# Probability and Statistics

## DECAP790

**Edited by**
**Sartaj Singh**

# Probability and Statistics

**Edited By:**
**Sartaj Singh**

# CONTENT

# Unit 01: Introduction to Probability

## Objectives

- Understand basics of statistics and probability.
- Learn concepts of set theory.
- Define basic terms of Sampling.
- Understand concept of conditional probability.
- Solve basic questions related to probability

## Introduction

**Probability and Statistics** are the two important concepts in Math's. Probability is all about chance. Whereas statistics is more about how we handle various data using different techniques. It helps to represent complicated data in a very easy and understandable way. The statistic has a huge application nowadays in data science professions. The professionals use the stats and do the predictions of the business. It helps them to predict the future profit or loss attained by the company. Probability denotes the possibility of the outcome of any random event. The meaning of this term is to check the extent to which any event is likely to happen.

For example, when we flip a coin in the air, what is the possibility of getting a head? The answer to this question is based on the number of possible outcomes. Here the possibility is either head or tail will be the outcome. So, the probability of a head to come as a result is 1/2.

The probability is the measure of the likelihood of an event to happen. It measures the certainty of the event. The formula for probability is given by;

**P(E) = Number of Favorable Outcomes/Number of total outcomes**

**P(E) = n(E)/n(S)**

Here,

n(E) = Number of event favorable to event E

n(S) = Total number of outcomes

## 1.1  What is Statistics?

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. It is a method of collecting and summarizing the data. This has many applications from a small scale to large scale. Whether it is the study of the population of the country or its economy, stats are used for all such data analysis.

Statistics has a huge scope in many fields such as sociology, psychology, geology, weather forecasting, etc. The data collected here for analysis could be quantitative or qualitative.

 Quantitative data are also of two types such as: discrete and continuous. Discrete data has a fixed value whereas continuous data is not a fixed data but has a range.

**Probability deals with predicting the likelihood of future events**, while statistics involves the analysis of the frequency of past events. Probability is primarily a theoretical branch of mathematics, which studies the consequences of mathematical definitions.

Task: What is difference between probability and statistics?

## 1.2  Terms Used in Probability and Statistics

There are various terms utilized in the probability and statistics concepts, Such as:

- Random Experiment
- Sample
- Random variables
- Expected Value
- Independence
- Variance
- Mean

Let us discuss these terms one by one.

### Random Experiment

An experiment whose result cannot be predicted, until it is noticed is called a random experiment. For example, when we throw a dice randomly, the result is uncertain to us. We can get any output between 1 to 6. Hence, this experiment is random.

### Sample Spacep

A sample space is the set of all possible results or outcomes of a random experiment. Suppose, if we have thrown a dice, randomly, then the sample space for this experiment will be all possible outcomes of throwing a dice, such as;

Example: Sample Space = {1, 2, 3, 4, 5, 6}

## Random Variables

The variables which denote the possible outcomes of a random experiment are called random variables. They are of two types:

1. Discrete Random Variables
2. Continuous Random Variables

Discrete random variables take only those distinct values which are countable. Whereas continuous random variables could take an infinite number of possible values.

## Independent Event

When the probability of occurrence of one event has no impact on the probability of another event, then both the events are termed as independent of each other. For example, if you flip a coin and at the same time you throw a dice, the probability of getting a 'head' is independent of the probability of getting a 6 in dice.

Task: Give examples of Discrete and continuous data values.

## Mean

Mean of a random variable is the average of the random values of the possible outcomes of a random experiment. In simple terms, it is the expectation of the possible outcomes of the random experiment, repeated again and again or n number of times. It is also called the expectation of a random variable.

## Expected Value

Expected value is the mean of a random variable. It is the assumed value which is considered for a random experiment.  It is also called expectation, mathematical expectation or first moment. For example, if we roll a dice having six faces, then the expected value will be the average value of all the possible outcomes, i.e. 3.5.

## Variance

Basically, the variance tells us how the values of the random variable are spread around the mean value. It specifies the distribution of the sample space across the mean.

Task:What is difference between dependent and independent event?

## 1.3   Elements of Set Theory

## Set - Definition

A set is an unordered collection of different elements. A set can be written explicitly by listing its elements using set bracket. If the order of the elements is changed or any element of a set is repeated, it does not make any changes in the set.

## Some Example of Sets

- A set of all positive integer
- A set of all the planets in the solar system
- A set of all the states in India
- A set of all the lowercase letters of the alphabet

## 1.4  Operations on sets

The symbol ∪ is employed to denote the union of two sets. Thus, the set $A \cup B$—read "*A* union *B*" or "the union of *A* and *B*"—is defined as the set that consists of all elements belonging to either set *A* or set *B* (or both).

For Example,

Suppose that Committee *A*, consisting of the 5 members

Jones, Blanshard, Nelson, Smith, and Hixon, meets with Committee *B*,

Consisting of the 5 members Blanshard, Morton, Hixon, Young, and Peters. Clearly, the union of Committees *A* and *B* must then consist of 8 members rather than 10—namely, Jones, Blanshard, Nelson, Smith, Morton, Hixon, Young, and Peters.

The intersection operation is denoted by the symbol ∩. The set $A \cap B$—read "*A* intersection *B*" or "the intersection of *A* and *B*"—is defined as the set composed of all elements that belong to both *A* and *B*.

Thus, the intersection of the two committees in the foregoing example is the set consisting of Blanshard and Hixon.

If *E* denotes the set of all positive even numbers and *O* denotes the set of all positive odd numbers, then their union yields the entire set of positive integers, and their intersection is the empty set. Any two sets whose intersection is the empty set are said to be disjoint.

When the admissible elements are restricted to some fixed class of objects *U*, *U* is called the universal set (or universe). Then for any subset *A* of *U*, the complement of *A* (symbolized by *A'* or $U - A$) is defined as the set of all elements in the universe *U* that are not in *A*. For example, if the universe consists of the 26 letters of the alphabet, the complement of the set of vowels is the set of consonants.

Example:

In analytic geometry, the points on a Cartesian grid are ordered pairs $(x, y)$ of numbers. In general, $(x, y) \neq (y, x)$; ordered pairs are defined so that $(a, b) = (c, d)$ if and only if both $a = c$ and $b = d$. In contrast, the set $\{x, y\}$ is identical to the set $\{y, x\}$ because they have exactly the same members.
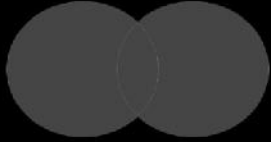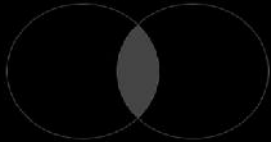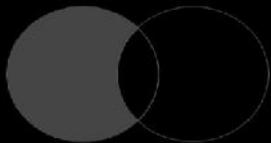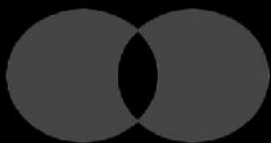
The Cartesian product of two sets *A* and *B*, denoted

By $A \times B$,

is defined as the set consisting of all ordered pairs $(a, b)$ for which $a \in A$ and $b \in B$. For example, if $A = \{x, y\}$ and $B = \{3, 6, 9\}$, then $A \times B = \{(x, 3), (x, 6), (x, 9), (y, 3), (y, 6), (y, 9)\}$.

Task: Show diagrammatically difference between union and intersection

| Set Operation | Venn Diagram | Interpretation |
|---|---|---|
| Union |  | $A \cup B$, is the set of all values that are a member of A, or B, or both. |
| Intersection |  | $A \cap B$, is the set of all values that are members of both A and B. |
| Difference |  | $A \setminus B$, is the set of all values of A that are not members of B |
| Symmetric Difference |  | $A \triangle B$, is the set of all values which are in one of the sets, but not both. |

## Set Theory Symbols

| Symbol | Name | Example | Explanation |
|---|---|---|---|
| { } | Set | $A = \{1,3\}$ <br> $B = \{2,3,9\}$ <br> $C = \{3,9\}$ | Collection of objects |
| $\cap$ | Intersect | $A \cap B = \{3\}$ | Belong to both set A and set B |
| $\cup$ | Union | $A \cup B = \{1,2,3,9\}$ | Belong to set A or set B |
| $\subset$ | Proper Subset | $\{1\} \subset A$ <br> $C \subset B$ | A set that is contained in another set |
| $\subseteq$ | Subset | $\{1\} \subseteq A$ <br> $\{1,3\} \subseteq A$ | A set that is contained in or equal to another set |
| $\not\subset$ | Not a Proper Subset | $\{1.3\} \not\subset A$ | A set that is not contained in another set |
| $\supset$ | Superset | $B \supset C$ | Set B includes set C |
| $\in$ | Is a member | $3 \in A$ | 3 is an element in set A |
| $\notin$ | Is not a member | $4 \notin A$ | 4 is not an element in set A |

*Probability and Statistics*

## 1.5 <u>What Is Conditional Probability?</u>

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

For example:

Event A is that an individual applying for college will be accepted. There is an 80% chance that this individual will be accepted to college.

Event B is that this individual will be given dormitory housing. Dormitory housing will only be provided for 60% of all of the accepted students.

P (Accepted and dormitory housing) = P (Dormitory Housing | Accepted) P (Accepted) = (0.60)*(0.80) = 0.48.

A conditional probability would look at these two events in relationship with one another, such as the probability that you are both accepted to college, *and* you are provided with dormitory housing.

Conditional probability can be contrasted with unconditional probability. Unconditional probability refers to the likelihood that an event will take place irrespective of whether any other events have taken place or any other conditions are present.

## Understanding Conditional Probability

As previously stated, conditional probabilities are contingent on a previous result. It also makes a number of assumptions. For example, suppose you are drawing three marbles—red, blue, and green—from a bag. Each marble has an equal chance of being drawn. What is the conditional probability of drawing the red marble after already drawing the blue one?

First, the probability of drawing a blue marble is about 33% because it is one possible outcome out of three. Assuming this first event occurs, there will be two marbles remaining, with each having a 50% chance of being drawn. So the chance of drawing a blue marble after already drawing a red marble would be about 16.5% (33% x 50%).

As another example to provide further insight into this concept, consider that a fair die has been rolled and you are asked to give the probability that it was a five. There are six equally likely outcomes, so your answer is 1/6. But imagine if before you answer, you get extra information that the number rolled was odd. Since there are only three odd numbers that are possible, one of which is five, you would certainly revise your estimate for the likelihood that a five was rolled from 1/6 to 1/3.

This *revised* probability that an event *A* has occurred, considering the additional information that another event *B* has definitely occurred on this trial of the experiment, is called the *conditional probability of A given B* and is denoted by P(A|B).

**Conditional Probability Formula**

*P(B|A) = P(A and B) / P(A)*

*Or:*

*P(B|A) = P(A∩B) / P(A)*

**Another Example of Conditional Probability**

As another example, suppose a student is applying for admission to a university and hopes to receive an academic scholarship. The school to which they are applying accepts 100 of every 1,000 applicants (10%) and awards academic scholarships to 10 of every 500 students who are accepted (2%). Of the scholarship recipients, 50% of them also receive university stipends for books, meals, and housing. For our ambitious student, the chance of them being accepted then receiving a scholarship is .2% (.1 x .02). The chance of them being accepted, receiving the scholarship, then also receiving a stipend for books, etc. is .1% (.1 x .02 x .5).

Task: How conditional probability is different from normal probability

**Independent Events**

Events can be "Independent", meaning each event is **not affected** by any other events.

**Example: Tossing a coin.**

Each toss of a coin is a perfect isolated thing.

What it did in the past will not affect the current toss.

The chance is simply 1-in-2, or 50%, just like ANY toss of the coin.

So each toss is an **Independent Event**.

**Dependent Events**

But events can also be "dependent" ... which means they **can be affected by previous events**.

**Example: Marbles in a Bag**

2 blue and 3 red marbles are in a bag.

What are the chances of getting a blue marble?

The chance is **2 in 5**

**But after taking one out** the chances change!

So the next time:

If we got a **red** marble before, then the chance of a blue marble next is **2 in 4**

If we got a **blue** marble before, then the chance of a blue marble next is **1 in 4**

This is because we are **removing** marbles from the bag.

So the next event **depends on** what happened in the previous event, and is called **dependent**

**Example: Drawing 2 Kings from a Deck**

**Event A** is drawing a King first, and **Event B** is drawing a King second.

For the first card the chance of drawing a King is 4 out of 52 (there are 4 Kings in a deck of 52 cards):

P(A) = 4/52

But after removing a King from the deck the probability of the 2nd card drawn is **less** likely to be a King (only 3 of the 51 cards left are Kings):

P(B|A) = 3/51

And so:

**P(A and B) = P(A) x P(B|A)** = (4/52) x (3/51) = 12/2652 = **1/221**

So the chance of getting 2 Kings is 1 in 221, or about 0.5%

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

Probability of event A given B has occured

Probability of event A occured and event B occured

Probability of event B

## 1.6  **Mutually Exclusive Events**

In probability theory, two events are said to be mutually exclusive if they cannot occur at the same time or simultaneously. In other words, mutually exclusive events are called disjoint events. If two events are considered disjoint events, then the probability of both events occurring at the same time will be zero. If A and B are the two events, then the probability of disjoint of event A and B is written by:

Probability of Disjoint (or) Mutually Exclusive Event = P ( A and B) = 0

In probability, the specific addition rule is valid when two events are mutually exclusive. It states that the probability of either event occurring is the sum of probabilities of each event occurring. If A and B are said to be mutually exclusive events then the probability of an event A occurring or the probability of event B occurring is given as P(A) + P(B), i.e.,

P (A or B) = P (A) + P (B)

Some of the examples of the mutually exclusive events are:

Example: When tossing a coin, the event of getting head and tail are mutually exclusive. Because the probability of getting head and tail simultaneously is 0.

In a six-sided die, the events "2" and "5" are mutually exclusive. We cannot get both the events 2 and 5 at the same time when we threw one die.

In a deck of 52 cards, drawing a red card and drawing a club are mutually exclusive events because all the clubs are black.

If the events A and B are not mutually exclusive, the probability of getting A or B is given as:

P (A or B) = P(A) + P(B) – P (A and B)

**Dependent and Independent Events**

Two events are said to be dependent if the occurrence of one event changes the probability of another event. Two events are said to be independent events if the probability of one event that does not affect the probability of another event. If two events are mutually exclusive, they are not independent. Also, independent events cannot be mutually exclusive.

