time series is still non-stationary, then other transformations to the data such as taking the natural log(if the time series is always positive) might be required.

Example: Conducting the ADF Test

A financial analyst wishes to conduct an ADF test on the log of 20-year real GDP from 1999 to 2019. The result of the tests is shown below:

Deterministic	γ	δ_0	δ_1	Lags	5%CV	1%CV
None	−0.004 (−1.665)			8	−1.940	−2.570
Constant	−0.008 (−1.422)	0.010 (1.025)		4	−2.860	−3.445
Trend	−0.084 (−4.376)	0.188 (−4.110)		3	−3.420	−3.984

The output of the ADF reports the results at the different number deterministic terms (first column), and the last three columns indicate the number of lags according to AIC and the 5% and 1% critical values that are appropriate to the underlying sample size and the deterministic terms. The quantities in the parenthesis (below the parameters) are the test statistics.

Determine whether the time series contains a unit root.

Solution

The hypothesis statement of the ADF test is:

$H_0 : \gamma = 0$ (The time series is a random walk)

$H_1 : \gamma < 0$ (the time series is a covariance stationary)

We begin with choosing the appropriate model. At 10%, the trend model has an absolute value of the statistic greater than the CV at 1% and 5% significance level; thus, we choose a model with the trend deterministic term.

Therefore, for this model, the null hypothesis is rejected at a 99% confidence level since |-4.376|>|-3.984|. Note that the null hypothesis is also rejected at a 95% confidence level.

Moreover, if the model was constant or no-deterministic, the null hypothesis will fail to be rejected. This reiterates the importance of choosing an appropriate model.

The Seasonal Differencing

Seasonal differencing is an alternative method of modeling the seasonal time series with a unit root. Seasonal differencing is done by subtracting the value in the same period in the previous year to remove the deterministic seasonalities, the unit root, and the time trends.

Consider the following quarterly time series with deterministic seasonalities and non-zero growth rate:

$$Y_t = \beta_0 + \beta_1 t + \gamma_1 D_{1t} + \gamma_2 D_{2t} + \gamma_3 D_{3t} + \epsilon_t$$

Where $\epsilon_t \sim WN(0, \sigma^2)$.

Denote a seasonal $\Delta_4 Y_t = Y_t - Y_{t-4}$

$$\begin{aligned} \Rightarrow \Delta_4 Y_t =& (\beta_0 + \beta_1 t + \gamma_1 D_{1t} + \gamma_2 D_{2t} + \gamma_3 D_{3t} + \epsilon_t) \\ & - (\beta_0 + \beta_1 (t-4) + \gamma_1 D_{1t-4} + \gamma_2 D_{2t-4} + \gamma_3 D_{3t-4} + \epsilon_{t-4}) \\ =& \beta_1 (t - (t-4)) - [\gamma_1 (D_{1t} - D_{1t-4}) + \gamma_1 (D_{12} - D_{2t-4}) + \gamma_1 (D_{3t} - D_{3t-4})] + \epsilon_t \\ & - \epsilon_{t-4} \end{aligned}$$

But

$$\gamma_j (D_{1j} - D_{1j-4}) = 0$$

Because $D_{1j} = D_{1j-4}$ by the definition of the seasonal differencing. So that:

$$\Rightarrow \Delta_4 Y_t = \beta_1(t-(t-4)) + \epsilon_t - \epsilon_{t-4}$$

Therefore,

$$\Delta_4 Y_t = 4\beta_1 + \epsilon_t - \epsilon_{t-4}$$

Intuitively, this an MA(1) model, which is covariance stationary. The seasonal differenced time series is described as the year to year change in Y_t or year to year growth in case of logged time series.

Spurious Regression

Spurious regression is a type of **regression** that gives misleading statistical evidence of a linear relationship between independent non-stationary variables. This is a problem in time series analysis, but this can be avoided by making sure each of the time series in question is stationary by using methods such as first differencing and log transformation (in case the time series is positive)

Condition for Differencing in Time Series

Practically, many financial and economic time series are plausibly persistent but stationary. Therefore, differencing is only required when there is clear evidence of unit root in the time series. Moreover, when it is difficult to distinguish whether time series is stationary or not, it is a good statistical practice to generate models at both levels and the differences.

For example, we wish to model the interest rate on government bonds using an AR(3) model. The AR(3) is estimated on the levels and the differences (if we assume the existence of unit root) are modeled by AR(2) since the AR is reduced by one due to differencing. By considering the models at all levels allows us to choose the best model when the time series are highly persistent.

Forecasting

Forecasting in non-stationary time series is analogous to that of stationary time series. That is, the forecasted value at time T is the expected value of Y_{T+h}.

Consider a linear time trend:

$$Y_T = \beta_0 + \beta_1 T + \epsilon_t$$

Intuitively,

$$Y_{T+h} = \beta_0 + \beta_1(T + h) + \epsilon_{t+h}$$

Taking the expectation, we get:

$$\begin{aligned} & E_T(Y_{T+h}) = E_T(\beta_0) + E_T(\beta_1(T + h) + E_T(\epsilon_{t+h}) \\ \Rightarrow & E_T(Y_{T+h}) = \beta_0 + \beta_1(T + h) \end{aligned}$$

This is true because of both β_0 and $\beta_1(T + h)$ are constants while $\epsilon_{t+h} \sim WN(0, \sigma^2)$.

Forecasting in Seasonal Time Series

Recall that the seasonal time series can be modeled using the dummy variables. Consequently, we need to track the period of the forecast we desire. The annual time series is given by:

$$Y_T = \beta_0 + \sum_{j=1}^{s-1} \gamma_j D_{jt} + \epsilon_t$$

The first-step forecast is:

$$E_T(Y_{T+1}) = \beta_0 + \gamma_j$$

Where:

$j = (T + 1) \bmod s$ is the forecasted period and that the forecast and the coefficient on the omitted periods is 0.

For instance, for quarterly seasonal time series that excludes the dummy variable for the fourth quarter (Q_4), then the forecast for period 116 is given by:

$$\begin{aligned} & E_T(Y_{T+1}) = \beta_0 + \gamma_j \\ & E_T(Y_{T+1}) = \beta_0 + \gamma_{(116+1)(\bmod 4)} = \beta_0 + \gamma_1 \end{aligned}$$

Therefore, the h-step ahead forecast are by tracking the period of $T+h$ so that:

$$E_T(Y_{T+h}) = \beta_0 + \gamma_j$$

Where:

$$j = (T + h) \bmod s$$

Forecasting in Log Models

Under the log model, you should note that:

$$E(Y_{T+h}) \neq E(\ln Y_{T+h})$$

If the residuals are Gaussian white noise, that is:

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

Then the properties of the log-normal can be used for forecasting. If

$X \sim N(0, \sigma^2)$, then define $W = e^X \sim Log(\mu, \sigma^2)$. Also recall that the mean of a log-normal distribution is given by:

$$E(W) = e^{\mu + \frac{\sigma^2}{2}}$$

Using this analogy, for a log-linear time trend model:

$$\ln Y_{T+h} = \beta_0 + \beta_1(Y_{T+h}) + \epsilon_{T+h}$$

The forecast at time T+h,

$$E_T(\ln Y_{T+h}) = \beta_0 + \beta_1(Y_{T+h})$$

The variance of the shock is σ^2 so that:

$$\ln Y_{T+h} \sim (\beta_0 + \beta_1(Y_{T+h}), \sigma^2)$$

Thus,

$$E_T(Y_{T+h}) = e^{\beta_0+\beta_1(Y_{T+h})+\frac{\sigma^2}{2}}$$

Forecasting Confidence Intervals

Confidence intervals are constructed to reflect the uncertainty of the forecasted value. The confidence interval is dependent on the variance of the forecasted error, which is defined as:

$$\epsilon_{T+h} = Y_{T+h} - E_T(Y_{T+h})$$

i.e., it is the difference between the actual value and the forecasted value.

Consider the linear time trend model:

$$Y_{T+h} = \beta_0 + \beta_1(T + h) + \epsilon_{T+h}$$

Clearly,

$$E_T(Y_{T+h}) = \beta_0 + \beta_1(T + h)$$

And the forecast error is ϵ_{T+h}

If we wish to construct a 95% confidence interval, given that the forecast error is Gaussian white noise, then the confidence interval is given by:

$$E_T(Y_{T+h}) \pm 1.96\sigma$$

σ is not known and thus can be **estimated by the variance of the forecast error**.