**Mutually Exclusive Events Probability Rules**

In probability theory, two events are mutually exclusive or disjoint if they do not occur at the same time.

A clear case is the set of results of a single coin toss, which can end in either heads or tails, but not for both. While tossing the coin, both outcomes are collectively exhaustive, which suggests that at least one of the consequences must happen, so these two possibilities collectively exhaust all the possibilities.

Though, not all mutually exclusive events are commonly exhaustive. For example, the outcomes 1 and 4 of a six-sided die, when we throw it, are mutually exclusive (both 1 and 4 cannot come as result at the same time) but not collectively exhaustive (it can result to distinct outcomes such as 2,3,5,6).

From the definition of mutually exclusive events, certain rules for the probability are concluded.

Addition Rule: P (A + B) = 1

Subtraction Rule: P (A U B)′ = 0

Multiplication Rule: P (A ∩ B) = 0

There are different varieties of events also. For instance, think a coin that has a Head on both the sides of the coin or a Tail on both sides.

It doesn't matter how many times you flip it, it will always occur Head (for the first coin) and Tail (for the second coin). If we check the sample space of such experiment, it will be either { H } for the first coin and { T } for the second one. Such events have single point in the sample space and are called **"Simple Events"**. Such kind of two sample events is always mutually exclusive.

`

Task: What are mutually exclusive events?
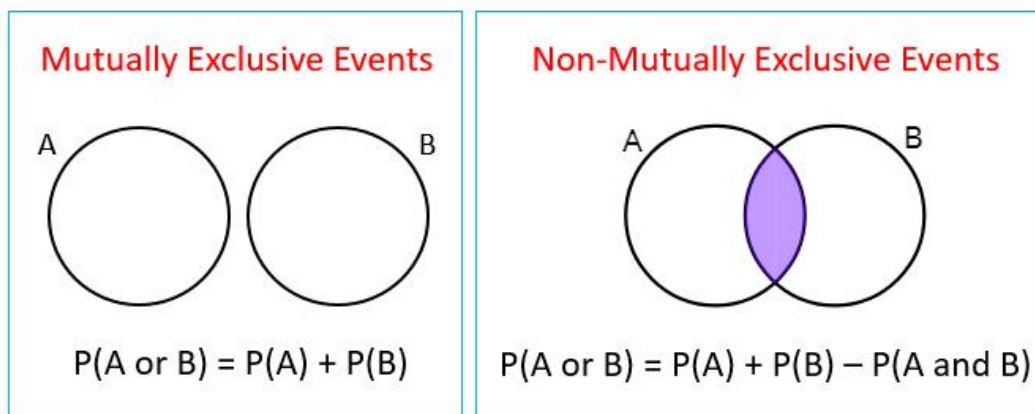


## Conditional Probability for Mutually Exclusive Events

Conditional probability is stated as the probability of an event A, given that another event B has occurred. Conditional Probability for two independent events B has given A is denoted by the expression P( B | A) and it is defined using the equation

P(B|A)= P (A∩B))/P(A)

Redefine the above equation using multiplication rule: P (A ∩ B) = 0

P(B|A)= 0/P(A)

So the conditional probability formula for mutually exclusive events is:

**P (B | A) = 0**

Examples with Solutions

Here the sample problem for mutually exclusive events is given in detail. Go through once to learn easily.

**Question 1: What is the probability of a die showing a number 3 or number 5?**

Solution: Let,

P(3) is the probability of getting a number 3

P(5) is the probability of getting a number 5

P(3) = 1/6 and P(5) = 1/6

So,

P(3 or 5) = P(3) + P(5)

P(3 or 5) = (1/6) + (1/6) = 2/6

P(3 or 5) = 1/3

Therefore, the probability of a die showing 3 or 5 is 1/3.

## 1.7  Pair wise independence

The events are called pairwise independent **if any two events in the collection are independent of each other**, while saying that the events are mutually independent (or collectively independent) intuitively means that each event is independent of any combination of other events in the collection.

- Pairwise means forming all possible pairs — two items at a time — from a set. For example, in the set {1,2,3} all possible pairs are (1,2),(2,3),(1,3).

- The events are called pairwise independent if any two events in the collection are independent of each other while saying that the events are mutually independent (or collectively independent) intuitively means that each event is independent of any combination of other events in the collection.

## 1.8  What Is Bayes' Theorem?

Bayes' theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence. In finance, Bayes' theorem can be used to rate the risk of lending money to potential borrowers.

Bayes' theorem is also called Bayes' Rule or Bayes' Law and is the foundation of the field of Bayesian statistics

Applications of the theorem are widespread and not limited to the financial realm. As an example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test.

Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities.

Prior probability, in Bayesian statistical inference, is the probability of an event before new data is collected. This is the best rational assessment of the probability of an outcome based on the current knowledge before an experiment is performed.

Posterior probability is the revised probability of an event occurring after taking into consideration new information. Posterior probability is calculated by updating the prior probability by using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

Bayes' theorem thus gives the probability of an event based on new information that is, or may be related, to that event.

The formula can also be used to see how the probability of an event occurring is affected by hypothetical new information, supposing the new information will turn out to be true.

For instance, say a single card is drawn from a complete deck of 52 cards. The probability that the card is a king is four divided by 52,

Which equals 1/13 or approximately 7.69%.

Remember that there are four kings in the deck. Now, suppose it is revealed that the selected card is a face card. The probability the selected card is a king, given it is a face card, is four divided by 12, or approximately 33.3%, as there are 12 face cards in a deck.

## Formula For Bayes' Theorem

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

**where:**

$P(A) = $ The probability of A occurring

$P(B) = $ The probability of B occurring

$P(A|B) = $ The probability of A given B

$P(B|A) = $ The probability of B given A

$P\left(A \cap B)\right) = $ The probability of both A and B occurring

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

P(Fire|Smoke) means how often there is fire when we can see smoke
P(Smoke|Fire) means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know
"backwards" P(Smoke|Fire)

*Probability and Statistics*

📃Example:

- **Dangerous fires are rare (1%)**
- **But smoke is fairly common (10%) due to barbecues,**
- **And 90% of dangerous fires make smoke**

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$P(Fire | Smoke) = \frac{P(Fire) \ P(Smoke | Fire)}{P(Smoke)}$$

$$= \frac{1\% \ x \ 90\%}{10\%}$$

$$= 9\%$$

So it is still worth checking out any smoke to be sure.

📃 Example: Picnic Day

You are planning a picnic today, but the morning is cloudy

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

**What is the chance of rain during the day?**

We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.

The chance of Rain given Cloud is written P(Rain | Cloud)

So let's put that in the formula:

$$P(Rain | Cloud) = \frac{P(Rain) \ P(Cloud | Rain)}{P(Cloud)}$$

- P(Rain) is Probability of Rain = 10%
- P(Cloud | Rain) is Probability of Cloud, given that Rain happens = 50%
- P(Cloud) is Probability of Cloud = 40%

$$P(Rain | Cloud) = \frac{0.1 \ x \ 0.5}{0.4} \ = .125$$

Or a 12.5% chance of rain. Not too bad, let's have a picnic!

## 1.9   How to Use Bayes Theorem for Business and Finance

In finance and business circles, corporate financial specialists have been applying Bayes' Theorem for centuries.

Consider these applications:

In evaluating interest rates. Companies rely on interest rates for multiple reasons - borrowing money, investing in the fixed income market, and trading in currencies overseas. Any unexpected shifts in interest rate values can hit a company hard in the pocketbook, and can negatively impact profits and revenues. With Bayes Theorem and estimated probabilities, companies can better evaluate systematic changes in interest rates, and steer their financial resources to take maximum advantage.

With net income. Businesses are keen to be on top of their net income streams, or the profit a business earns after subtracting expenses out of the equation. Net income is highly vulnerable to external events, like legal proceedings, weather, the cost of necessary equipment and materials, and geopolitical events, for starters. Plugging probability scenarios into the net income equation when these scenarios arise gives financial decision makers a stronger platform when managing resources and making critical decisions.

For extending credit. Under the Bayes Theorem conditional probability model, financial companies can make better decisions and better evaluate the risk of lending cash to unfamiliar or even existing borrowers. For example, an existing client may have had a good previous track record of repaying loans, but lately the client has been slow in playing This additional information, based on probability theory, can lead the company to treat the slow payment history as a red flag, and either hike interest rates on the loan or reject it altogether.

Example:

**The Enigma code**
In 1774, the brilliant French mathematician Pierre-Simon Laplace expanded upon Bayes' theorem, before the theorem all but disappeared from sight until the 20th Century, when British codebreaker Alan Turing used it during the Second World War to help crack the 'unbreakable' Enigma code, a development that helped the Allies win the war.

Turing developed a system based on Bayesian theory that enabled him to guess a stretch of letters in an Enigma message, calculate the probabilities, and add more clues as they arrived. With this method he could reduce the number of wheel settings to be tested, which subsequently led him to cracking the code.

With the advent of the computer age, the use of Bayesian theory has exploded, into such areas as artificial intelligence, robotics, law, imaging technologies and medical diagnostics. In 1996, Bill Gates said that Microsoft's competitive advantage was its use of Bayesian networks. Bayes techniques are also used in spam filters, voice recognition systems, recommendation systems and in Google search.

Despite Bayes' theorem being a clever mathematical formula, the good news is that you don't need to be a mathematician to be able to apply Bayesian thinking to investing or your everyday life.

## Summary

**Probability and Statistics** are the two important concepts in Math's. Probability is all about chance.

Whereas statistics is more about how we handle various data using different techniques.

It helps to represent complicated data in a very easy and understandable way.

The statistic has a huge application nowadays in data science professions.

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

In probability theory, two events are mutually exclusive or disjoint if they do not occur at the same time.

A set is an unordered collection of different elements. A set can be written explicitly by listing its elements using set bracket. If the order of the elements is changed or any element of a set is repeated, it does not make any changes in the set

Random Experiment:An experiment whose result cannot be predicted, until it is noticed is called a random experiment. For example, when we throw a dice randomly, the result is uncertain to us. We can get any output between 1 to 6. Hence, this experiment is random.

Sample Space: A sample space is the set of all possible results or outcomes of a random experiment.

## **Keywords**

Expected value is the mean of a random variable. It is the assumed value which is considered for a random experiment Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

In probability theory, two events are mutually exclusive or disjoint if they do not occur at the same time.

A set is an unordered collection of different elements. A set can be written explicitly by listing its elements using set bracket. If the order of the elements is changed or any element of a set is repeated, it does not make any changes in the set

Bayes' theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability

## **Self Assessment**

1. _____ is more about how we handle various data using different techniques.

A. Statistics

B. Probability

C. Random Experiemnt

D. Sample Space

2. _____is all about chance.

A. Statistics

B. Probability

C. Random Experiemnt

D. Sample Space

3. A _____ is a trial, or observation that can be repeated numerous times under the same

conditions.

A. Statistics

B. Probability

C. Random Experiment

D. Sample Space

4. The _____ of a random experiment is the collection of all possible outcomes.

A. Statistics

B. Probability

C. Random Experiment

D. Sample Space

5. _____are two or more sets that have no elements in common, therefore the intersection is an empty set.

A. Disjoint sets

B. Union

C. Set difference

D. Intersection

6. In mathematics, a set A is a _____of a set B if all elements of A are also elements of B.

A. Disjoint sets

B. Subset

C. Set difference

D. Intersection

7. Tossing a coin is

A. Dependent event

B. Independent event

C. Null

D. All of these

8. _____is the probability of an event occurring given that another event has already occurred.

A. Conditional probability

B. Unconditional probability

C. Random probability

D. All of these

9. Suppose we have 5 blue marbles and 5 red marbles in a bag.  We pull out one marble, which may be blue or red.  Now there are 9 marbles left in the bag. This is example of

A. Dependent event

B. Independent event

C. Null

D. All of these

10. Events are said to be _____if they cannot occur together.

A. Mutually exclusive

B. Exclusive

C. Mutually

D. All of these

11. The result of an experiment is known as _____.

A. Random variable

B. Event

C. Sample space

D. All of these

12. _____are nothing but all the  sample points

A. Exhaustive events

B. Mutually exclusive

C. Exclusive

D. Mutually

13. _____is the measure of the likelihood that an event will occur.

A. Probability

B. Statistics

C. Sample space

D. Random Experiment

14. The _____Theorem is a mathematic model, based on statistics and probability that aims to

calculate the probability of one scenario based on its relationship with another scenario.

A. Multiplication

B. Addition

C. Bayes

D. Random theorem

15. The initial probability is based on the present level of information.

A. Prior Probability

B. Posterior Probability

C. Previous Probability

D. All of these

## Answers for Self Assessment

| 1. | A | 2. | B | 3. | C | 4. | D | 5. | A |
|----|---|----|---|----|---|----|---|----|---|
| 6. | B | 7. | B | 8. | A | 9. | A | 10. | A |
| 11. | B | 12. | A | 13. | A | 14. | C | 15. | A |

　　**Lovely Professional University**

## Review Questions

Q1. What is the probability of getting a 2 or a 5 when a die is rolled?

Q2.What is difference between probability and statistics?

Q3.Explain conditional probability with example?

Q4.  How Probability and statistics are related to set theory of mathematics?

Q5.  Why, mutually exclusive events are called disjoint events.

Q6. What is Bayes theorem and How to Use Bayes Theorem for Business and Finance.

Q7.Give example to differentiate independent and dependent events?

Q8 what is random experiment and random variables.

## Further Readings

An Introduction to Probability and Statistics

Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

First Course in Probability, A

Book by Sheldon M. Ross

Schaums Theory and Problems of Statistics

Book by Murray R. Spiegel

Introduction to Probability, Statistics, and Random ...

Book by HosseinPishro-Nik

## Web Links

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability

*Dr. Pritpal Singh, Lovely Professional University*

# Unit 02: Introduction to Statistics and Data Analysis

## Objectives

- Understand basic definitions of statistical inference.
- Understand various sampling techniques.
- Learn the concept of experimental design.
- Understand concept of sampling techniques.
- learn concept of sample and population

## Introduction:

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting, and presenting empirical data. Statistics is a highly interdisciplinary field; research in statistics finds applicability in virtually all scientific fields and research questions in the various scientific fields motivate the development of new statistical methods and theory. In developing methods and studying the theory that underlies the methods statisticians draw on a variety of mathematical and computational tools. Two fundamental ideas in the field of statistics are uncertainty and variation. There are many situations that we encounter in science (or more generally in life) in which the outcome is uncertain. In some cases the uncertainty is because the outcome in question is not determined yet (e.g., we may not know whether it will rain tomorrow) while in other cases the uncertainty is because although the outcome has been determined already we are not aware of it (e.g., we may not know whether we passed a particular exam).