Intuitively, the confidence intervals for any model can be computed depending on the individual forecast error $\epsilon_{T+h} = Y_{T+h} - E_T(Y_{T+h})$.

Example: Forecasting and Forecasting Confidence Intervals

A linear time trend model is estimated on annual government bond interest rates from the year 2000 to 2020. The model's equation is given by:

$$R_t = 0.25 + 0.000154t + \hat{\epsilon}_t$$

The standard deviation of the forecasting error is estimated to be $\hat{\sigma} = 0.0245$. What is the 95% confidence interval for the second year if the forecasting residual errors (residuals) is a Gaussian white noise?

(Note that for the first time period t=2000 and the last time period is t=2020)

Solution

The second year starting from 2000 is 2002. So,

$$E_T(R_{2002}) = 0.25 + 0.000154 \times 2002 = 0.2808308$$

The 95% confidence interval is given by:

$$\begin{aligned} &E_T(Y_{T+h}) \pm 1.96\sigma \\ &= 0.28083 \pm 1.96 \times 0.0245 \\ &= [0.2328108, 0.3288508] \end{aligned}$$

So the 95% confidence interval for the interest rate is between 1.029% and 10.68%.

Question 1

The seasonal dummy model is generated on the quarterly growth rates of mortgages. The model is given by:

$$Y_t = \beta_0 + \sum_{j=1}^{s-1} \gamma_j D_{jt} + \epsilon_t$$

The estimated parameters are $\hat{\gamma}_1 = 6.25, \hat{\gamma}_2 = 50.52, \hat{\gamma}_3 = 10.25$ and $\hat{\beta}_0 = -10.42$ using the data up to the end of 2019. What is the forecasted value of the growth rate of the mortgages in the second quarter of 2020?

A. 40.10

B. 34.56

C. 43.56

D. 36.90

The correct answer is A.

We need to define the set of dummy variables:

$$D_{jt} = \begin{cases} 1, & \text{for } Q_2 \\ 0, & \text{for } Q_1, Q_3 \text{ and } Q_4 \end{cases}$$

So,

$$E(\hat{Y}_{Q_2}) = \beta_0 + \sum_{j=1}^{3} \gamma_j D_{jt} = -10.42 + 0 \times 6.25 + 1 \times 50.52 + 0 \times 10.25 = 40.1$$

Question 2

A mortgage analyst produced a model to predict housing starts (given in thousands) within

California in the US. The time series model contains both a trend and a seasonal component and is given by the following:

$$Y_t = 0.2t + 15.5 + 4.0 \times D_{2t} + 6.4 \times D_{3t} + 0.5 \times D_{4t}$$

The trend component is reflected in variable time(t), where (t) month and seasons are defined as follows:

Season	Months	Dummy
Winter	December, January, and February	
Spring	March, April, and May	D_{2t}
Summer	June, July, and August	D_{3t}
Fall	September, October, and November	D_{4t}

The model started in April 2019; for example, $y_{(T+1)}$ refers to May 2019.

What does the model predict for March 2020?

A. 21,700 housing starts

B. 22,500 housing starts

C. 24,300 housing starts

D. 20,225 housing starts

The correct answer is A.

The model is given as:

$$Y_t = 0.2t + 15.5 + 4.0 \times D_{2t} + 6.4 \times D_{3t} + 0.5 \times D_{4t}$$

Important: Since we have three dummies and an intercept, quarterly seasonality is reflected by the intercept (15.5) plus the three seasonal dummy variables (D_2, D_3, and D_4).

If Y_{T+1} = May 2019, then March 2020 = $Y_T + 11$

Finally, note that March falls under D_{2t}

$$y_{T+11} = 0.20 \times 11 + 15.5 + 4.0 \times 1 = 21.7$$

Thus, the model predicts 21,700 housing starts in March 2020.

Reading 23: Measuring Return, Volatility, and Correlation

After completing this reading, you should be able to:

- Calculate, distinguish, and convert between simple and continuously compounded returns.
- Define and distinguish between volatility, variance rate, and implied volatility.
- Describe how the first two moments may be insufficient to describe non-normal distributions.
- Explain how the Jarque-Bera test is used to determine whether returns are normally distributed.
- Describe the power law and its use for non-normal distributions.
- Define correlation and covariance and differentiate between correlation and dependence.
- Describe properties of correlations between normally distributed variables when using a one-factor model.

Measurement of Returns

A return is a profit from an investment. Two common methods used to measure returns include:

1. Simple Returns Method
2. Continuously Compounded Returns Method.

The Simple Returns Method

Denoted R_t the simple return is given by:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Where

P_{t-1}=Price of an asset at time t (current time)

P_{t-1}=Price of an asset at time t-1 (past time)

The time scale is arbitrary or shorter period such monthly or quarterly. Under the simple returns method, the returns over multiple periods is the product of the simple returns in each period. Mathematically given by:

$$1 + R_T = \prod_{t=i}^{T} (1 + R_t)$$

$$\Rightarrow R_T = (\prod_{t=i}^{T} (1 + R_t)) - 1$$

Example: Calculating the Simple Returns

Consider the following data.

Time	Price
0	100
1	98.65
2	98.50
3	97.50
4	95.67
5	96.54

Calculate the simple return based on the data for all periods.

Solution

We need to calculate the simple return over multiple periods which is given by:

$$1 + R_T = \prod_{t=i}^{T} (1 + R_t)$$

Consider the following table:

Time	Price	R_t	$1 + R_t$
0	100	−	−
1	98.65	−0.0135	0.9865
2	98.50	−0.00152	0.998479
3	97.50	−0.01015	0.989848
4	95.67	−0.01877	0.981231
5	96.54	0.009094	1.009094
		Product	0.9654

Note that

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

So that that

$$R_1 = \frac{P_1 - P_0}{P_0} = \frac{98.65 - 100}{100} = -0.0135$$

And

$$R_2 = \frac{P_2 - P_1}{P_1} = \frac{98.50 - 98.65}{98.65} = -0.00152$$

And so on.

Also note that:

$$\prod_{t=1}^{5} (1 + R_t) = 0.9865 \times 0.998479 \times \ldots \times 1.009094 = 0.9654$$

So,

$$1 + R_T = 0.9654 \Rightarrow R_T = -0.0346 = -3.46\%$$

Continuously Compounded Returns Method

Denoted by r_t. Compounded returns is the difference between the natural logarithm of the price of

assets at time t and t-1. It is given by:

$$r_t = \ln P_t - \ln P_{t-1}$$

Computing the compounded returns over multiple periods is easy because it is just the sum of returns of each period. That is:

$$r_T = \sum_{t=1}^{T} r_t$$

Example: Calculating Continuously Compounded Returns

Consider the following data.

Time	Price
0	100
1	98.65
2	98.50
3	97.50
4	95.67
5	96.54

What is the continuously compounded return based on the data over all periods?

Solution

The continuously compounded return over the multiple periods is given by

$$r_T = \sum_{t=1}^{T} r_t$$

Where

$$r_t = \ln P_t - \ln P_{t-1}$$

Consider the following table:

Time	Price	$r_t = \ln P_t - \ln P_{t-1}$
0	100	–
1	98.65	−0.01359
2	98.50	−0.00152
3	97.50	0.0102
4	95.67	−0.01895
5	96.54	0.009053
	Sum	−0.03521

Note that

$$r_1 = \ln P_1 - \ln P_0 = \ln 98.65 - \ln 100 = -0.01359$$
$$r_2 = \ln P_2 - \ln P_1 = \ln 98.50 - \ln 98.65 = -0.00152$$

And so on.

Also,

$$r_T = \sum_{t=1}^{5} r_t = -0.01359 + -0.00152 + \cdots + 0.009053 = -0.03521 = -3.521\%$$

Relationship between the Compounded and Simple Returns

Intuitively, the compounded returns is an approximation of the simple return. The approximation, however, is prone to significant error over longer time horizons, and thus compounded returns are suitable for short time horizons.

The relationship between the compounded returns and the simple returns is given by the formula:

$$1 + R_t = e^{r_t}$$

Example: Conversion Between the Simple and Compound Returns

What is the equivalent simple return for a 30% continuously compounded return?

Solution.