Probability is a mathematical language used to discuss uncertain events and probability plays a key role in statistics. Any measurement or data collection effort is subject to a number of sources of variation. By this we mean that if the same measurement were repeated, then the answer would likely change. Statisticians attempt to understand and control (where possible) the sources of variation in any situation.

> Two fundamental ideas in the field of statistics are uncertainty and variation

## 2.1   Statistical inference

**Statistical inference** is the process of using data analysis to infer properties of an underlying distribution of probability. [Inferential statistical analysis infers properties of a population, for example by **testing hypotheses** and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population. In machine learning, the term *inference* is sometimes used instead to mean "make a prediction, by evaluating an already trained model"; in this context inferring properties of the model is referred to as *training* or *learning* (rather than *inference*), and using a model for prediction is referred to as *inference* (instead of *prediction*); see also predictive inference.

**Models and assumptions**

Any statistical inference requires some assumptions. A **statistical model** is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to draw inference. Descriptive statistics are typically used as a preliminary step before more formal inferences are drawn.

**Degree of models/assumptions**

Statisticians distinguish between three levels of modeling assumptions;

**Fully parametric**: The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters.

For example, one may assume that the distribution of population values is truly Normal, with unknown mean and variance, and that datasets are generated by 'simple' random sampling. The family of generalized linear models is a widely used and flexible class of parametric models.

**Non-parametric**: The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal.

For example, every continuous probability distribution has a median, which may be estimated using the sample median or the Hodges–Lehmann–Sen estimator, which has good properties when the data arise from simple random sampling.

**Semi-parametric**: This term typically implies assumptions 'in between' fully and non-parametric approaches. For example, one may assume that a population distribution has a finite mean. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but not make any parametric assumption describing the variance around that mean (i.e. about the presence or possible form of any heteroscedasticity). More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components. One component is treated parametrically and the other non-parametrically. The well-known Cox model is a set of semi-parametric assumptions.

Task: How statistical inference is used in analysis?

## 2.2   Population and Sample

In statistics as well as in quantitative methodology, the set of data are collected and selected from a statistical population with the help of some defined procedures. There are two different types of data sets namely, **population and sample**. So basically when we calculate the mean deviation, variance and standard deviation, it is necessary for us to know if we are referring to the

entire population or to only sample data. Suppose the size of the population is denoted by 'n' then the sample size of that population is denoted by n -1. Let us take a look of population, data sets and sample data sets in detail.

## Population

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a **parameter**. For example, all people living in India indicates the population of India.

There are different types of population. They are:

- Finite Population
- Infinite Population
- Existent Population
- Hypothetical Population

Let us discuss all the types one by one.

### Finite Population

The finite population is also known as a countable population in which the population can be counted. In other words, it is defined as the population of all the individuals or objects that are finite. For statistical analysis, the finite population is more advantageous than the infinite population. Examples of finite populations are employees of a company, potential consumer in a market.

### Infinite Population

The infinite population is also known as an uncountable population in which the counting of units in the population is not possible. Example of an infinite population is the number of germs in the patient's body is uncountable.

### Existent Population

The existing population is defined as the population of concrete individuals. In other words, the population whose unit is available in solid form is known as existent population. Examples are books, students etc.

### Hypothetical Population

The population in which whose unit is not available in solid form is known as the hypothetical population. A population consists of sets of observations, objects etc that are all something in common. In some situations, the populations are only hypothetical.

Examples are an outcome of rolling the dice, the outcome of tossing a coin

## Sample

It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic. Sampling is the process of selecting the sample from the population. For example, some people living in India is the sample of the population.

Basically, there are two types of sampling. They are:

- Probability sampling
- Non-probability sampling
-

### Probability Sampling

In probability sampling, the population units cannot be selected at the discretion of the researcher. This can be dealt with following certain procedures which will ensure that every unit of the population consists of one fixed probability being included in the sample. Such a method is also called random sampling. Some of the techniques used for probability sampling are:

- Simple random sampling

- Cluster sampling
- Stratified Sampling
- Disproportionate sampling
- Proportionate sampling
- Optimum allocation stratified sampling
- Multi-stage sampling

**Non Probability Sampling**

In non-probability sampling, the population units can be selected at the discretion of the researcher. Those samples will use the human judgments for selecting units and has no theoretical basis for estimating the characteristics of the population. Some of the techniques used for non-probability sampling are

- Quota sampling
- Judgment sampling
- Purposive sampling

Task: What is different between probability and non-probability sampling

**Population and Sample Examples**

All the people who have the ID proofs is the population and a group of people who only have voter id with them is the sample.

All the students in the class are population whereas the top 10 students in the class are the sample.

All the members of the parliament is population and the female candidates present there is the sample.

**Simple random sampling:** One of the best probability sampling techniques that helps in saving time and resources, is the Simple Random Sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.

For example, in an organization of 500 employees, if the HR team decides on conducting team building activities, it is highly likely that they would prefer picking chits out of a bowl. In this case, each of the 500 employees has an equal opportunity of being selected.

**Cluster sampling:** Cluster samplingis a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback. For example, if the United States government wishes to evaluate the number of immigrants living in the Mainland US, they can divide it into clusters based on states such as California, Texas, Florida, Massachusetts, Colorado, Hawaii, etc. This way of conducting a survey will be more effective as the results will be organized into states and provide insightful immigration data.

**Systematic sampling:** Researchers use the systematic sampling methodto choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.

For example, a researcher intends to collect a systematic sample of 500 people in a population of 5000. He/she numbers each element of the population from 1-5000 and will choose every 10th individual to be a part of the sample (Total population/ Sample Size = 5000/500 = 10).

**Stratified random sampling:** Stratified random samplingis a method in which the researcher divides the population into smaller groups that don't overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.

For example, a researcher looking to analyze the characteristics of people belonging to different annual income divisions will create strata (groups) according to the annual family income. Eg – less than $20,000, $21,000 – $30,000, $31,000 to $40,000, $41,000 to $50,000, etc. By doing this, the researcher concludes the characteristics of people belonging to different income groups. Marketers can analyze which income groups to target and which ones to eliminate to create a roadmap that would bear fruitful results.

**Convenience sampling:** This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as convenience sampling, because of the researcher's ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it's purely done based on proximity and not representativeness.

This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

For example, startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the mall entrance and giving out pamphlets randomly.

**Judgmental or purposive sampling:** Judgmental or purposive samplesare formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master's degree. The selection criteria will be: "Are you interested in doing your masters in …?" and those who respond with a "No" are excluded from the sample.

**Snowball sampling:** Snowball samplingis a sampling method that researchers apply when the subjects are difficult to trace. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.

**Quota sampling:** In Quota sampling the selection of members in this sampling technique happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.

## 2.3    Difference between Population and Sample

Some of the key differences between population and sample are clearly given below:

**Lovely Professional University**

| Comparison | Population | Sample |
|---|---|---|
| Meaning | Collection of all the units or elements that possess common characteristics | A subgroup of the members of the population |
| Includes | Each and every element of a group | Only includes a handful of units of population |
| Characteristics | Parameter | Statistic |
| Data Collection | Complete enumeration or census | Sampling or sample survey |
| Focus on | Identification of the characteristics | Making inferences about the population |

Task: In which situation snowball sampling is used?

## 2.4   Measures of  Locations

The three most common measures of location are **the mean, the median, and the mode.**

In statistics, the central tendency is the descriptive summary of a data set. Through the single value from the dataset, it reflects the centre of the data distribution. Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset. Generally, the central tendency of a dataset can be defined using some of the measures in statistics.

**Definition**

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.

**Measures of Central Tendency**

The central tendency of the dataset can be found out using the three important measures namely mean, median and mode.

## CENTRAL TENDENCY

Central Tendency

Mean — Median — Mode

**Mean**

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean. Some other measures of mean used to find the central tendency are as follows:

Geometric Mean

Harmonic Mean

Weighted Mean

It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs. Calculating the mean value is completely easy.

The histogram given below shows that the mean value of symmetric continuous data and the skewed continuous data.

### Histogram of skewed continuous

Mean 36624



### Histogram of symmetric continuous

Mean 100.67
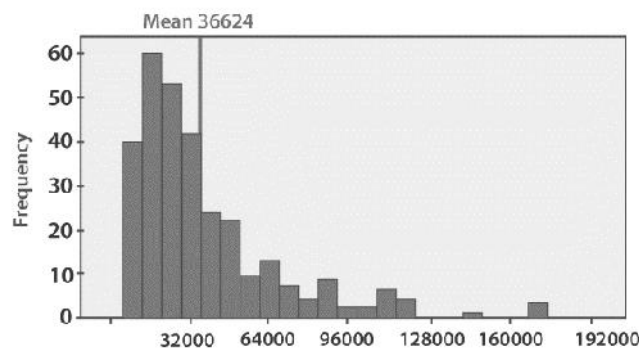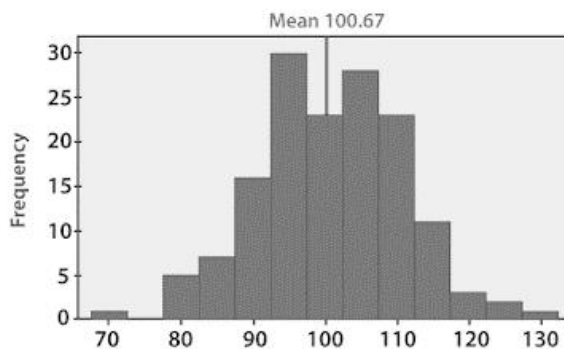
*Probability and Statistics*

In symmetric data distribution, the mean value is located accurately at the centre. But in the skewed continuous data distribution, the extreme values in the extended tail pull the mean value away from the centre. So it is recommended that the mean can be used for the symmetric distributions.

**Median**

Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.

Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2

| Median odd |
|------------|
| 23 |
| 21 |
| 18 |
| 16 |
| 15 |
| 13 |
| 12 |
| 10 |
| 9 |
| 7 |
| 6 |
| 5 |
| 2 |

Here 12 is the middle or median number that has 6 values above it and 6 values below it.

Now, consider another example with an even number of observations that are arranged in descending order – 40, 38, 35, 33, 32, 30, 29, 27, 26, 24, 23, 22, 19, and 17

| Median even |
|-------------|
| 40 |
| 38 |
| 35 |
| 33 |
| 32 |
| 30 |
| 29 |
| 27 |
| 26 |
| 24 |
| 23 |
| 22 |
| 19 |
| 17 |

28

When you look at the given dataset, the two middle values obtained are 27 and 29.

Now, find out the mean value for these two numbers.

i.e.,(27+29)/2 =28

Therefore, the median for the given data distribution is 28.

**Mode**

The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5

| Mode |
|------|
| 5 |
| 5 |
| 5 |
| 4 |
| 4 |
| 3 |
| 2 |
| 2 |
| 1 |

Since the mode represents the most common value. Hence, the most frequently repeated value in the given dataset is 5.

Based on the properties of the data, the measures of central tendency are selected.

If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset.

If you have skewed distribution, the best measure of finding the central tendency is the median.

If you have the original data, then both the median and mode are the best choice of measuring the central tendency.

If you have categorical data, the mode is the best choice to find the central tendency.

> Task: In what situation median is more effective than mean?

## 2.5   Measures of variability

**Variance**

According to layman's words, the variance is a measure of how far a set of data are dispersed out from their mean or average value. It is denoted as '$\sigma^2$'.

**Properties of Variance**

It is always non-negative since each term in the variance sum is squared and therefore the result is either positive or zero.

Variance always has squared units. For example, the variance of a set of weights estimated in kilograms will be given in kg squared. Since the population variance is squared, we cannot compare it directly with the mean or the data themselves.

**Standard Deviation**

The spread of statistical data is measured by the standard deviation. Distribution measures the deviation of data from its mean or average position. The degree of dispersion is computed by the method of estimating the deviation of data points. It is denoted by the symbol, '$\sigma$'.

**Properties of Standard Deviation**

It describes the square root of the mean of the squares of all values in a data set and is also called the root-mean-square deviation.

The smallest value of the standard deviation is 0 since it cannot be negative.

When the data values of a group are similar, then the standard deviation will be very low or close to zero. But when the data values vary with each other, then the standard variation is high or far from zero.

**Variance and Standard Deviation Formula**

As discussed, the variance of the data set is the average square distance between the mean value and each data value. And standard deviation defines the spread of data values around the mean.

| | Population | Sample |
|---|---|---|
| **Variance** | $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ | $S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ |
| **Standard deviation** | $\sigma = \sqrt{\dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$ | $S = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ |

The formulas for the variance and the standard deviation for both population and sample data set are given below:

**Variance Formula:**

$\sigma^2$ = Population variance

N = Number of observations in population

$X_i$ = ith observation in the population

$\mu$ = Population mean

The sample variance formula is

$s^2$ = Sample variance

n = Number of observations in sample

$x_i$ = ith observation in the sample

$\bar{x}$ = Sample mean

**Standard Deviation Formula**

The population standard deviation formula is g

$\sigma$ = Population standard deviation

Similarly, the sample standard deviation formula is:

Here,

s = Sample standard deviation

**Variance and Standard deviation Relationship**

Variance is equal to the average squared deviations from the mean, while standard deviation is the number's square root. Also, the standard deviation is a square root of variance. Both measures exhibit variability in distribution, but their units vary: Standard deviation is expressed in the same units as the original values, whereas the variance is expressed in squared units.

**Example**

Question: If a die is rolled, then find the variance and standard deviation of the possibilities.

**Solution:** When a die is rolled, the possible outcome will be 6. So the sample space, n = 6 and the data set = { 1;2;3;4;5;6}.

To find the variance, first, we need to calculate the mean of the data set.

Mean, $\bar{x}$ = (1+2+3+4+5+6)/6 = 3.5

We can put the value of data and mean in the formula to get;

$\sigma^2 = \Sigma\ (x_i - \bar{x})^2/n$

$\sigma^2 = \frac{1}{6}\ (6.25+2.25+0.25+0.25+2.25+6.25)$

$\sigma^2 = 2.917$

Now, the standard deviation = $\sqrt{2.917}$ = 1.708

## 2.6   Discrete and continuous data

Discrete data is information that can only take certain values. These values don't have to be whole numbers (a child might have a shoe size of 3.5 or a company may make a profit of £3456.25 for example) but they are fixed values – a child cannot have a shoe size of 3.72!

The number of each type of treatment a salon needs to schedule for the week, the number of children attending a nursery each day or the profit a business makes each month are all examples of discrete data. This type of data is often represented using tally charts, bar charts or pie charts.

Continuous data is data that can take any value. Height, weight, temperature and length are all examples of continuous data. Some continuous data will change over time; the weight of a baby in its first year or the temperature in a room throughout the day. This data is best shown on a line graph as this type of graph can show how the data changes over a given period of time. Other continuous data, such as the heights of a group of children on one particular day, is often grouped into categories to make it easier to interpret.

| Examples | |
|---|---|
| **Discrete** | **Continuous** |
| • # of eggs in a basket<br>• # of kids in a class<br>• # of Facebook likes<br>• # of diaper changes in a day<br>• # of wins in a season<br>• # of votes in an election | • Weight difference to 8 decimals before and after cookie binge.<br>• Wind speed<br>• Water temperature<br>• Volts of electricity |

## 2.7   What is Statistical Modeling?

Statistical modeling refers to the data science process of applying statistical analysis to datasets. A statistical model is a mathematical relationship between one or more random variables and other non-random variables. The application of statistical modeling to raw data helps data scientists approach data analysis in a strategic manner, providing intuitive visualizations that aid in identifying relationships between variables and making predictions.

Common data sets for statistical analysis include Internet of Things (IoT) sensors, census data, public health data, social media data, imagery data, and other public sector data that benefit from real-world predictions.

**Statistical Modeling Techniques**

The first step in developing a statistical model is gathering data, which may be sourced from spreadsheets, databases, data lakes, or the cloud. The most common statistical modeling methods for analyzing this data are categorized as either supervised learning or unsupervised learning. Some popular statistical model examples include logistic regression, time-series, clustering, and decision trees.

Supervised learning techniques include regression models and classification models:

Regression model: a type of predictive statistical model that analyzes the relationship between a dependent and an independent variable. Common regression models include logistic, polynomial, and linear regression models. Use cases include forecasting, time series modeling, and discovering the causal effect relationship between variables.

Classification model: a type of machine learning in which an algorithm analyzes an existing, large and complex set of known data points as a means of understanding and then appropriately classifying the data; common models include models include decision trees, Naive Bayes, nearest neighbor, random forests, and neural networking models, which are typically used in Artificial Intelligence.

Unsupervised learning techniques include clustering algorithms and association rules:

K-means clustering: aggregates a specified number of data points into a specific number of groupings based on certain similarities.

Reinforcement learning: an area of deep learning that concerns models iterating over many attempts, rewarding moves that produce favorable outcomes and penalizing steps that produce undesired outcomes, therefore training the algorithm to learn the optimal process.

There are three main types of statistical models: parametric, nonparametric, and semi parametric:

Parametric: a family of probability distributions that has a finite number of parameters.

Nonparametric: models in which the number and nature of the parameters are flexible and not fixed in advance.

Semi parametric: the parameter has both a finite-dimensional component (parametric) and an infinite-dimensional component (nonparametric).

## 2.8   Experimental Design Definition

In Statistics, the experimental design or the design of experiment (DOE) is defined as the design of an information-gathering experiment in which a variation is present or not, and it should be performed under the full control of the researcher. This term is generally used for controlled experiments. These experiments minimise the effects of the variable to increase the reliability of results. In this design, the process of an experimental unit may include a group of people, plants, animals, etc.

**Types of Experimental Designs**

There are different types of experimental designs of research. They are:

Pre-experimental Research Design

True-experimental Research Design

Quasi-Experimental Research Design

In this article, we are going to discuss these different experimental designs for research with examples.

**Pre-experimental Research Design**

The simplest form of experimental research design in Statistics is the pre-experimental research design. In this method, a group or various groups are kept under observations, after some factors are recognised for the cause and effect. This method is usually conducted in order to understand whether further investigations are needed for the targeted group. That is why this process is considered to be cost-effective. This method is classified into three types, namely,

Static Group Comparison

One-group Pretest-posttest Experimental Research Design

One-shot Case Study Experimental Research Design

**True-experimental Research Design**

This is the most accurate form of experimental research design as it relies on the statistical hypothesis to prove or disprove the hypothesis. This is the most commonly used method implemented in Physical Science. True experimental research design is the only method that establishes the cause and effect relationship within the groups. The factors which need to be satisfied in this method are:

Random variable

Variable can be manipulated by the researcher

Control Groups (A group of participants are familiar to the experimental group, but the experimental rules do not apply to them)

Experimental Group (Research participants where experimental rules are applied)

**Quasi-Experimental Design**

A quasi-experimental design is similar to the true experimental design, but there is a difference between the two.

In true experiment design, the participants of the group are randomly assigned. So, every unit has an equal chance of getting into the experimental group.

In a quasi-experimental design, the participants of the groups are not randomly assigned. So, the researcher cannot make a cause or effect conclusion. Thus, it is not possible to assign the participants into the group.

Apart from these types of experimental design research in statistics, there are other two methods used in the research process such as randomized block design and completely randomized design.

**Randomized Block Design**

The randomised block design is preferred in the case when the researcher is clear about the distinct difference among the group of objects. In this design, the experimental units are classified into subgroups of similar categories. Those groups are randomly assigned to the group of treatment. The blocks are classified in such a way in which the variability within each block should be less than the variability among the blocks. This block design is quite efficient as it reduces the variability and produces a better estimation

## 2.9    Importance of Graphs & Charts

Walk into almost any business meeting and you'll see one of these talked about at some point. What is it? It's either a graph or a chart describing something about the business. It could be a chart showing the progress the team is making on a big project. Or it could be a graph showing the sales of the business and comparing it with the sales of the competition. Either way, these graphs and charts make the information much easier to digest and understand.

Example: A **graph** or a **chart** may be defined as a visual presentation of data. For example, a utility company uses a column chart to help its customers see just how much energy they've used during the last billing cycle. A bakery may use a pie chart to show how many breads it sells when compared to its other products, such as cheesecakes and apple pies.
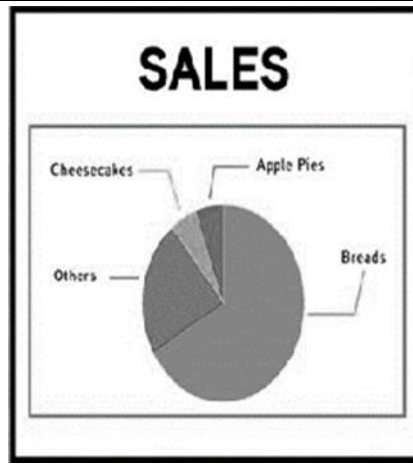
## Graphs & Charts in Business

Businesses have many uses for graphs and charts. There are many types of graphs and charts, making it easy for a business to choose the one that fits their needs the most. Let's take a look at some choices of graphs and charts available to businesses.

The first that we will look into is called a **column chart**. This type of chart has vertical columns. The height of each column, for example, tells you how much the corresponding item is. If, instead of columns, the chart has horizontal bars, that is called a bar chart. Businesses can use column or bar charts to compare products or to show how much is used each day. This type of chart lends itself well as a comparison tool, as it's easy to visually see which item's column or bar is taller or longer. This chart, for example, shows the number of ice cream sales this past week.
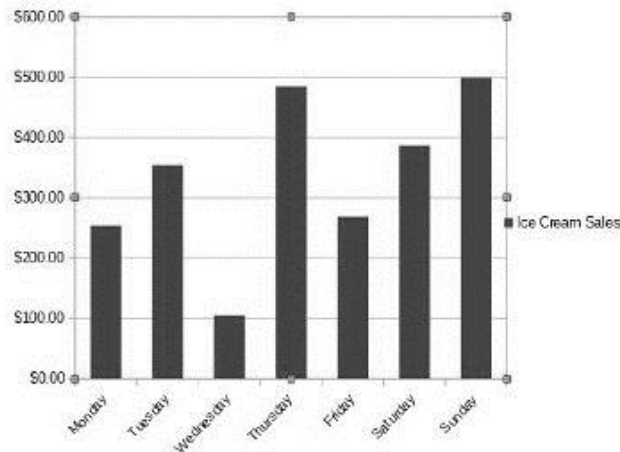


*A column chart*

You can easily see that Sunday had the largest number of ice cream sales, while Wednesday had the least.

**Line graphs** are those graphs that show your data as a line. Each successive data point is connected to the previous. This type of graph is best suited for data that is continuous in nature, such as showing the operating temperature of a computer chip over time. For example, a computer company can use a line graph to show the temperature of its processor as a person is using it for both easy word processing tasks and then some intensive gaming tasks. You'll see the line curve up and then down.

**Pie charts** are those charts that look like pies, hence the name. Each pie is split into slices with each slice representing one particular group of data. The size of the slice shows you how much of that group of data you have. Businesses can use pie charts to show their market size, such as this one showing how much of the market a particular shoe company has

## Summary

- **Statistical inference** is the process of using data analysis to infer properties of an underlying distribution of probability.
- **Sampling** is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population
- A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from. **The size of the sample is always less than the total size of the population.**
- Experimental design is the **process of carrying out research in an objective and controlled fashion** so that precision is maximized and specific conclusions can be drawn regarding a hypothesis statement.
- A discrete variable is a variable whose **value** is obtained by counting. A continuous variable is a variable whose value is obtained by measuring. ... A continuous random variable X takes all values in a given interval of numbers.

## Keywords

- **Sampling** is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population
- A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from. **The size of the sample is always less than the total size of the population**
- In the most basic form of probability sampling (i.e., a simple random sample), every member of the population has an equal chance of being selected into the study.
- . Non-probability sampling, on the other hand, **does not involve "random" processes for selecting participants**

## Self Assessment

1. _____is the process of using data analysis to infer properties of an underlying distribution of probability.
A. Statistical inference
B. Conditional probability
C. Sampling
D. Quota

2. In this case each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected.
A. Simple Random Sampling
B. Systematic Sampling
C. Stratified sampling
D. Clustered Sampling

3. In which Individuals are selected at regular intervals from the sampling frame
A. Simple Random Sampling
B. Systematic Sampling

C. Stratified sampling

D. Clustered Sampling

4. In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic

A. Simple Random Sampling

B. Systematic Sampling

C. Stratified sampling

D. Clustered Sampling

5. In a_____, subgroups of the population are used as the sampling unit, rather than individuals.

A. Simple Random Sampling

B. Systematic Sampling

C. Stratified sampling

D. Clustered Sampling

6. _____is perhaps the easiest method of sampling, because participants are selected based on availability and willingness to take part.

A. Convenience sampling

B. Quota sampling

C. Judgment Sampling

D. Snowball sampling

7. _____can be effective when a sampling frame is difficult to identify

A. Convenience sampling

B. Quota sampling

C. Judgment Sampling

D. Snowball sampling

8. _____has the advantage of being time-and cost-effective to perform whilst resulting in a range of responses

A. Convenience sampling

B. Quota sampling

C. Judgment Sampling

D. Snowball sampling

9. This method of sampling is often used by market researchers For example Interviewers are given a lot of subjects of a specified type to attempt to recruit.

A. Convenience sampling

B. Quota sampling

C. Judgment Sampling

D. Snowball sampling

10. The "_____" is the "middle" value in the list of numbers.
    A. Median
    B. Mode
    C. Mean
    D. Range

11. The_____is the sum of the data values divided by the number of data items.
    A. Median
    B. Mode
    C. Mean
    D. Range

12. Value of the random sample that occurs at the greatest frequency
    A. Median
    B. Mode
    C. Mean
    D. Range

13. _____also known as the categorical data.
    A. Qualitative data
    B. Quantative data
    C. Discrete data
    D. All of these

14. _____is also known as numerical data which represents the numerical value.
    A. Qualitative data
    B. Quantative data
    C. Discrete data
    D. All of these

15. _____is information that can only take certain values.
    A. Qualitative data
    B. Quantative data
    C. Discrete data
    D. All of these

## Answers for Self Assessment

| 1. | A | 2. | A | 3. | B | 4. | C | 5. | D |
|----|---|----|---|----|---|----|---|----|---|
| 6. | A | 7. | D | 8. | C | 9. | B | 10. | A |
| 11. | C | 12. | B | 13. | A | 14. | B | 15. | C |

## Review Questions

1. Why probability sampling method is any method of sampling that utilizes some form of random selection?
2. Explain this statement in detail "Non-probability sampling is defined as a sampling technique in which the researcher selects samples based on the subjective judgment of the researcher rather than random selection".
3. How Statistical inference is used in using data analysis?
4. What are different measures of location explain with example of each?
5. What are different measures of variability explain with example of each?
6. Whyit is said that Experimental design is the process of carrying out research in an objective and controlled fashion?
7. How do you know if data is discrete or continuous?
8. Give three examples of discrete data and continuous data?
9. How do you determine sample and population?

## Further Readings

- An Introduction to Probability and Statistics
- Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi
- First Course in Probability, A Book by Sheldon M. Ross
- Schaums Theory and Problems of Statistics
- Book by Murray R. Spiegel
- Introduction to Probability, Statistics, and Random ...
- Book by HosseinPishro-Nik

## Web Links

- https://www.tutorialspoint.com
- www.webopedia.com
- https://www.britannica.com/science/probability

# Unit 03: Mathematical Expectations

## Objectives

- Understand basics of Mathematical Expectation.
- Learn concepts of Dispersion.
- Understand Concept of skewness and Kurtosis.
- Understand concept of Expected Values.
- Solve basic questions related to probability

## Introduction

Probability is used to denote the happening of a certain event, and the occurrence of that event, based on past experiences. The mathematical expectation is the events which are either impossible or a certain event in the experiment. Probability of an impossible event is zero, which is possible only if the numerator is 0. Probability of a certain event is 1 which is possible only if the numerator and denominator are equal.

Expected value

In statistics and probability analysis, the expected value is **calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all of those values**. By calculating expected values, investors can choose the scenario most likely to give the desired outcome.

Mathematical expectation, also known as the expected value, which is the summation of all possible values from a random variable.

Expected value is **the average value of a random variable over a large number of experiments**.

Examples: If our random variable were the number obtained by rolling a fair 3-sided die,

The expected value would be (1 * 1/3) + (2 * 1/3) + (3 * 1/3) = 2

## 3.1 Mathematical Expectation

Mathematical expectation, also known as the expected value, which is the summation of all possible values from a random variable. It is also known as the product of the probability of an event occurring, denoted by P(x), and the value corresponding with the actually observed occurrence of the event.