Using the formula:

$$1 + R_t = e^{r_t}$$
$$\Rightarrow R_t = e^{r_t} - 1 = e^{0}.3 - 1 = 0.3499 = 34.99\%$$

It is worth noting that compound returns are always less than the simple return. Moreover, simple returns are never less than -100%, unlike compound returns, which can be less than -100%. For instance, the equivalent compound return for -65% simple return is:

$$r_t = \ln(1 - 0.65) = -104.98\%$$

Measurement of Volatility and Risk

The volatility of a variable denoted as σ is the standard deviation of returns. The standard deviation of returns measures the volatility of the return over the time period at which it is captured.

Consider the linear scaling of the mean and variance over the period at which the returns are measured. The model is given by:

$$r_t = \mu + \sigma e_t$$

Where $E(r_t) = \mu$ is the mean of the return, $V(r_t) = \sigma^2$ is the variance of the return. e_t is the shocks, which is assumed to be iid distributed with the mean 0 and variance of 1. Moreover, the return is assumed to be also iid and normally distributed with the mean μ^2 i.e. $r_t \sim^{iid} N(\mu, \sigma^2)$. Note the shock can also be expressed as $\epsilon_t = \sigma e_t$ where: $\epsilon_t \sim N(0, \sigma^2)$.

Assume that we wish to calculate the returns under this model for 10 working days (two weeks). Since the model deals with the compound returns, we have:

$$\sum_{i=1}^{10} r_{t+i} = \sum_{i=1}^{10} (\mu + \sigma e_{t+i}) = 10\mu + \sigma \sum_{i=1}^{10} e_{t+i}$$

So that the mean of the return over the 10 days is 10μ and the variance also is $10\sigma^2$ since e_t is iid. The volatility of the return is, therefore:

$$\sqrt{10}\sigma$$

Therefore, the variance and the mean of return are scaled to the holding period while the volatility is scaled to the square root of the holding period. This feature allows us to convert volatility between different periods.

For instance, given daily volatility, we would to have yearly (annualized) volatility by scaling it by $\sqrt{252}$. That is:

$$\sigma_{annual} = \sqrt{252 \times \sigma^2_{daily}}$$

Note that 252 is the conventional number of trading days in a year in most markets.

Example: Calculating the Annualized Volatility

The monthly volatility of the price of gold is 4% in a given year. What is the annualized volatility of the gold price?

Solution

Using the scaling analogy, the corresponding annualized volatility is given by:

$$\sigma_{annual} = \sqrt{12 \times 0.04^2} = 13.86\%$$

Variance Rate

The variance rate, also termed as variance, is the square of volatility. Similar to mean, variance rate is linear to holding period and hence can be converted between periods. For instance, an annual variance rate from a monthly variance rate is given by

$$\sigma^2_{annual} = 12 \times \sigma^2_{monthly}$$

The variance of returns can be approximated as:

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t-1}^{T} (r_t - \hat{\mu})^2$$

Where $\hat{\mu}$ is the sample mean of return, and T is the sample size.

Example: Calculating the Variance of Return

The investment returns of a certain entity for five consecutive days is 6%, 5%, 8%,10% and 11%. What is the variance estimator of returns?

Solution

We start by calculating the sample mean:

$$\hat{\mu} = \frac{1}{5}(0.06 + 0.05 + 0.08 + 0.10 + 0.11) = 0.08$$

So that the variance estimator is:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t-1}^{T}(r_t - \hat{\mu})^2$$

$$= \frac{1}{5}[(0.06 - 008)^2 + (0.05 - 0.08)^2 + (0.08 - 0.08)^2 + (0.10 - 0.08)^2 + (0.11 - 0.08)^2] = 0.0052 = 0.52\%$$

The Implied Volatility

Implied volatility is an alternative measure of volatility that is constructed using options valuation. The options (both put and call) have payouts that are nonlinear functions of the price of the underlying asset. For instance, the payout from the put option is given by:

$$\max(K - P_T)$$

where P_T is the price of the underlying asset, K being the strike price, and T is the maturity period. Therefore, the price payout from an option is sensitive to the variance of the return on the asset.

The Black-Scholes-Merton model is commonly used for option pricing valuation. The model relates

the price of an option to the risk-free rate of interest, the current price of the underlying asset, the strike price, time to maturity, and the variance of return.

For instance, the price of the call option can be denoted by:

$$C_t = f(r_f, T, P_t, \sigma^2)$$

Where:

r_f= Risk-free rate of interest

T=Time to maturity

P_t=Current price of the underlying asset

σ^2=Variance of the return

The implied volatility σ relates the price of an option with the other three parameters. The implied volatility is an annualized value and does not need to be converted further.

The volatility index (VIX) measures the volatility in the S&P 500 over the coming 30 calendar days. VIX is constructed from a variety of options with different strike prices. VIX applies to a large variety of assets such as gold, but it is only applicable to highly liquid derivative markets and thus not applicable to most financial assets.

The Financial Returns Distribution

The financial returns are assumed to follow a normal distribution. Typically, a normal distribution is thinned-tailed, does not have skewness and excess kurtosis. The assumption of the normal distribution is sometimes not valid because a lot of return series are both skewed and mostly heavy-tailed.

To determine whether it is appropriate to assume that the asset returns are normally distributed, we use the Jarque-Bera test.

The Jarque-Bera Test

Jarque-Bera test tests whether the skewness and kurtosis of returns are compatible with that of normal distribution.

Denoting the skewness by S and kurtosis by k, the hypothesis statement of the Jarque-Bera test is stated as:

H_0 : S = 0 and k=3 (the returns are normally distributed)

vs

H_1 : S ≠ 0 and k ≠ 3 (the returns are not normally distributed)

The test statistic (JB) is given by:

$$JB = (T-1)\left(\frac{\hat{S}^2}{6} + \frac{(\hat{k}-3)^2}{24}\right)$$

Where T is the sample size.

The basis of the test is that, under normal distribution, the skewness is asymptotically normally distributed with the variance of 6 so that the variable $\frac{\hat{S}^2}{6}$ is chi-squared distributed with one degree of freedom (χ_1^2) and kurtosis is also asymptotically normally distributed with the mean of 3 and variance of 24 so that $\frac{(\hat{k}-3)^2}{24}$ is also (χ_1^2) variable. Coagulating these arguments given that these variables are independent, then:

$$JB \sim \chi_2^2$$

The Decision Rule of the JB Test

When the test statistic is greater than the critical value, then the null hypothesis is rejected. Otherwise, the alternative hypothesis is true. We use the χ_2^2 table with the appropriate degrees of freedom:

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22

For example, the critical value of a χ_2^2 at a 5% confidence level is 5.991, and thus, if the computed test statistic is greater than 5.991, the null hypothesis is rejected.

Example: Conducting a JB Test

Investment return is such that it has a skewness of 0.75 and a kurtosis of 3.15. If the sample size is 125, what is the JB test statistic? Does the data qualify to be normally distributed at a 95% confidence level?

Solution

The test statistic is given by:

$$JB = (T-1)\left(\frac{\hat{S}^2}{6} + \frac{(\hat{k}-3)^2}{24}\right) = (125-1)\left(\frac{0.75^2}{6} + \frac{(3.15-3)^2}{24}\right) = 11.74$$

Since the test statistic is greater than the 5% critical value (5.991), then the null hypothesis that the data is normally distributed is rejected.

The Power Law

The power law is an alternative method of determining whether the returns are normal or not by

studying the tails. For a normal distribution, the tail is thinned, such that the probability of any return greater than kσ decreases sharply as *k increases*. Other distributions are such that their tails decrease relatively slowly, given a large deviation.

The power law tails are such that, the probability of observing a value greater than a given value x defined as:

$$P(X > x) = kx^{-\alpha}$$

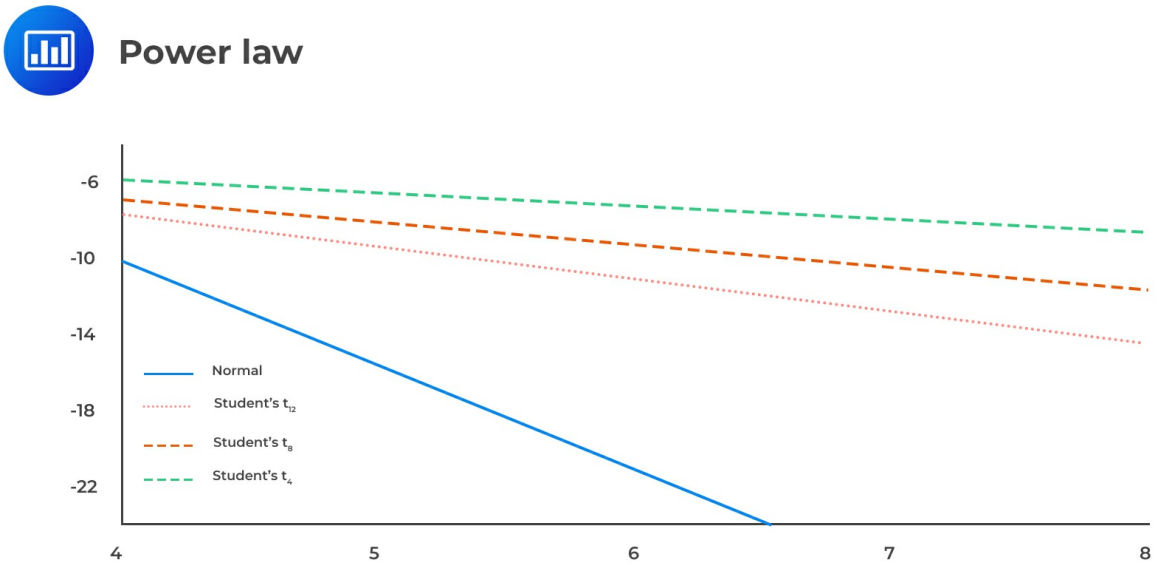
Where k and α are constants.

The tail behavior of distributions is effectively compared by considering the natural log (ln(P(X>x))) of the tail probability. From the above equation:

$$\ln \text{prob}(X > x) = \ln k - \alpha \ln x$$

To test whether the above equation holds, a graph of ln prob(X > x) plotted against lnx.

For a normal distribution, the plot is quadratic in x, and hence it decays quickly, meaning that they have thinned tails. For other distributions such as Student's t distribution, the plots are linear to x, and thus, the tails decay at a slow rate, and hence they have fatter tails (produce values that are far from the mean).

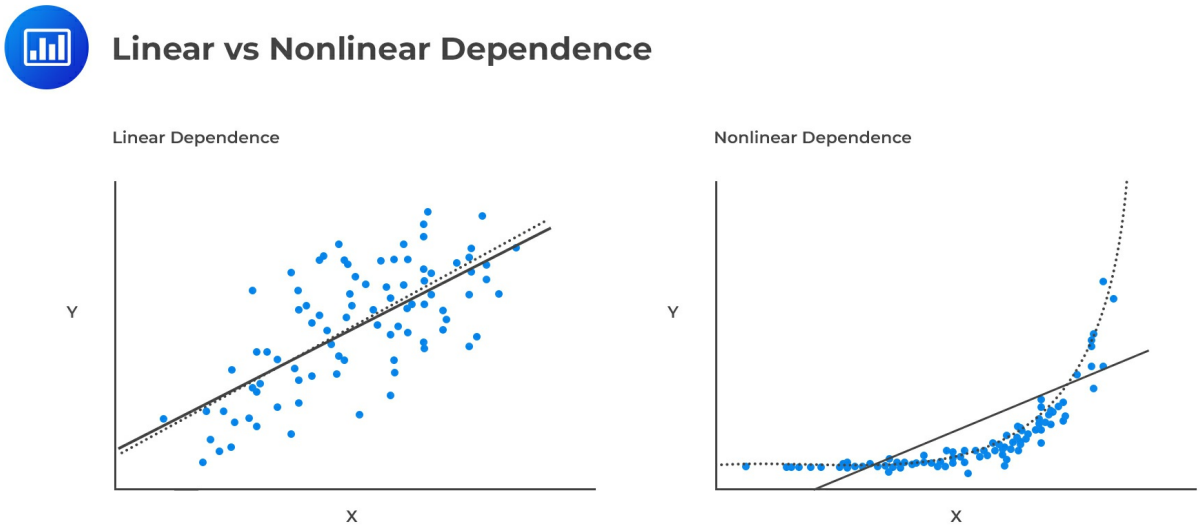


Dependence and Correlation of Random Variables.

The two random variables X and Y are said to be independent if their joint density function is equal to the product of their marginal distributions. Formally stated:

$$f_{X,Y} = f_X(x).f_Y(y)$$

Otherwise, the random variables are said to be dependent. The dependence of random variables can be linear or nonlinear.



The linear relationship of the random variables is measured using the correlation estimator called Pearson's correlation.

Recall that given the linear equation:

$$Y_i = \alpha + \beta_i X_i + \epsilon_i$$

The slope β is related to the correlation coefficient ρ. That is, if $\beta = 0$, then the random variables X_i and Y_i are uncorrelated. Otherwise, $\beta \neq 0$. Infact, if the variances of the random variables are engineered such that they are both equal to unity ($\sigma_X^2 = \sigma_Y^2 = 1$), the slope of the regression equation is equal to the correlation coefficient ($\beta = \rho$). Thus, the regression equation reflects how the correlation measures the linear dependence.

Nonlinear dependence is complex and thus cannot be summarized using a single statistic.

Measures of Correlation

The correlation is mostly measured using the rank correlation (Spearman's rank correlation) and Kendal's τ correlation coefficient. The values of the correlation coefficient are between -1 and 1. When the value of the correlation coefficient is 0, then the random variables are independent; otherwise, a positive (negative) correlation indicates an increasing (a decreasing) relationship between the random variables.

Rank Correlation

The rank correlation uses the ranks of observations of random variables X and Y. That is, rank correlation depends on the linear relationship between the ranks rather than the random variables themselves.

The ranks are such that 1 is assigned to the smallest value, 2 to the next value, and so on until the largest value is assigned n.

When a rank repeats itself, an average is computed depending on the number of repeated variables, and each is assigned the averaged rank. Consider the ranks 1,2,3,3,3,4,5,6,7,7. Rank 3 is repeated

three times, and rank 7 is repeated two times. For the repeated 3's, the averaged rank is $\frac{(3+4+5)}{3} = 4$. For the repeated 7's the averaged rank is $\frac{(9+10)}{2} = 8.5$. Note that we are averaging the ranks, which the repeated ranks could have to assume if they were not repeated. So the new ranks are:1,2,4,4,4,4,5,6,8.5,8.5.

Now, denote the rank of X by R_X and that of Y by R_Y then the rank correlation estimator is given by:

$$\hat{\rho}_s = \frac{\text{Cov}(\hat{R_X}, R_Y)}{\sqrt{\hat{V}(R_X)}\sqrt{\hat{V}(R_Y)}}$$

Alternatively, when all the ranks are distinct (no repeated ranks), the rank correlation estimator is estimated as:

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n}(R_{X_i} - R_{Y_i})^2.}{n(n^2-1)}$$

The intuition of the last formula is that when a highly ranked value of X is paired with corresponding ranked values of Y, then the value of $R_{X_i} - R_{Y_i}$ is very small and thus, correlation tends to 1. On the other, if the smaller rank values of X are marched with larger rank values of Y, then $R_{X_i} - R_{Y_i}$ is relatively larger and thus, correlation tends to -1.

When the variables X and Y have a linear relationship, linear and rank, correlations have equal value. However, rank correlation is inefficient compared to linear correlation and only used for confirmational checks. On the other hand, rank correlation is insensitive to outliers because it only deals with the ranks and not the values of X and Y.

Example: Calculating the Rank Correlation

Consider the following data.

i	X	Y
1	0.35	2.50
2	1.73	6.65
3	−0.45	−2.43
4	−0.56	−5.04
5	4.03	3.20
6	3.21	2.31

What is the value of rank correlation?

Solution

Consider the following table where the ranks of each variable have been filled and the square of their difference in ranks.

i	X	Y	R_X	R_Y	$(R_X - R_Y)^2$
1	0.35	2.50	4	3	1
2	1.73	6.65	3	1	4
3	−0.45	−2.43	5	5	0
4	−0.56	−5.04	6	6	0
5	4.03	3.20	1	2	1
6	3.21	2.31	2	4	4
				Sum	10

Since there are no repeated ranks, then the rank correlation is given by:

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n}(R_{X_i} - R_{Y_i})^2.}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 10}{6(6^2-1)} = 1 - 0.2857 = 0.7143$$

The Kendal's Tau (τ)

Kendal's Tau is a non-parametric measure of the relationship between two random variables, say, X and Y. Kendal's τ compares the frequency of concordant and discordant pairs.