For a random variable expected value is a useful property. E(X) is the expected value and can be computed by the summation of the overall distinct values that is the random variable. The mathematical expectation is denoted by the formula:

E(X)= Σ (x1p1, x2p2, …, xnpn),

Where, x is a random variable with the probability function, f(x),

P is the probability of the occurrence,

And n is the number of all possible values.

The mathematical expectation of an indicator variable can be 0 if there is no occurrence of an event A, and the mathematical expectation of an indicator variable can be 1 if there is an occurrence of an event A.

For example, a dice is thrown, the set of possible outcomes is { 1,2,3,4,5,6} and each of this outcome has the same probability 1/6. Thus, the expected value of the experiment will be 1/6*(1+2+3+4+5+6) = 21/6 = 3.5. It is important to know that "expected value" is not the same as "most probable value" and, it is not necessary that it will be one of the probable values.

**Properties of Expectation**

First property: If X and Y are the two variables, then the mathematical expectation of the sum of the two variables is equal to the sum of the mathematical expectation of X and the mathematical expectation of Y.

Or

E(X+Y)=E(X)+E(Y)

The first property is that of the additional theorem. This property states that if there is an X and Y, then the sum of those two random variables are equal to the sum of the mathematical expectation of the individual random variables.

Second property: The mathematical expectation of the product of the two random variables will be the product of the mathematical expectation of those two variables, but the condition is that the two variables are independent in nature. In other words, the mathematical expectation of the product of the n number of independent random variables is equal to the product of the mathematical expectation of the n independent random variables

Or

E(XY)=E(X)E(Y)

This property of the mathematical expectation states that if there is an X and Y, then the product of those two random variables are equal to the product of the mathematical expectation of the individual random variables.

Third property: The mathematical expectation of the sum of a constant and the function of a random variable is equal to the sum of the constant and the mathematical expectation of the function of that random variable.

Or,

Examples $E(a+f(X))=a+E(f(X))$,

Where, a is a constant and f(X) is the function.

## 3.2    Random Variable Definition

A random variable is a rule that assigns a numerical value to each outcome in a sample space. Random variables may be either discrete or continuous. A random variable is said to be discrete if it assumes only specified values in an interval. Otherwise, it is continuous. We generally denote the random variables with capital letters such as X and Y. When X takes values 1, 2, 3, …, it is said to have a discrete random variable.

As a function, a random variable is needed to be measured, which allows probabilities to be assigned to a set of potential values. It is obvious that the results depend on some physical variables which are not predictable. Say, when we toss a fair coin, the final result of happening to be heads or tails will depend on the possible physical conditions. We cannot predict which outcome will be noted. Though there are other probabilities like the coin could break or be lost, such consideration is avoided.

Variate

A variate can be defined as a generalization of the random variable. It has the same properties as that of the random variables without stressing to any particular type of probabilistic experiment. It always obeys a particular probabilistic law.

A variate is called discrete variate when that variate is not capable of assuming all the values in the provided range.

If the variate is able to assume all the numerical values provided in the whole range, then it is called continuous variate.

### Types of Random Variable

As discussed in the introduction, there are two random variables, such as:

- Discrete Random Variable
- Continuous Random Variable

Let's understand these types of variables in detail along with suitable examples below.

### Discrete Random Variable

A discrete random variable can take only a finite number of distinct values such as 0, 1, 2, 3, 4, and so on. The probability distribution of a random variable has a list of probabilities compared with each of its possible values known as probability mass function.

In an analysis, let a person be chosen at random, and the person's height is demonstrated by a random variable. Logically the random variable is described as a function which relates the person to the person's height. Now in relation with the random variable, it is a probability distribution that enables the calculation of the probability that the height is in any subset of likely values, such as the likelihood that the height is between 175 and 185 cm, or the possibility that the height is either less than 145 or more than 180 cm. Now another random variable could be the person's age which could be either between 45 years to 50 years or less than 40 or more than 50.

### Continuous Random Variable

A numerically valued variable is said to be continuous if, in any unit of measurement, whenever it can take on the values a and b. If the random variable X can assume an infinite and uncountable set of values, it is said to be a continuous random variable. When X takes any value in a given interval (a, b), it is said to be a continuous random variable in that interval.

Examples

Formally, a continuous random variable is such whose cumulative distribution function is constant throughout. There are no "gaps" in between which would compare to numbers which have a limited probability of occurring. Alternately, these variables almost never take an accurately prescribed value c but there is a positive probability that its value will rest in particular intervals which can be very small.

Task: What is difference between discrete and random variable.

## 3.3 Central Tendency

In statistics, the central tendency is the descriptive summary of a data set. Through the single value from the dataset, it reflects the centre of the data distribution. Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset. Generally, the central tendency of a dataset can be defined using some of the measures in statistics.

### Definition

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.

### Measures of Central Tendency

The central tendency of the dataset can be found out using the three important measures namely mean, median and mode.



### Mean

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean. Some other measures of mean used to find the central tendency are as follows:

- Geometric Mean
- Harmonic Mean
- Weighted Mean

It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs. Calculating the mean value is completely easy.

The histogram given below shows that the mean value of symmetric continuous data and the skewed continuous data.

Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics.

The central tendency is one of the most quintessential concepts in statistics. Although it does not provide information regarding the individual values in the dataset, it delivers a comprehensive summary of the whole dataset.

## Histogram of skewed continuous
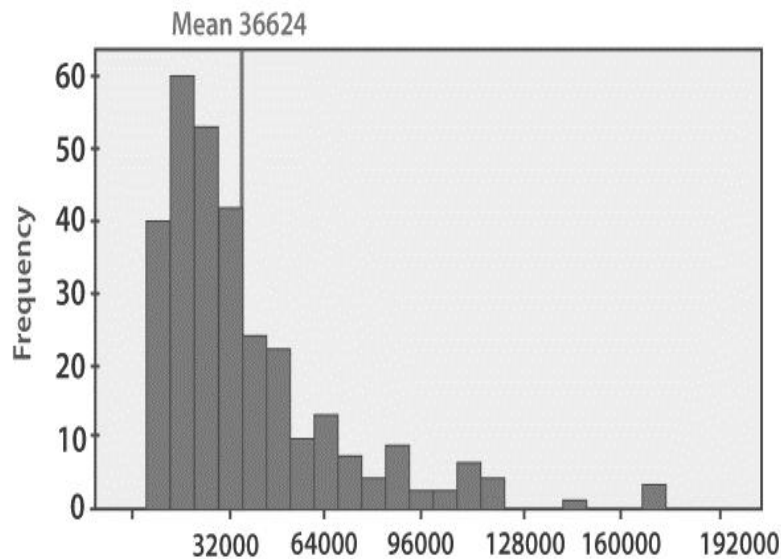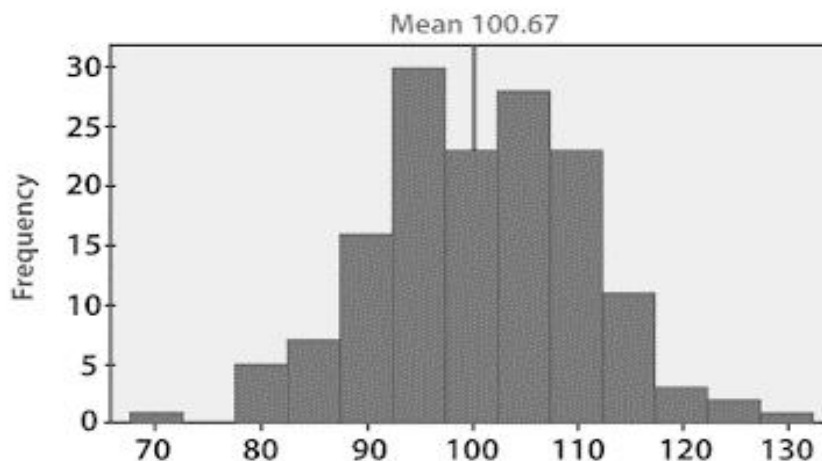
Mean 36624

## Histogram of symmetric continuous

Mean 100.67

In symmetric data distribution, the mean value is located accurately at the centre. But in the skewed continuous data distribution, the extreme values in the extended tail pull the mean value away from the centre. So it is recommended that the mean can be used for the symmetric distributions.

**Median**

Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.

For example, Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2

| Median odd |
| --- |
| 23 |
| 21 |
| 18 |
| 16 |
| 15 |
| 13 |
| 12 |
| 10 |
| 9 |
| 7 |
| 6 |
| 5 |
| 2 |

Here 12 is the middle or median number that has 6 values above it and 6 values below it.

For example, Now, consider another example with an even number of observations that are arranged in descending order – 40, 38, 35, 33, 32, 30, 29, 27, 26, 24, 23, 22, 19, and 17

| Median even |
| --- |
| 40 |
| 38 |
| 35 |
| 33 |
| 32 |
| 30 |
| 29 |
| 27 |
| 26 |
| 24 |
| 23 |
| 22 |
| 19 |
| 17 |

28

When you look at the given dataset, the two middle values obtained are 27 and 29.

Now, find out the mean value for these two numbers.

i.e.,(27+29)/2 =28

Therefore, the median for the given data distribution is 28.

**Mode**

The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5

| Mode |
| :---: |
| 5 |
| 5 |
| 5 |
| 4 |
| 4 |
| 3 |
| 2 |
| 2 |
| 1 |

Since the mode represents the most common value. Hence, the most frequently repeated value in the given dataset is 5.

Based on the properties of the data, the measures of central tendency are selected.

- If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset.
- If you have skewed distribution, the best measure of finding the central tendency is the median.
- If you have the original data, then both the median and mode are the best choice of measuring the central tendency.
- If you have categorical data, the mode is the best choice to find the central tendency.

Task: What are different conditions to use measures of central tendency?

## 3.4    What is Skewness and Why is it Important?

What is Skewed Data?

The measure of the asymmetry of a distribution of probability that is ideally symmetric and is given by the third standardized moment is skewness. In simple words, skew is the measure of how much a random variable'sprobability distributionvaries from the normal distribution.

**Lovely Professional University**

When both sides of the distribution are not distributed equally then this is known as Skewed Data. It is not a symmetrical distribution.

📋 Task: To quickly see if the data is skewed, we can use a histogram.

**Types of Skewness**

Well, the normal distribution is the distribution of the probability without any skewness.

There are two types of skewness, apart from this:

- Positive Skewness
- Negative Skewness

**Positive Skewness**

A positively skewed distribution (often referred to as Right-Skewed) is a distribution type where most values are concentrated to the left tail of the distribution whereas the right tail of the distribution is longer. A positively skewed distribution is the complete opposite of a negatively skewed distribution

A Positively Skewed Curve

In contrast to normally distributed data, where all central trend measurements (mean, median, and mode) are equal to each other, with positively skewed data, the observations are dispersed. The general relationship between the central tendency measures in a positively skewed distribution can be expressed using the following inequalities:

**Mean > Median > Mode**

**Negative Skewness**

A negatively skewed distribution (often referred to as Left-Skewed) is a kind of distribution where more values are on the right side of the distribution graph whereas the left tail of its distribution graph is longer.

A Negatively Skewed Curve

Apart from normally distributed data, where all central trend measurements (mean, median, and mode) are equal to each other, with negatively skewed data, the measurements are dispersed. The general relationship between central trend measures in the negatively skewed distribution can be displayed using the following inequality:

**Mode > Median > Mean**



**How to Find Skewness of Data?**

One measure of skewness would be to subtract the mean from the mode, then divide the difference by the Standard Deviation of the data. This is called Pearson's first coefficient of skewness. We have a dimensionless quantity as the explanation for dividing the difference. This explains why there is positive skewness in data skewed to the right. The mean is greater than the mode if the data set is skewed to the right, so subtracting the mode from the mean gives a positive number. A similar argument shows why there is negative skewness in data skewed to the left.

To calculate the asymmetry of a data set, Pearson's second coefficient of skewness is also used. We deduct the mode from the median for this value, multiply this number by 3 and then divide it by the Standard Deviation.

**Note:** If the data shows a strong mode, Pearson's first coefficient of skewness is useful. Pearson's second coefficient can be preferable if the data has a poor mode or several modes, as it does not depend on mode as a central tendency measure.

**Uses of Skewed Data**

In various contexts, skewed data arises very naturally. Incomes are skewed to the right because the mean can be significantly influenced by even a few people making millions of dollars, and there are no negative incomes. Similarly, details related to a product's lifetime, such as a light bulb brand, is skewed to the right. Here, zero is the smallest that a lifetime can be, and long-lasting light bulbs can give the data a positive skew.

**What is Skewness in Statistics?**

In statistics, if one asks what skewness is, it is the degree of asymmetry found in a distribution of probability. Distributions can exhibit to varying degrees right (positive) skewness or left (negative) skewness. Zero skewness exhibits a natural distribution (bell curve).

What Does Skewness Tell You?

Investors note skewness when judging a return distribution because it, like kurtosis, considers the extremes of the data set rather than focusing solely on the average. Short- and medium-term investors in particular need to look at extremes because they are less likely to hold a position long enough to be confident that the average will work itself out.

For example, Investors commonly use standard deviation to predict future returns, but the standard deviation assumes a normal distribution. As few return distributions come close to normal, skewness is a better measure on which to base performance predictions. This is due to skewness risk.

Skewness risk is the increased risk of turning up a data point of high skewness in a skewed distribution. Many financial models that attempt to predict the future performance of an asset assume a normal distribution, in which measures of central tendency are equal. If the data are skewed, this kind of model will always underestimate skewness risk in its predictions. The more skewed the data, the less accurate this financial model will be.

## 3.5   What is Kurtosis?

This is a statistical procedure used in reporting the distribution. Unlike skewness which differentiates extreme values between one tail and another, kurtosis computes the absolute values in each tail. Large kurtosis is present in the distributions that possess tail data surpassing the tails of the normal distribution. Conversely, the distributions that exhibit less extreme tail data in comparison to the tails of the normal distribution have low kurtosis. Investors interpret high kurtosis of the return distribution as a signal that they will face frequent and more extreme returns than the typical + or - three standard deviations from the mean that the normal distribution of returns predicts. This is called kurtosis risk.

**How is Kurtosis Used?**

Kurtosis is computed from the combination of the tails of a distribution relative to the center of the distribution. Graphing a set of approximately normal data through a histogram displays a bell peak and a majority of data within + or three standard deviations of the mean. These tails, however, go beyond the + or 3 standard deviations of the normal bell-curved distribution in the presence of high kurtosis. Kurtosis doesn't measure the peakedness of distribution but instead describes the shape of the distribution's tail in relation to its form.