Consider the set of random variables X_i and Y_i. These pairs are said to be concordant for all $i \neq j$ if the ranks of the components agree. That is, $X_i > X_j$ when $Y_i > Y_j$ or $X_i < X_j$ when $Y_i < Y_j$. That is, they

are concordant if they agree on the same directional position (consistent). When the pairs disagree, they are termed as discordant. Note that ties are neither concordant nor discordant.

Intuitively, random variables with a high number of concordant pairs have a strong positive correlation, while those with a high number of discordant pairs are negatively correlated.

The Kendal's Tau is defined as:

$$\hat{\tau} = \frac{n_c - n_d}{\frac{n(n-1)}{2}} = \frac{n_c}{n_c + n_d + n_t} - \frac{n_d}{n_c + n_d + n_t}$$

Where

n_c=number of concordant pairs

n_d=number of discordant pairs

n_t=number of ties

It is easy to se that Kendal's Tau is equivalent to the difference between the probabilities of concordance and discordance. Moreover, when all the pairs are concordant, $\hat{\tau} = 1$ and when all pairs are discordant, $\hat{\tau} = -1$.

Example: Calculating the Kendall's Tau

Consider the following data (same as the example above).

i	X	Y
1	0.35	2.50
2	1.73	6.65
3	−0.45	−2.43
4	−0.56	−5.04
5	4.03	3.20
6	3.21	2.31

What is Kendall's τ correlation coefficient?

Solution

The first step is to rank each data:

i	X	Y	R_X	R_Y
1	0.35	2.50	4	3
2	1.73	6.65	3	1
3	−0.45	−2.43	5	5
4	−0.56	−5.04	6	6
5	4.03	3.20	1	2
6	3.21	2.31	2	4

Next is to arrange ranks in order of rank X, then the concordant (C) pairs are the number of ranks greater than the given rank of Y, and discordant pairs are the number of ranks less than the given rank of Y.

R_X	R_Y	C	D
1	2	4	1
2	4	2	2
3	1	3	0
4	3	2	0
5	5	1	0
6	6		
Total		12	3

Note that, C=4, are the number of ranks greater than 2 (4,3,5 and 6) below it. Also, D=1 is the number of ranks less than 2 below it. This is continued up to the second last row since there are no more ranks to look up.

So, $n_c = 12$ and $n_d = 3$

$$\Rightarrow \hat{\tau} = \frac{n_c - n_d}{\frac{n(n-1)}{2}} = \frac{12-3}{\frac{6(6-1}{2}} = \frac{9}{15} = 0.600$$

Practice Question

Suppose that we know from experience that $\alpha = 3$ for a particular financial variable, and we observe that the probability that $X > 10$ is 0.04.

Determine the probability that X is greater than 20.

A. 125%

B. 0.5%

C. 4%

D. 0.1%

The correct answer is **B.**

From the given probability, we can get the value of constant k as follows:

$$\begin{aligned} \text{prob}(X > x) &= kx^{(-\alpha)} \\ 0.04 &= k(10)^{(-3)} \\ k &= 40 \end{aligned}$$

Thus,

$$P(X > 20) = 40(20)^{(-3)} = 0.005 \text{ or } 0.5\%$$

Note: The power law provides an alternative to assuming normal distributions.

Reading 24: Simulation and Bootstrapping

After completing this reading, you should be able to:

- Describe the basic steps to conduct a Monte Carlo simulation.
- Describe ways to reduce the Monte Carlo sampling error.
- Explain the use of antithetic and control variates in reducing Monte Carlo sampling error.
- Describe the bootstrapping method and its advantage over the Monte Carlo simulation.
- Describe pseudo-random number generation.
- Describe situations where the bootstrapping method is ineffective.
- Describe the disadvantages of the simulation approach to financial problem-solving.

Introduction and Definitions

Simulation is a way of modeling random events to match real-world outcomes. By observing **simulated** results, researchers gain insight into real problems. Examples of the application of the simulation are the calculation of option payoff and determining the accuracy of an estimator. Some of the simulation methods are the Monte Carlo Simulation (Monte Carlo) and the Bootstrapping.

Monte Carlo Simulation approximates the expected value of a random variable using the numerical methods. The Monte Carlo generates the random variables from an assumed data generating process (DGP), and then it applies a function(s) to create realizations from the unknown distribution of the transformed random variables. This process is repeated (to improve the accuracy), and the statistic of interest is then approximated using the simulated values.

Bootstrapping is a type of simulation where it uses the observed variables to simulate from the unknown distribution that generates the observed variables. In other words, bootstrapping involves the combination of the observed data and the simulated values to create a new sample that is related but different from the observed data.

The notable similarity between Monte Carlo and bootstrapping is that both aim at calculating the expected value of the function by using simulated data (often by use of a computer).

Also, the contrasting feature in these methods is that in Monte Carlo simulation, a data generating process (DGP) is entirely used to simulate the data. However, in bootstrapping, observed data is used to generate the simulated data without specifying an underlying DGP.

Simulation of Random Variables

The simulation requires the generation of random variables from an assumed distribution, mostly using a computer. However, computer-generated numbers are not necessarily random and thus termed as **pseudo-random numbers**. Pseudo numbers are produced by the complex deterministic functions (pseudo number generators, PNGs), which seem to be random. The initial values of pseudo numbers are termed as a **seed value**, which is usually unique but generates similar random variables when PNRG runs.

The ability of the simulated variables from PRNGs to replicate makes it possible to use pseudo numbers across multiple experiments because the same sequence of random variables can be generated using the same seed value. Therefore, we can use this feature to choose the best model or reproduce the same results in the future in case of regulatory requirements. Moreover, the corresponding random variables can be generated using different computers.

Simulating Random Variables from a Specific Distribution

Simulating random variables from a specific distribution is initiated by first generating a random number from a uniform distribution (0,1). After that, the cumulative distribution of the distribution we are trying to simulate is used to get the random values from that distribution. That is, we first generate a random number U from U(0,1) distribution, then, we use the generated random number to simulate a random variable X with the pdf **f(x)** by using the CDF, **F(x)**.

Let U be the probability that X takes a value less than or equal to x, that is,

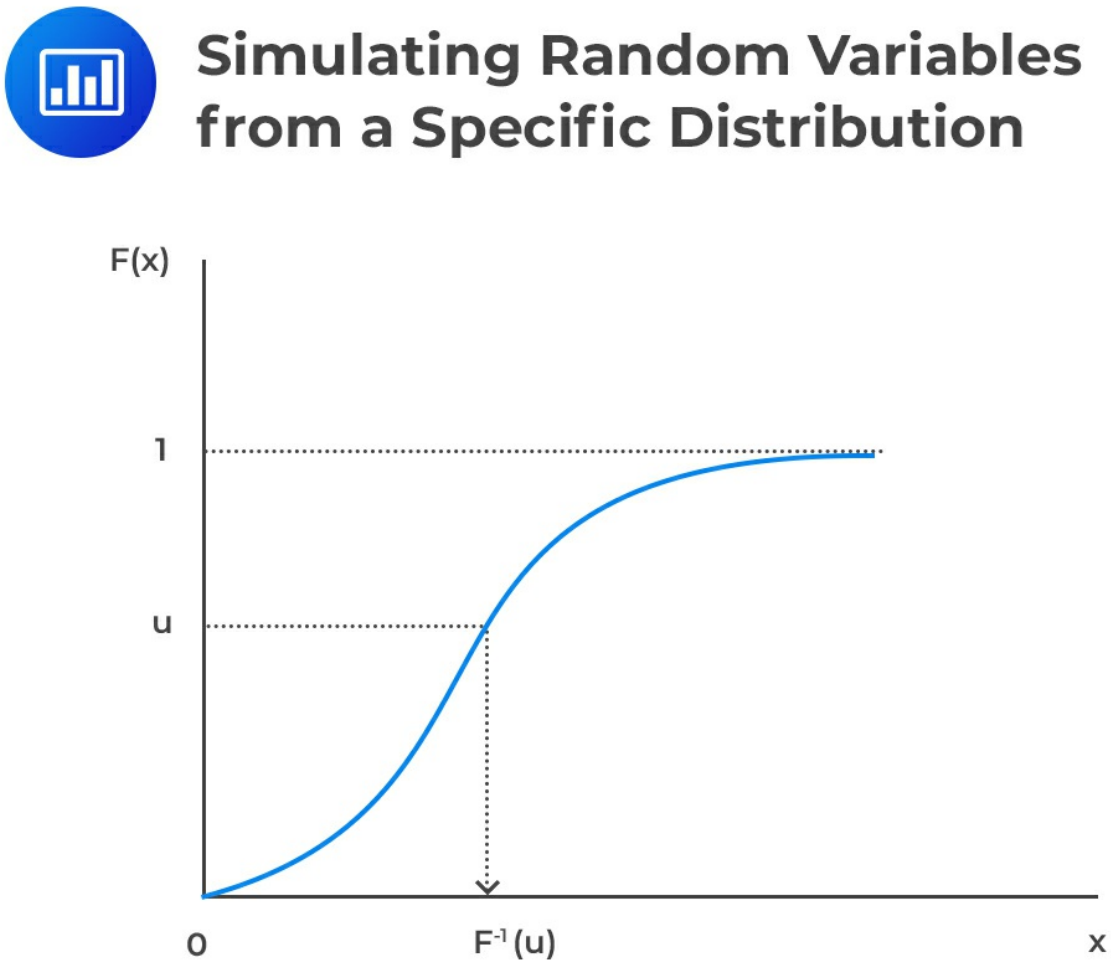
$$U = P(\leq x) = F(x)$$

Then we can derive the random variable x as:

$$x = F^{-1}(u)$$

To put this in a more straightforward perspective, the algorithm for simulating random variable from a specific distribution involves:

1. Generating a random variable u from the uniform distribution $U(0,1)$
2. Compute $x = F^{-1}(u)$



Note that the random variable X has a CDF F(x) as shown below:

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

Example: Generating Random Variables from Exponential Distribution

Assume that we want to simulate three random variables from an exponential distribution with a parameter $\lambda = 0.2$ using the value 0.112, 0.508, and 0.005 from U(0,1).

Solution

This question assumes that the uniform random variable has been generated. The inverse of the CDF of exponential distribution is given by:

$$F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x)$$

So, in this case:

$$F^{-1}(x) = -\frac{1}{0.2}\ln(1-x)$$
$$x = -\frac{1}{0.2}\ln(1-u)$$

So the random variables are:

$$x_1 = -\frac{1}{0.2}\ln(1-u_1) = -5\ln(1-0.112) = 2.37567$$
$$x_2 = -\frac{1}{0.2}\ln(1-u_2) = -5\ln(1-0.508) = 14.1855$$
$$x_3 = -\frac{1}{0.2}\ln(1-u_3) = -5\ln(1-0.005) = 0.10025$$

The random variables are 2.37567, 14.1855 and 0.10025

Monte Carlo Simulation

Monte Carlo simulation is used to estimate the population moments or functions. The Monte Carlo is as follows:

Assume that X is a random variable that can be simulated and let g(X) be a function that can be

evaluated at the realizations of X. Then, the simulation generates multiple copies of g(X) by simulating draws from $X = x_j$ and calculate $g_i = g(x_i)$.

This process is then repeated b times so that a set of iid variables is generated from the unknown distribution g(X), which can then be used to estimate the desired statistic.

For instance, if we wish to estimate the mean of the generated random variables, then the mean is given by:

$$\hat{E}(g(X)) = \frac{1}{b}\sum_{i=1}^{b} g(X_i)$$

This is true because the generated variables are iid, and then the process is repeated b times. Consequently, by the law of large number (LLN),

$$\lim_{b\to\infty} \hat{E}(g(X)) = E(g(x))$$

Also, the Central Limit Theorem applies to the estimated mean so that:

$$Var[\hat{E}(g(X))] = \frac{\sigma_g^2}{b}$$

Where $\sigma_g^2 = Var(g(X))$

The second moment, which is the variance (standard variance estimator) is estimated as:

$$\hat{\sigma}_g^2 = \frac{1}{b}\sum_{i=1}^{b} (g(X_i) - E[\hat{g}(X)])^2$$

From CLT, the standard error of the simulated expectation is given by:

$$\sqrt{\frac{\sigma_g^2}{b}} = \frac{\sigma_2}{\sqrt{b}}$$

The standard error of the simulated expectation measures the level of accuracy of the estimation; thus, the choice of b determines the accuracy of the simulation.,

Another quantity that can be calculated from the simulation is the α-quantile by arranging the **b** draws in ascending order then selecting the value $b\alpha$ of the sorted set.

Moreover, using the simulation, we determine the finite sample properties of the estimated parameters. Assume that the sample size n is large enough so that approximation by CLT is adequate. Now, consider a finite-sample distribution of a parameter $\hat{\theta}$. Using the assumed DGP, n random samples are generated so that:

$$X = [x_1, x_2, \dots, x_n]$$

We need to estimate a parameter $\hat{\theta}$.

We would need to simulate new data set and estimate the parameter b times: $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_b)$ from the finite-sample distribution of the estimator of θ. From these values, we can rule out the properties of the estimator $\hat{\theta}$. For instance, the bias defined as:

$$\text{Bias}(\theta) = E(\hat{\theta}) - \theta$$

That can be approximated as:

$$(\hat{\text{Bias}})(\theta) = \frac{1}{b}\sum_{i=1}^{b}(\hat{\theta}_i - \theta)$$

Having the basics of the Monte Carlo simulation, its basic logarithm is as follows:

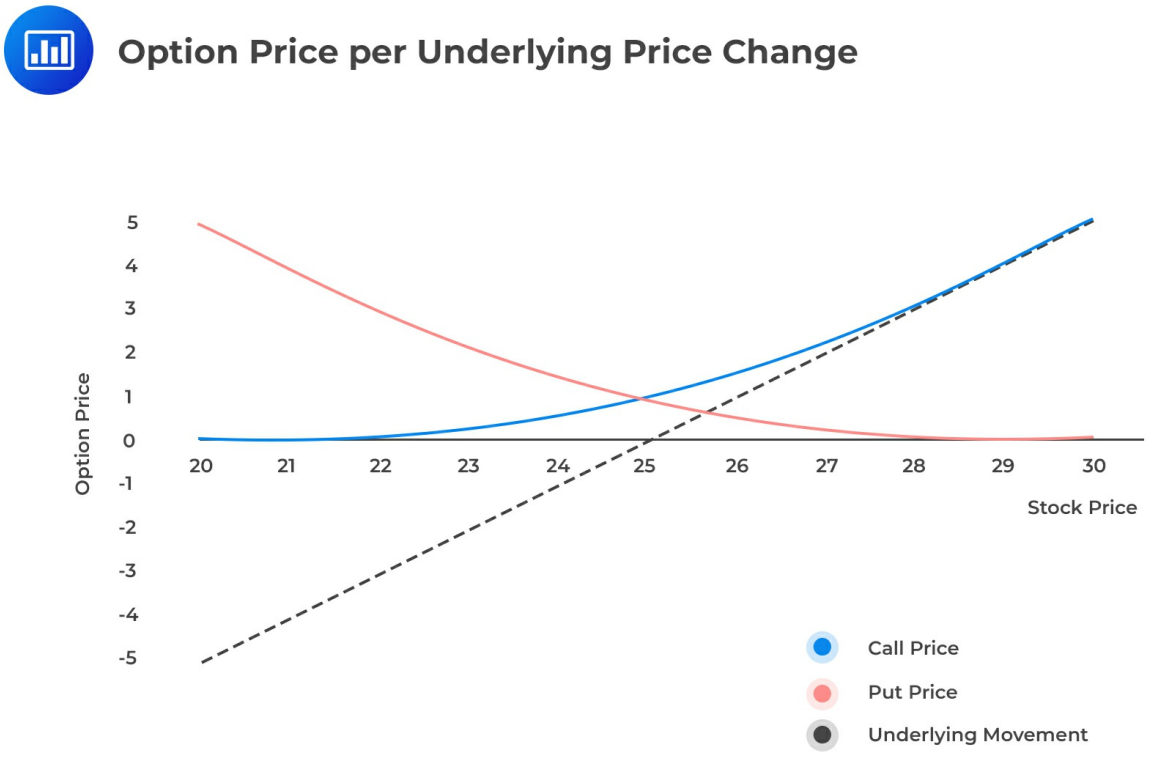
i. Generate the data: $x_i = [x_{1i}, x_{2i}, \dots, x_{ni}]$ by using the assumed DGP.
ii. Compute the desired function or statistic $g_i = g(x_i)$.
iii. Iterate steps 1 and 2 b times.
iv. From the replications $\{g_1, g_2, \dots, g_b\}$, calculate the statistic of interest.
v. Determine the accuracy of the estimated quantity by calculating the standard error. If the standard error is huge, increase the number of b-replications to obtain the smallest error possible.