This means that distribution can have an infinite peak and a low kurtosis and an infinite kurtosis but a perfectly flat top. It measures only tailedness.

**Types of Kurtosis**

A set of data can display up to three categories of kurtosis whose measures are compared against a bell curve. These categories are as follows:

**Mesokurtic distribution**. In this distribution, the kurtosis statistic is the same as that of the bell curve, and so the distribution's extreme value characteristic is the same as the one belonging to a normal distribution.

**Leptokurtic distribution**. This has a higher kurtosis than that of a mesokurtic distribution. It has elongated tails and is skinny because of the outliers stretching the horizontal axis of the histogram graph causing the most of the data to result in a narrow vertical range. This has led to the leptokurtic distribution being viewed as a concentrated towards the mean. However, some extreme outliers lead to concentration appearance.

**Platykurtic distribution**. These distributions possess short tails. Uniform distributions have broad peaks although the beta (.5, 1) has an infinitely pointy peak. These two distributions are platykurtic since their extreme values are lower than that of the bell curve. Investors view these distributions as stable and predictable because they don't often become extreme returns.

Task: How Skewness is different from Kurtosis?

## 3.6    What is Dispersion in Statistics?

Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which a numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.

### Measures of Dispersion

In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

### Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

### Absolute Measure of Dispersion

An absolute measure of dispersion contains the same unit as the original data set. Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

1.  **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
2.  **Variance:** Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance. Variance $(\sigma^2) = \sum(X-\mu)^2/N$
3.  **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. $= \sqrt{\sigma}$.
4.  **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
5.  **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

### Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

- Co-efficient of Range
- Co-efficient of Variation
- Co-efficient of Standard Deviation
- Co-efficient of Quartile Deviation
- Co-efficient of Mean Deviation
- Co-efficient of Dispersion

The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages. The dispersion coefficient is also used when two series with different measurement units are compared. It is denoted as C.D.

The common coefficients of dispersion are:

| C.D. In Terms of | Coefficient of dispersion |
|---|---|
| Range | C.D. = (Xmax − Xmin) / (Xmax + Xmin) |
| Quartile Deviation | C.D. = (Q3 − Q1)/(Q3 + Q1) |
| Standard Deviation (S.D.) | C.D. = S.D./Mean |
| Mean Deviation | C.D. = Mean deviation/Average |

## 3.7  Solved Example on Measures of Dispersion

For example

Problem: Below is the table showing the values of the results for two companies A, and B.

Measures of dispersion

|  | Company A | Company B |
|---|---|---|
| Number of employees | 900 | 1000 |
| Average daily wage | Rs. 250 | Rs. 220 |
| Variance in the distribution of wages | 100 | 144 |

Which of the company has a larger wage bill?

Calculate the coefficients of variations for both of the companies.

Calculate the average daily wage and the variance of the distribution of wages of all the employees in the firms A and B taken together.

Solution:

For Company A

No. of employees = n1 = 900, and average daily wages = ȳ1 = Rs. 250

We know, average daily wage = Total wages/Total number of employees

Or, Total wages = Total employees × average daily wage = 900 × 250 = Rs. 225000 … (i)

For Company B

No. of employees = n2 = 1000, and average daily wages = ȳ2 = Rs. 220

So, Total wages = Total employees × average daily wage = 1000 × 220 = Rs. 220000 … (ii)

Comparing (i), and (ii), we see that Company A has a larger wage bill.

For Company A

Variance of distribution of wages = $\sigma_1^2$ = 100

C.V. of distribution of wages  = 100 x standard deviation of distribution of wages/ average daily wages

Or, C.V. A = 100 × √100/250 = 100 × 10/250 = 4 … (i)

For Company B

Variance of distribution of wages = $\sigma_{22}$ = 144

C.V. B = 100 × √144∕220 = 100 × 12∕220 = 5.45 … (ii)

Comparing (i), and (ii), we see that Company B has greater variability.

For Company A and B, taken together

The average daily wages for both the companies taken together

ȳ = (n1 ȳ 1 + n2 ȳ 2)∕( n1 + n2) = (900 × 250 + 1000 × 220) ÷ (900 + 1000) = 445000∕1900 = Rs. 234.21

The combined variance, $\sigma_2$ = (1∕ n1 + n2) ÷ [n1 ($\sigma_1$ 2 + d1 2) + n2 ($\sigma_2$ 2 + d2 2)]

Here, d1 = ȳ1 −ȳ = 250 – 234.21 = 15.79, d2 = ȳ2 −ȳ = 220 – 234.21 = – 14.21.

Hence, $\sigma_2$ = [900 × (100 + 15.792) + 1000 × (144 + – 14.212)]∕(900 + 1000)

or, $\sigma_2$ = (314391.69 + 345924.10)∕1900 = 347.53.

## 3.8    Differences Between Skewness and Kurtosis

**Skewness**, in basic terms, implies off-center, so does in statistics, it means lack of symmetry. With the help of skewness, one can identify the shape of the distribution of data. **Kurtosis**, on the other hand, refers to the pointedness of a peak in the distribution curve. The main difference between skewness and kurtosis is that the former talks of the degree of symmetry, whereas the latter talks of the degree of peakedness, in the frequency distribution.

| BASIS FOR COMPARISON | SKEWNESS | KURTOSIS |
|---|---|---|
| Meaning | Skewness alludes the tendency of a distribution that determines its symmetry about the mean. | Kurtosis means the measure of the respective sharpness of the curve, in the frequency distribution. |
| Measure for | Degree of lopsidedness in the distribution. | Degree of tailedness in the distribution. |
| What is it? | It is an indicator of lack of equivalence in the frequency distribution. | It is the measure of data, which is either peaked or flat in relation to the normal distribution. |
| Represents | Amount and direction of the skew. | How tall and sharp the central peak is |

# Summary

Mathematical expectation, also known as the expected value**,** which is the summation of all possible values from a random variable.

**Skewness** refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.

Kurtosis is a statistical measure that **defines how heavily the tails of a distribution differ from the tails of a normal distribution**. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values

Dispersion is a statistical term that **describes the size of the distribution of values expected for a particular variable** and can be measured by several different statistics, such as range, variance, and standard deviation.

A **measure of central tendency** is a single value that attempts to describe a set of data by identifying the central position within that set of data.

The mode is the **value that appears most often in a set of data values**. ... Like the statistical mean and median, the mode is a way of expressing, in a (usually) single number, important information about a random variable or a population.

In statistics and probability theory, the **median** is the value separating the higher half from the lower half of a data sample.

# Keywords

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

Dispersion is a statistical term that describes the size of the distribution of values expected for a particular variable

The mode is the value that appears most often in a set of data values

# Self Assessment

1. _____, also known as the expected value.

   A. Mathematical expectation.

   B. Random Variables

   C. Continuous variable

   D. All of these

2. Mathematically, a _____ is a real-valued function whose domain is a sample space S

   of a random experiment.

   A. Mathematical expectation.

   B. Random Variable

   C. Continuous variable

   D. All of these

3. A variable which assumes infinite values of the sample space is a _____.

   A. Mathematical expectation.

B. Random Variable

C. Continuous random variable

D. All of these

4. _____is a descriptive summary of a dataset through a single value that reflects the center of the data distribution.

A. Random variable

B. Central Tendency

C. Sample space

D. All of the above

5. The _____ is the "middle" value in the list of numbers.

A. Median

B. Mode

C. Mean

D. Range

6. The _____ is the score that occurs most frequently in a set of data.

A. Median

B. Mode

C. Mean

D. Range

7. _____means the extent to which a numerical data is likely to vary about an average value.

A. Dispersion
B. Left Skewed
C. Right Skewed
D. All of the above

8. The value of skewness for a _____skewed distribution is greater than zero.

A. Positive
B. Negative
C. Neutral
D. All of the Above

9. _____measures the degree of peakedness of a frequency distribution.

A. Kurtosis

B. Skewness

C. Dispersion

D. All of the above

**Lovely Professional University**

10. When the peak of a curve becomes relatively high then that curve is called _____

A. Leptokurtic

B. Platykurtic.

C. Mesokurtic.

D. All of the above

11. When the curve is flat-topped, then it is called _____

A. Leptokurtic

B. Platykurtic.

C. Mesokurtic.

D. All of the above

12. _____is a visual display of data and statistical results.

A. Dispersion
B. Kurtosis
C. Graphical representation
D. All of the above

13. _____is a new technology that provides the users the tools to store the summarized

information.

A. Data Warehousing
B. Data mining
C. Data Dispersion
D. All of the above

14. Finding hidden patterns from data is called as

A. Data Warehousing
B. Data mining
C. Data Dispersion
D. All of the above

15.    The value of skewness for a _____skewed distribution is less than zero.
A.    Positive
B.    Negative
C.    Neutral
D.    All of the Above

## Answers for Self Assessment

| 1. | A | 2. | B | 3. | C | 4. | B | 5. | A |
|---|---|---|---|---|---|---|---|---|---|
| 6. | B | 7. | A | 8. | A | 9. | B | 10. | A |
| 11. | A | 12. | B | 13. | C | 14. | A | 15. | B |

## Review Questions

1. Why Mathematical expectation, also known as the expected value?
2. What is Skewness and Why is it Important?
3. What kurtosis tells us about distribution?
4. What is difference between kurtosis and skewness of data?
5. How Dispersion is measured? Explain it with example.
6. What is acceptable skewness and kurtosis?
7. How do you interpret skewness and kurtosis?
8. What do you do when your data is not normally distributed?
9. How do you know if your data is normally distributed?

## Further Readings

- An Introduction to Probability and Statistics

- Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

- First Course in Probability, A

- Book by Sheldon M. Ross

- Schaums Theory and Problems of Statistics

- Book by Murray R. Spiegel

- Introduction to Probability, Statistics, and Random ...

- Book by Hossein Pishro-Nik

## Web Links

- https://www.tutorialspoint.com
- www.webopedia.com
- https://www.britannica.com/science/probability

# Unit 04: MOMENTS

Contents

Objectives

Introduction

4.1     What is Chebyshev's Inequality?

4.2     Moments of a random variable

4.3     Raw vs central moment

4.4     Moment-generating function

4.5     What is Skewness and Why is it Important?

4.6     What is Kurtosis?

4.7     Cumulants

Summary

Keywords

Self-assessment questions

Answers for Self Assessment

Review Questions

## Objectives

- Understand basics of Moments in statistics.
- Learn concepts of Chebyshev's Inequality.
- Understand Concept of skewness and Kurtosis.
- Understand concept of Moment generating functions.
- Solve basic questions related to Chebyshev's Inequality.

## Introduction

In mathematics, the moments of a function are quantitative measures related to the shape of the function's graph. If the function represents mass, then the first moment is the center of the mass, and the second moment is the rotational inertia. If the function is a probability distribution, then the first moment is the expected value, the second central moment is the variance, the third standardized moment is the skewness, and the fourth standardized moment is the kurtosis. The mathematical concept is closely related to the concept of moment in physics. Moments are popularly used to describe the qualities of a distribution. They represent a convenient and unifying method for summarizing many of the most commonly used measures such as measures of tendency, variation, skewness and kurtosis.

Moments are statistical measures that for measuring certain characteristics of the distribution. Moments can be raw moments, central moments and moments about any arbitrary point. In probability theory, Chebyshev's inequality guarantees that, for a wide class of probability distributions, no more than a certain fraction of values can be more than a certain distance from the mean.

## 4.1 What is Chebyshev's Inequality?

Chebyshev's inequality is a probability theory that guarantees that within a specified range or distance from the mean, for a large range of probability distributions, no more than a specific fraction of values will be present. In other words, only a definite fraction of values will be found within a specific distance from the mean of a distribution.

The formula for the fraction for which no more than a certain number of values can exceed is 1/K2; in other words, 1/K2 of a distribution's values can be more than or equal to K standard deviations away from the mean of the distribution. Further, it also holds that 1–(1/K2) of a distribution's values must be within, but not including, K standard deviations away from the mean of the distribution.

Understanding Chebyshev's Inequality

Chebyshev's inequality states that within two standard deviations away from the mean contains 75% of the values, and within three standard deviations away from the mean contains 88.9% of the values. It holds for a wide range of probability distributions, not only the normal distribution.

Example

However, when applied to the normal distribution, Chebyshev's inequality is less precise than the 65-95-99.7 rule; yet, it is important to keep in mind that the theory applies to a far broader range of distributions. It should be noted that standard deviations equal to or less than one are not valid for Chebyshev's inequality formula.

The fraction of any set of numbers lying within k standard deviations of those numbers of the mean of those numbers is at least

$1-1/k^2$

Where k=the within number the standard deviation and k must be greater than 1

Example

Problem Statement −

Use Chebyshev's theorem to find what percent of the values will fall between 123 and 179 for a data set with mean of 151 and standard deviation of 14.

Solution −

We subtract 151-123 and get 28, which tells us that 123 is 28 units below the mean.

We subtract 179-151 and also get 28, which tells us that 151 is 28 units above the mean.

Those two together tell us that the values between 123 and 179 are all within 28 units of the mean. Therefore the "within number" is 28.

So we find the number of standard deviations, k, which the "within number", 28, amounts to by dividing it by the standard deviation −

k=the within numberthe standard deviation=28/14=2

So now we know that the values between 123 and 179 are all within 28 units of the mean, which is the same as within k=2 standard deviations of the mean. Now, since k > 1 we can use Chebyshev's formula to find the fraction of the data that are within k=2 standard deviations of the mean. Substituting k=2 we have −

$1-1/k^2 = 1-1/2^2 = 1-1/4 = 3/4$

**So 3/4 of the data lie between 123 and 179. And since 3/4=75% that implies that 75% of the data values are between 123 and 179.**

**Applications of Chebyshev's Inequality**

Example

Numerical Example-1:

Suppose that it is known that the number of products formed in a factory during a week is a random variable with a mean of 50. If the variance of a week production is equal to 25, then what can be said about the productivity that it will be between 40 and 60?

**Solution:**

Step-1: Mean ($\mu$) = 50, Variance ($\sigma^2$) = 25 $\Rightarrow\sigma$= 5

Step-2: Required probability: P (40 < X < 60)

= P (40 < X < 60) = P(-10 < X-50 < 10) = P(| X-50 | < 10)

Step-3: Now, by using the Chebyshev's theorem, we have P(| X-$\mu$ | <k$\sigma$) $\geq$ 1-1/k2

Find k by compare with general equation, therefore k$\sigma$ = 10 $\Rightarrow$k(5) =10 $\Rightarrow$ k=

Step-4: Apply the Chebyshev's Theorem to find the required probability:

$\geq$ 1-1/k² $\geq$ 1-(1/4) $\geq$ 3/4  $\geq$ 0.75

Step-5: Present the results

**Therefore, the lower bound of the probability that the productivity lies between 40 and 60 is equal to 0.75.**

Example

A symmetric die is thrown 600 times. Find the lower bound for the probability of getting 80 to 120 sixes.