Example: Using the Monte Carlo Simulation to Estimate the Price of a Call Option

Recall that the price of a call option is given by:

$$\max(0, S_T - K)$$

S_T is the price of the underlying stock at the time of maturity T, and K is the strike price. The price of the call option is a non-linear function of the underlying stock price at the expiration date, and thus, we can model the price of the call option.



Assuming that the log of the stock price is normally distributed, then the price of the stock can be modeled as the sum of the initial stock price, a mean and normally distributed error. Mathematically stated as:

$$s_T = s_0 + T\left(r_f - \frac{\sigma^2}{2}\right)\sqrt{T}x_i$$

Where

s_0= the initial stock price

T = time to maturity in years

r_f = annualized time to maturity

σ^2 = variance of the stock return

x_i = simulated values from $N(0, \sigma^2)$

From the formula above, to simulate the price of the underlying stock requires the estimation of the stock volatility.

Using the simulated price of the stock, the price of the option can be calculated as:

$$c = e^{(-r_f T)} \max(S_T - K, 0)$$

And thus the mean of the price of the call option can be estimated as:

$$\hat{E}(c) = \bar{c} = \frac{1}{b} \sum_{i=1}^{b} c_i$$

Where c_i is the simulated payoffs of the call option. Note that, using the equation, $s_T = s_0 + T\left(r_f - \frac{\sigma^2}{2}\right)\sqrt{T}x_i$, the simulated stock prices can be expressed as:

$$S_{Ti} = e^{s_0 + T\left(r_f - \frac{\sigma^2}{2}\right) + \sqrt{T}x_i}$$

And thus

$$g(x_i) = c_i = e^{(-rT)} \max\left(e^{s_0 + T(r_f - \frac{\sigma^2}{2}) + \sqrt{T}x_i} - K, 0\right)$$

The standard error of the call option price is given by:

$$s.e\left(\hat{E}(c)\right) = \sqrt{\frac{\hat{\sigma}_g^2}{b}} = \frac{\hat{\sigma}_g}{\sqrt{b}}$$

Where $\hat{\sigma}_g^2$

$$\hat{\sigma}_g^2 = \frac{1}{b}\sum_{\forall i}(c_i - \hat{c})^2$$

Given that we calculate the standard error, we can calculate the confidence intervals for the estimated mean of the call option price. For instance, the 95% confidence interval; is given by:

$$\hat{E}(c) \pm 1.96 \text{ s.e } (\hat{E}(c))$$

Reducing Monte Carlo Sampling Error

Sampling error in Monte Carlo simulation is reduced by two complementary methods:

1. Antithetic Variables, and
2. Control Variates.

These methods can be used simultaneously.

To set the mood, recall that the estimation of expected values in simulation depends on the Law of Large Numbers (LLN) and that the standard error of the estimated expected value is proportional to $1/\sqrt{b}$. Therefore, the accuracy of the simulation depends on the variance of the simulated quantities.

Antithetic Variables

Recall variance between two random variables X and Y is given by:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

Otherwise, if the variables are independent, then:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Moreover, if the covariance between the variables is negative (or negatively correlated), then:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$$

The antithetic variables use the last result. The antithetic variables reduce the sampling error by

incorporating the second set of variables that are generated in such a way that they are negatively correlated with the initial iid simulated variables. That is, each simulated variable is paired with an antithetic variable so that they occur in pairs and are negatively correlated.

If U_1 is a uniform random variable, then:

$$F^{-1}(U_1) \sim F_x$$

Denote an antithetic variable U_2 which is generated using:

$$U_2 = 1 - U_1$$

Note that U_2 is also a uniform random variable so that:

$$F^{-1}(U_2) \sim F_x$$

Then by definition of antithetic variables, the correlation between U_1 and U_2 is negative as well as their mappings onto the CDF F_x.

Using the antithetic random variables is analogous to typical Monte Carlo simulation only that values are constructed in pairs $[\{U_1, 1-U_1\}, \{U_2, 1-U_2\}, \ldots, \{U_{\frac{b}{2}}, 1-U_{\frac{b}{2}}\}]$ which are then transformed to have the desired distribution using the inverse CDF.

Note that the number of simulations is b/2 since the simulation values are in pairs. The antithetic variables reduce the sampling error only if the function g(X) is monotonic in x so that $Corr(x_i, -x_i) = Corr(g(x_i), g(-x_i))$.

Notably, the antithetic random variables reduce the sampling error through the correlation coefficient. Note that usually sampling error using b iid simulated values, is

$$\frac{\sigma_g}{\sqrt{b}}$$

But by introducing the antithetic random variables, then the standard error is given by:

$$\frac{\sigma_g\sqrt{1+\rho}}{\sqrt{b}}$$

Clearly, the standard error decreases when the correlation coefficient, $\rho < 0$.

Control Variates

Control variates reduce the sampling error by incorporating values that have a mean of zero and correlated to simulation. The control variates have a mean of zero so that it does not bias the approximation. Given that the control variate and the desire function are correlated, an effective combination (optimal weights) of the control variate and the initial simulation value to reduce the variance of the approximation.

Recall that expected value is approximated as:

$$\hat{E}[g(X)] = \frac{1}{b}\sum_{i=1}^{b} g(x_i)$$

Since this estimate is consistent, we can break down to:

$$\hat{E}[g(X)] = E[g(X)] + \eta_i$$

Where η_i is a mean zero error. That is: $E(\eta_i) = 0$

Denote the control variate by $h(X_i)$ so that by definition, $E[h(X_i)] = 0$ and that it is correlated with η_i.

An ideal control variate should be less costly to construct and that it should be highly correlated with $g(X)$ so that the optimal combination parameter β_0 that minimizes the estimation errors can be approximated by the regression equation:

$$g(x_i) = \beta_0 + \beta_1 h(X_i)$$

Disadvantages of Simulation

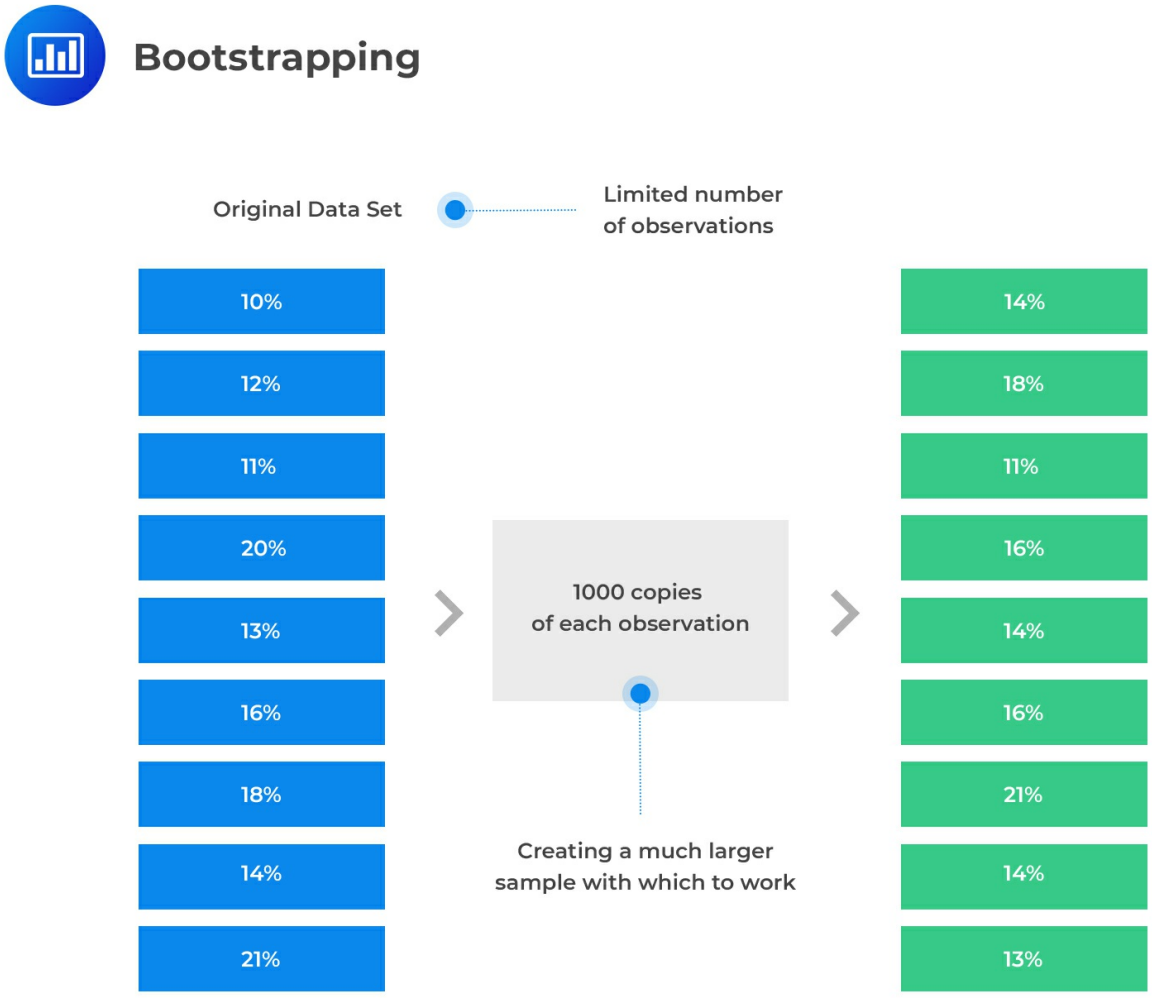
- Monte Carlo Simulation can result in unreliable approximates of moments if the DGPs used do not adequately describe the observed data. This mostly occurs due to misspecifications

of the DGP.

- Simulation can be costly, especially when you are running multiple simulation experiments because it can be time-consuming.

Bootstrapping

As stated earlier, bootstrapping is a type of simulation where it uses the observed variables to simulate from the unknown distribution that generates the observed variable. However, note that bootstrapping does not directly model the observed data or suggest any assumption about the distribution, but rather, the unknown distribution in which the sample is drawn is the origin of the observed data.



There are two types of bootstraps:

i. iid Bootstraps
ii. Circular Blocks Bootstraps (CBB)

iid Bootstrap

iid bootstraps select the samples that are constructed with replacement from the observed data. Assume that a simulation sample of size m is created from the observed data with n observations. iid bootstraps construct observation indices by randomly sampling with replacing from the values 1,2,..., n. These random indices are then used to draw the observed data to be included in the simulated data

(bootstrap sample).

For instance, assume we want to draw 10 observations from a sample of 50 data points: $\{x_1, x_2, x_3, \dots, x_{50}\}$. The first simulation could use $\{x_1, x_{12}, x_{23} x_{11}, x_{32}, x_{43} x_1, x_{22}, x_2, x_{22}\}$observations and second simulation could use $\{x_{50}, x_{21}, x_{23} x_{19}, x_{32}, x_{49} x_{41}, x_{22}, x_{12}, , x_{39}\}$ and so on until the desired number of simulations is reached.

In other words, iid bootstrap is analogous to Monte Carlo Simulation, where bootstrap samples are used instead of simulated samples. Under iid bootstrap, the expected values are estimated as:

$$\hat{E}[g(X)] = \frac{1}{b} \sum_{i=1}^{b} g(x_{1,j,}^{BS} x_{2,j,}^{BS}, \dots, x_{m,j,}^{BS})$$

Where

$x_{i,j,}^{BS}$ = observation i from observation j

b = total number of bootstraps samples

The iid bootstrap is suitable when observations used are independent over time, and thus using it in financial analysis is unsuitable because most of the financial data is dependent.

In short, the logarithm of generating a sample using the iid bootstrap include:

i. Create a random set of m integers $(i_1, i_2, \dots, i_m)$ from (1,2,…,n) with replacement.
ii. Construct the bootstrap sample as $x_{i_1}, x_{i_2}, \dots, x_{i_m}$

Circular Block Bootstrap (CBB)

The circular block bootstrap differs from the iid bootstrap in that instead of sampling each data point with replacement, it samples the blocks of size q with replacement. For instance, assume that we have 50 observations which are sampled into five blocks (q=5), each with 10 observations.

The blocks are sampled with replacement until the desired sample size is produced. In the case that the number of observations in sampled blocks is larger than the required sample size, some of the observations are omitted in the last block.

The size of the number of blocks should be large enough to reflect the dependence of observations but not too large to exclude some crucial blocks. Conventionally, the size of the blocks is the square root of the sample size ($\sqrt{n}$).

The general steps of generating sample using the CBB are:

i. Decide on the size of block q-more preferably, the block size should be equal to the square root of the sample size, i.e $\sqrt{n}$.
ii. Select the first block index i from (1,2,…,n) and transfer $\{x_i, x_{i+1}, \ldots, x_{i+q}\}$ to the bootstrap sample where the indices larger (i>n) wrap around.
iii. Incase the bootstrap sample has less than m elements, repeat step (ii) above.
iv. In case the bootstrap sample has more than m elements, omit the values from the end of the bootstrap sample until the sample size is m.

Application of Bootstrapping

One of the applications of bootstrapping is the estimation of the p-value at risk in financial markets. Recall the p-value at risk (p-VaR) is defined as:

$$\underset{\text{Var}}{\text{argmin}} \Pr(L > \text{VaR}) = 1 - p$$

Where:

L = loss of the portfolio over a given period, and

1-p = the probability that the loss occurs.

If the loss is measured in percentages of a particular portfolio, then p-VaR can be seen as a quantile of the return distribution. For instance, if we wish to calculate a one-year VaR of a portfolio, then we will simulate a one-year data (252 days) and then find the quantile of the simulated annual returns.

The VaR is then calculated by sorting the bootstrapped annual returns from lowest to highest and then determining (1-p)b, which is basically the empirical 1-p quantile of the annual returns.

Situations Where Bootstrap Will be Ineffective

The following are the two situations where bootstraps will not be sufficiently effective:

- In cases where there are outliers in the data, hence there is a likelihood that the bootstrap's conclusion will be affected.
- Non-independent data – When a bootstrap is applied, the assumption the data are independent of one another.

Disadvantages of Bootstrapping

- Bootstrapping uses the whole data to generate a simulated sample and thus may make the simulated sample unreliable when the past and the present data are different. For example, the present state of a financial market might be different from the past.
- Bootstrapping of historical data can be unreliable due to changes in the market so that the present is different from the past. For instance, if we are bootstrapping market interest rates, there might be huge discrepancies due to past and present market forces, which cause the interest rate to fluctuate significantly.

Comparison between Monte Carlo Simulation and Bootstrapping

Monte Carlo simulation uses an entire statistical model that incorporates the assumption on the distribution of the shocks, and therefore, the results are inaccurate if the model used is poor even when the replications are significantly large.

On the other hand, bootstrapping does not specify the model but instead assumes the past resembles the present of the data. In other words, the bootstrapping incorporates the aspect of the dependence of the observed data to reflect the sampling variation.

Both Monte Carlo Simulation and bootstrapping are affected by the "Black Swan" problem, where the resulting simulations in both methods closely resemble historical data. In other words, simulations tend to focus on historical data, and thus, the simulations are not so different from what it has been observed.

Practice Question

Which of the following statements correctly describes an antithetic variable?

A. They are variables that are generated to have a negative correlation with the initial simulated sample.
B. They are mean zero values that are correlated to the desired statistic that is to be computed from through simulation.
C. They are the mean zero variables that are negatively correlated with the initial simulated sample.
D. None of the above

Solution

The correct answer is **A.**

Control variates are used to reduce the sampling error in the Monte Carlo simulation. They are constructed to have a negative correlation with the initial simulated sample so that the overall standard error of approximation is reduced.