Solution:

Step-1: A symmetric die is thrown 600 times, so it follows Binomial Distribution and p=1/6.

Step-2: Now, by using the binomial distribution, we have to calculate the mean and variance of the random variables using the given below formula:

$$MEAN: \quad \mu = np$$
$$Variance: \quad \sigma^2 = npq$$
$$SD: \quad \sigma = \sqrt{npq}$$

Mean = np = 600*(1/6) = 100

Variance = npq = 600*(1/6)*(5/6) = 500/6

Step-3: Required probability: P(80 < X < 120)

P(80 < X < 120) = P(-20 <X-100 < 20) = P(| X-100 | < 20)

Step-4: Now, by using the Chebyshev's theorem, we have P(| X-$\mu$ | <k$\sigma$) $\geq$1-1/k²

Find k by compare with general equation, therefore k$\sigma$ = 20 $\Rightarrow$k($\sqrt{(500/6)}$) = 10 $\Rightarrow$ k = 20$\sqrt{(6/500)}$

Step-5: Apply Chebyshev's Theorem to find the required probability:

$\geq$ 1-1/k2 $\geq$ 1-500/2400 $\geq$ 19/24 $\geq$ 0.79

Step-6: Present the results

**Therefore, the lower bound of the probability of getting sixes between 80 and 120 is equal to 0.79.**

Task: Explain applications of Chebyshev's Theorem

## 4.2 <u>Moments of a random variable</u>

In mathematics, the moments of a function are quantitative measures related to the shape of the function's graph. If the function represents mass, then the first moment is the center of the mass, and the second moment is the rotational inertia. If the function is a probability distribution, then the first moment is the expected value, the second central moment is the variance, the third standardized moment is the skewness, and the fourth standardized moment is the kurtosis. The mathematical concept is closely related to the concept of moment in physics.

Significance of the moments

It is possible to define moments for random variables in a more general fashion than moments for real-valued functions — see moments in metric spaces. The moment of a function, without further explanation, usually refers to the above expression with c = 0.

For the second and higher moments, the central moment (moments about the mean, with c being the mean) are usually used rather than the moments about zero, because they provide clearer information about the distribution's shape. The n-th moment about zero of a probability density function f(x) is the expected value of $X^n$ and is called a raw moment or crude moment.[3] The moments about its mean μ are called central moments; these describe the shape of the function, independently of translation.

The shape of any distribution can be described by its various 'moments'. The first four are:

1) The mean, which indicates the central tendency of a distribution.

2) The second moment is the variance, which indicates the width or deviation.

3) The third moment is the skewness, which indicates any asymmetric 'leaning' to either left or right.

4) The fourth moment is the Kurtosis, which indicates the degree of central 'peakedness' or, equivalently, the 'fatness' of the outer tails.

First moment- Mean

Measure the location of the central point.

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Second moment- Standard Deviation (SD, σ(Sigma)):

Measure the spread of values in the distribution OR how far from the normal.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}$$

σ = (Variance) ^.5

Small SD: Numbers are close to mean

High SD  : Numbers are spread out

For normal distribution:

Within 1 SD: 68.27% values lie

Within 2 SD: 95.45% values lie

Within 3 SD: 99.73% values lie

Advantages over Mean Absolute Deviation(MAD):

1. Mathematical properties- Continuous, differentiable.

2. SD of a sample is more consistent estimate for a population- When drawing repeated samples from a normally distributed population, the standard deviations of samples are less spread out as compare to mean absolute deviations.

Third moment- Skewness

Measure the symmetry in the distribution.

$$Skew = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{(X_i - \bar{X})}{\sigma} \right]^3$$

Skewness=0 [Normal Distribution, Symmetric]

Other Formulas:

1. Skewness = (Mean-Mode)/SD

2. Skewness = 3*(Mean-Median)/SD

(Mode = 3*Median-2*Mean)

Transformations (to make the distribution normal):

a. Positively skewed (right): Square root, log, inverse

b. Negatively skewed (left) : Reflect and square[sqrt(constant-x)], reflect and log, reflect and inverse
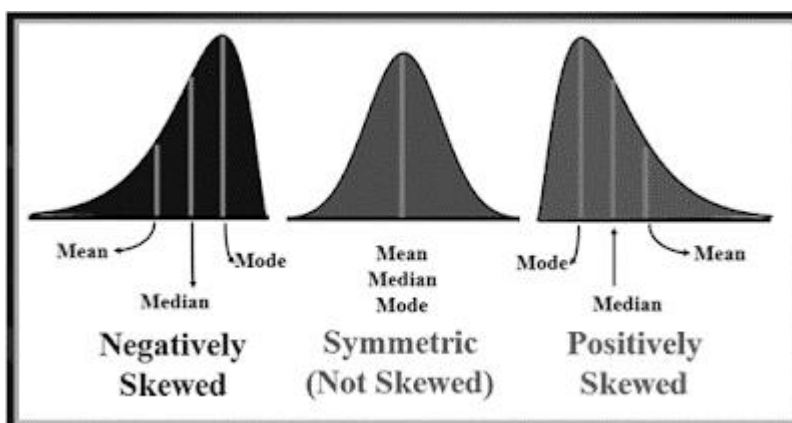


Fourth moment- Kurtosis:

Measure the amount in the tails.

$$Kurt = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^4$$

Kurtosis=3 [Normal Distribution]

Kurtosis<3 [Lighter tails]

Kurtosis>3 [Heavier tails]

Other Formulas:

Excess Kurtosis = Kurtosis - 3

Understanding:

Kurtosis is the average of the standardized data raised to fourth power. Any standardized values less than |1| (i.e. data within one standard deviation of the mean) will contribute petty to kurtosis.

The standardized values that will contribute immensely are the outliers.

High Kurtosis alerts about attendance of outliers.
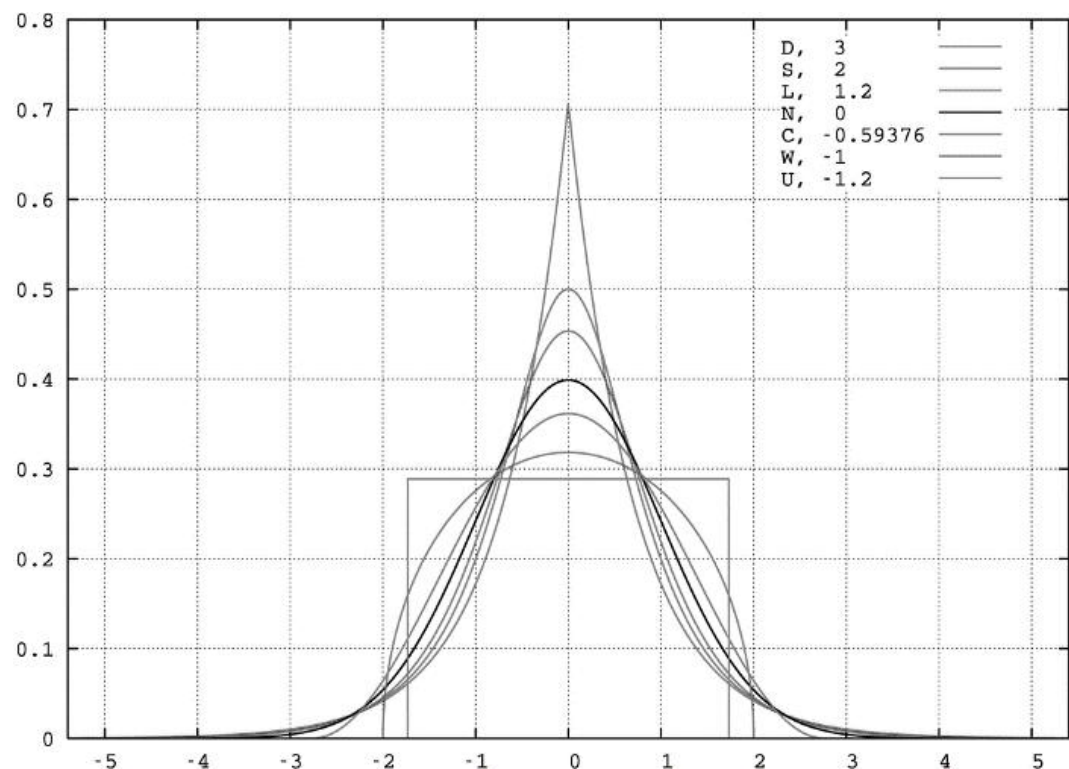


Excess Kurtosis for Distributions [Laplace (D)ouble exponential; Hyperbolic (S)ecant; (L)ogistic; (N)ormal; ©osine; (W)igner semicircle; (U)niform]

## 4.3 Raw vs central moment

Another approach helpful to find the summary measures for probability distribution is based on the 'moments'. We will discuss two types of moments.

i. Moments about the origin. (Origin may be zero or any other constant say A). It is also called as raw moments.

ii. Moments about the mean is called as central moments.

The **raw moments** (or 'moments about zero') $\mu_i{}'$ of a distribution are defined as

$$\mu_i{}' = \int_{-\infty}^{+\infty} x^i f(x)\,dx,$$

for continuous distributions with PDF *f(x)* and

$$\mu_i{}' = \sum_{k=0}^{n} x_k{}^i p_k$$

for discrete distributions with PMF *pi*.

The **central moments** (or 'moments about the mean') $\mu_i$ for $i \geq 2$ are defined as:

$$\mu_i = \int_{-\infty}^{+\infty} (x - \mu)^i f(x)\,dx$$

with analogue definitions for discrete variables. The lower central moments are directly related to the variance,

The second, third and fourth central moments can be expressed in terms of the raw moments as follows:

$$\mu_2 = \mu_2{}' - \mu^2$$

$$\mu_3 = \mu_3{}' - 3\mu\mu_2{}' + 2\mu^3$$

$$\mu_4 = \mu_4{}' - 4\mu\mu_3{}' + 6\mu^2\mu_2{}' - 3\mu^4$$

## 4.4 Moment-generating function

In probability theory and statistics, the **moment-generating function** of a real-valued random variable is an alternative specification of its probability distribution. Thus, it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions. There are particularly simple results for the moment-generating functions of distributions defined by the weighted sums of random variables. However, not all random variables have moment-generating functions.

As its name implies, the moment generating function can be used to compute a distribution's moments: the *n*th moment about 0 is the *n*th derivative of the moment-generating function, evaluated at 0.

In addition to real-valued distributions (univariate distributions), moment-generating functions can be defined for vector- or matrix-valued random variables, and can even be extended to more general cases.

The moment-generating function of a real-valued distribution does not always exist, unlike the characteristic function. There are relations between the behavior of the moment-generating function of a distribution and properties of the distribution, such as the existence of moments

Examples

Here are some examples of the moment-generating function and the characteristic function for comparison

| Distribution | Moment-generating function $M_X(t)$ | Characteristic function $\varphi(t)$ |
|---|---|---|
| Degenerate $\delta_a$ | $e^{ta}$ | $e^{ita}$ |
| Bernoulli $P(X=1)=$ | $1-p+pe^t$ | $1-p+pe^{it}$ |
| Geometric $(1-p)^{k-1}\,1$ | $\dfrac{pe^t}{1-(1-p)e^t}$ $\forall t < -\ln(1-p)$ | $\dfrac{pe^{it}}{1-(1-p)\,e^{it}}$ |
| Binomial $B(n,p)$ | $\left(1-p+pe^t\right)^n$ | $\left(1-p+pe^{it}\right)^n$ |
| Negative binomial $\mathrm{NB}(r,p)$ | $\left(\dfrac{pe^t}{1-e^t+pe^t}\right)^r$ | $\left(\dfrac{pe^{it}}{1-e^{it}+pe^{it}}\right)^r$ |
| Poisson $\mathrm{Pois}(\lambda)$ | $e^{\lambda(e^t-1)}$ | $e^{\lambda(e^{it}-1)}$ |
| Uniform (continuous) $\mathrm{U}(a,b)$ | $\dfrac{e^{tb}-e^{ta}}{t(b-a)}$ | $\dfrac{e^{itb}-e^{ita}}{it(b-a)}$ |
| Uniform (discrete) $\mathrm{DU}(a,b)$ | $\dfrac{e^{at}-e^{(b+1)t}}{(b-a+1)(1-e^t)}$ | $\dfrac{e^{ait}-e^{(b+1)it}}{(b-a+1)(1-e^{it})}$ |
| Laplace $L(\mu,b)$ | $\dfrac{e^{t\mu}}{1-b^2t^2},\ |t|<1/b$ | $\dfrac{e^{it\mu}}{1+b^2t^2}$ |
| Normal $N(\mu,\sigma^2)$ | $e^{t\mu+\frac{1}{2}\sigma^2t^2}$ | $e^{it\mu-\frac{1}{2}\sigma^2t^2}$ |

**Properties of MGFs**

- **Result 1 (MGF determines the distribution):** *If two random variables have the same MGF, they must have the same distribution.*

- **Result 2 (MGF of location-scale transformations ) :** *Let* **X** *be a random variable for which the MGF is* $M_1$ *and consider the random variable* $Y = aX + b$, *where* **a** *and* **b** *are given constants. Let the MGF of Y be denoted by* $M_2$. *Then for any value of* **t** *such that* $M_1(t)$ *exists,*

$$M_2(t) = e^{bt}M_1(at)$$

- **Result 3 (MGF of a sum of independent r.v.s) :** *Suppose that* $X_1,\dots,X_n$ *are* **n** *independent random variables and that* $M_i$ *is the MGF of* $X_i$. *Let* $Y = X_1 + \dots + X_n$ *and the MGF of Y be given by* **M**. *Then for any value of t such that* $M_i(t)$ *exists for all* $i = 1,2,\dots,n$,

$$M(t) = \prod_{i=1}^{n} M_i(t)$$

Task:  What is Moment generating function?

## 4.5 What is Skewness and Why is it Important?

What is Skewed Data?

The measure of the asymmetry of a distribution of probability that is ideally symmetric and is given by the third standardized moment is skewness. In simple words, skew is the measure of how much a random variable's probability distribution varies from the normal distribution.

When both sides of the distribution are not distributed equally then this is known as Skewed Data. It is not a symmetrical distribution.

To quickly see if the data is skewed, we can use a histogram.

**Lovely Professional University**

**Types of Skewness**

Well, the normal distribution is the distribution of the probability without any skewness.

There are two types of skewness, apart from this:

- Positive Skewness
- Negative Skewness

**Positive Skewness**

A positively skewed distribution (often referred to as Right-Skewed) is a distribution type where most values are concentrated to the left tail of the distribution whereas the right tail of the distribution is longer. A positively skewed distribution is the complete opposite of a negatively skewed distribution

A Positively Skewed Curve

In contrast to normally distributed data, where all central trend measurements (mean, median, and mode) are equal to each other, with positively skewed data, the observations are dispersed. The general relationship between the central tendency measures in a positively skewed distribution can be expressed using the following inequalities:

Mean >Median> Mode

Negative Skewness

A negatively skewed distribution (often referred to as Left-Skewed) is a kind of distribution where more values are on the right side of the distribution graph whereas the left tail of its distribution graph is longer.

A Negatively Skewed Curve

Apart from normally distributed data, where all central trend measurements (mean, median, and mode) are equal to each other, with negatively skewed data, the measurements are dispersed. The general relationship between central trend measures in the negatively skewed distribution can be displayed using the following inequality:

Mode >Median > Mean

How to Find Skewness of Data?

One measure of skewness would be to subtract the mean from the mode, then divide the difference by the Standard Deviation of the data. This is called Pearson's first coefficient of skewness. We have a dimensionless quantity as the explanation for dividing the difference. This explains why there is positive skewness in data skewed to the right. The mean is greater than the mode if the data set is skewed to the right, so subtracting the mode from the mean gives a positive number. A similar argument shows why there is negative skewness in data skewed to the left.

To calculate the asymmetry of a data set, Pearson's second coefficient of skewness is also used. We deduct the mode from the median for this value, multiply this number by 3 and then divide it by the Standard Deviation.

Note: If the data shows a strong mode, Pearson's first coefficient of skewness is useful. Pearson's second coefficient can be preferable if the data has a poor mode or several modes, as it does not depend on mode as a central tendency measure.

**Uses of Skewed Data**

In various contexts, skewed data arises very naturally. Incomes are skewed to the right because the mean can be significantly influenced by even a few people making millions of dollars, and there are no negative incomes. Similarly, details related to a product's lifetime, such as a light bulb brand, is skewed to the right. Here, zero is the smallest that a lifetime can be, and long-lasting light bulbs can give the data a positive skew.

**What is Skewness in Statistics?**

In statistics, if one asks what skewness is, it is the degree of asymmetry found in a distribution of probability. Distributions can exhibit to varying degrees right (positive) skewness or left (negative) skewness. Zero skewness exhibits a natural distribution (bell curve).



*What Does Skewness Tell You?*

Investors note skewness when judging a return distribution because it, like kurtosis, considers the extremes of the data set rather than focusing solely on the average. Short- and medium-term investors in particular need to look at extremes because they are less likely to hold a position long enough to be confident that the average will work itself out.

Investors commonly use standard deviation to predict future returns, but the standard deviation assumes a normal distribution. As few return distributions come close to normal, skewness is a better measure on which to base performance predictions. This is due to skewness risk.

Skewness risk is the increased risk of turning up a data point of high skewness in a skewed distribution. Many financial models that attempt to predict the future performance of an asset assume a normal distribution, in which measures of central tendency are equal. If the data are skewed, this kind of model will always underestimate skewness risk in its predictions. The more skewed the data, the less accurate this financial model will be.

## 4.6 What is Kurtosis?

This is a statistical procedure used in reporting the distribution. Unlike skewness which differentiates extreme values between one tail and another, kurtosis computes the absolute values in each tail. Large kurtosis is present in the distributions that possess tail data surpassing the tails of the normal distribution. Conversely, the distributions that exhibit less extreme tail data in comparison to the tails of the normal distribution have low kurtosis. Investors interpret high kurtosis of the return distribution as a signal that they will face frequent and more extreme returns than the typical + or - three standard deviations from the mean that the normal distribution of returns predicts. This is called kurtosis risk.

How is Kurtosis Used?

Kurtosis is computed from the combination of the tails of a distribution relative to the center of the distribution. Graphing a set of approximately normal data through a histogram displays a bell peak and a majority of data within + or three standard deviations of the mean. These tails, however, go beyond the + or 3 standard deviations of the normal bell-curved distribution in the presence of high kurtosis. Kurtosis doesn't measure the peakedness of distribution but instead describes the shape of the distribution's tail in relation to its form.

This means that distribution can have an infinite peak and a low kurtosis and an infinite kurtosis but a perfectly flat top. It measures only tailedness.

### Types of Kurtosis

A set of data can display up to three categories of kurtosis whose measures are compared against a bell curve. These categories are as follows:

Mesokurtic distribution. In this distribution, the kurtosis statistic is the same as that of the bell curve, and so the distribution's extreme value characteristic is the same as the one belonging to a normal distribution.

Leptokurtic distribution. This has a higher kurtosis than that of a mesokurtic distribution. It has elongated tails and is skinny because of the outliers stretching the horizontal axis of the histogram graph causing the most of the data to result in a narrow vertical range. This has led to the leptokurtic distribution being viewed as a concentrated towards the mean. However, some extreme outliers lead to concentration appearance.

Platykurtic distribution. These distributions possess short tails. Uniform distributions have broad peaks although the beta (.5, 1) has an infinitely pointy peak. These two distributions are platykurtic since their extreme values are lower than that of the bell curve. Investors view these distributions as stable and predictable because they don't often become extreme returns.





Mixed moments

Mixed moments are moments involving multiple variables.

Some examples are covariance, coskewness and cokurtosis. While there is a unique covariance, there are multiple co-skewnesses and co-kurtoses.

Higher moments

High-order moments are moments beyond 4th-order moments. As with variance, skewness, and kurtosis, these are higher-order statistics, involving non-linear combinations of the data, and can be used for description or estimation of further shape parameters. The higher the moment, the harder it is to estimate, in the sense that larger samples are required in order to obtain estimates of similar quality. This is due to the excess degrees of freedom consumed by the higher orders. Further, they can be subtle to interpret, often being most easily understood in terms of lower order moments – compare the higher derivatives of jerk and jounce in physics. For example, just as the 4th-order moment (kurtosis) can be interpreted as "relative importance of tails versus shoulders in causing dispersion" (for a given dispersion, high kurtosis corresponds to heavy tails, while low kurtosis corresponds to broad shoulders), the 5th-order moment can be interpreted as measuring "relative importance of tails versus center (mode, shoulders) in causing skew" (for a given skew, high 5th moment corresponds to heavy tail and little movement of mode, while low 5th moment corresponds to more change in shoulders).

## 4.7 <u>Cumulants</u>

In probability theory and statistics, the cumulants $\kappa_n$ of a probability distribution are a set of quantities that provide an alternative to the moments of the distribution. The moments determine the cumulants in the sense that any two probability distributions whose moments are identical will have identical cumulants as well, and similarly the cumulants determine the moments.

The first cumulant is the mean, the second cumulant is the variance, and the third cumulant is the same as the third central moment. But fourth and higher-order cumulants are not equal to central moments. In some cases theoretical treatments of problems in terms of cumulants are simpler than those using moments. In particular, when two or more random variables are statistically independent, the nth-order cumulant of their sum is equal to the sum of their nth-order cumulants. As well, the third and higher-order cumulants of a normal distribution are zero, and it is the only distribution with this property.

- **Moment generating functions** are a way to find **moments** like the mean($\mu$) and the variance($\sigma^2$). They are an alternative way to represent a probability distribution with a simple one-variable **function.**

- **Each probability distribution has a unique MGF**, which means they are especially useful for solving problems like finding the distribution for sums of random variables. They can also be used as a proof of the Central Limit Theorem.

Why the Cumulant Generating Function is Important?

The cumulant generating function is important because both it and the cumulants lend themselves so well to mathematical analysis, besides (in the case of the cumulants) being meaningful in their own right. They change in simple, easy to understand ways when their underlying PDF (probability density function) is changed, and they are easy to define on most spaces.

- The cumulant generating function is infinitely differentiable, and it passes through the origin. Its first derivative is monotonic function from the least to the greatest upper bounds of the probability distribution.

- Its second derivative is positive everywhere where it is defined.

- Cumulants accumulate: the $k^{th}$ cumulant of a sum of independent random variables is just the sum of the $k^{th}$ cumulants of the summands.

- Cumulants also have a scaling property: the nth cumulant of n X is $c^n$ times the nth cumulant of X.

## Summary

- Chebyshev's inequality is a probabilistic inequality. It provides an upper bound to the probability that the absolute deviation of a random variable from its mean will exceed a given threshold.

- Chebyshev's inequality is more general, stating that a minimum of just 75% of values must lie within two standard deviations of the mean and 88.89% within three standard deviations for a broad range of different probability distributions

- Moments are a set of statistical parameters to measure a distribution.

- Standard deviation is the square root of the variance: an indication of how closely the values are spread about the mean. A small standard deviation means the values are all similar. If the distribution is normal, 63% of the values will be within 1 standard deviation.

- Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess.

- While skewness signifies the extent of asymmetry, kurtosis  measures the degree of peakedness of a frequency distribution

## Keywords

- Moments are popularly used to describe the qualities of a distribution. They represent a convenient and unifying method for summarizing many of the most commonly used measures such as measures of tendency, variation, skewness and kurtosis

- Moments are statistical measures that for measuring  certain characteristics of the distribution. Moments can be raw moments, central moments and moments about any arbitrary point.

- Mode: Defines the most frequently occurring value in a dataset. In some cases, a dataset may contain multiple modes while some datasets may not have any mode at all.

- The first cumulant is the mean, the second cumulant is the variance, and the third cumulant is the same as the third central moment.

## Self-assessment questions

1. _____ inequality is a probabilistic inequality.

a)  Chebyshev's

b)  Bayes

c)  Charles

d)  All of the above

2. _____ are a set of statistical parameters to measure a distribution.

a)  Moments

b)  Kurtosis

c)  Skewness

d)  Variance

**3. First Moment is \_\_\_**

a) Mean

b) Median

c) Variance

d) Skewness

**4. Standard deviation is the square root of the _____-**

a) Mean

b)Median

c)Variance

d)Skewness

**5.Third Moment is**

a)Mean

b)Median

c)Variance

d)Skewness

**6._____ also known as the expected value, which is the summation of all possible values from a random variable.**

a)Mathematical expectation

b)Skewness

c)Kurtosis

d)Random variable

**7.Mathematically, a _____is a real-valued function whose domain is a sample space S of a random experiment.**

a)Mathematical expectation

b)Skewness

c)Kurtosis

d)Random variable

**8.In probability theory and statistics, the _____ of a real-valued random variable is an alternative specification of its probability distribution**

a.Mathematical expectation

b.Skewness

c.Kurtosis

d.Moment generating function

**9.Binomial distribution is**

a)Discrete distribution

b)Continuous distribution

**Lovely Professional University**

c)Normal distribution

d)Poison distribution

**10.Poison distribution is**

a)Discrete distribution

b)Continuous distribution

c)Normal distribution

d)Poison distribution

**11.In probability theory, a _____is a type of continuous probability distribution**

a)Geometric Distribution

b)Bayes distribution

c)Normal distribution

d)Poison distribution

**12.In probability theory and statistics, the _____ with parameters n and p is the discrete probability distribution**.

a)Binomial Distribution

b)Bayes distribution

c)Normal distribution

d)Poison distribution

**13.The middle value measure of central tendency in a dataset that is arranged in ascending order**

a)Mean

b)Median

c)Mode

d)All of the above

**14.In probability theory and statistics, the _____ of a probability distribution are a set of quantities that provide an alternative to the moments of the distribution.**

a)Cumulants

b)Mean

c)Median

d)Mode

**15.If skewness is _____the mean is smaller than the median and the distribution has a large tail of small values.**

a)Negative

b)Positive

c)Zero

d)All of the above

## Answers for Self Assessment

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 a | 2. a | 3 a | 4 c | 5 d | 6 a | 7 d | 8 d |
| 9 a | 10 a | 11 c | 12 a | 13 b | 14 a | 15 a | |

## Review Questions

1. What is the use of Chebyshev inequality?
2. What does Chebyshev's inequality measure?
3. What does moments mean in statistics?
4. What is the use of moments in statistics?
5. How lower central moments are directly related to the variance, skewness and kurtosis.
6. What are first and second moments?
7. Why skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean?
8. How does skewness effect mean?
9. Explain concept of kurtosis with example?
10. What is acceptable skewness and kurtosis?

## Further Readings

- An Introduction to Probability and Statistics
- Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi
- First Course in Probability, A
- Book by Sheldon M. Ross
- Schaums Theory and Problems of Statistics
- Book by Murray R. Spiegel
- Introduction to Probability, Statistics, and Random ...
- Book by Hossein Pishro-Nik

**Online links**

- https://www.tutorialspoint.com
- www.webopedia.com
- https://www.britannica.com/science/probability

# Unit05:Relation between Moments

| CONTENTS |
|---|
| Objectives |
| Introduction |
| 5.1    Discrete and Continuous Data |
| 5.2    Difference between Discrete and Continuous Data |
| 5.3    Moments in Statistics |
| 5.4    Scale and Origin |
| 5.5    Effects of change of origin and change of scale |
| 5.6    Skewness |
| 5.7    Kurtosis Measures |
| 5.8    Why Standard Deviation Is an Important Statistic |
| Summary |
| Keywords |
| Self Assessment |
| Answers for Self Assessment |
| Review Questions |
| Further Readings |

## Objectives

- Understand basics of Moments.
- Learn concepts of change of origin
- Understand Concept of measuring skewness and Kurtosis.
- Understand concept of change of scale
- Solve basic questions related to Pearson coefficient.

## Introduction

Central Tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics. The change of origin means that some value has been either added or subtracted to the observations. The change of scale means that some value is either multiplied or divided to the observations

For example –

Effect of change of origin on the mean. If the mean of observations is 7 and 3 is added to all observations, then new mean also increases by 3, and if we subtract 3, the mean decreases by 3.

For example –

Effect of change of scale on mean. If each observation is multiplied by x, then the new mean becomes x times of initial mean.  If it is divided divide with x then the new mean = initial mean/x.

## 5.1    Discrete and Continuous Data

Discrete Data

Discrete Data can only take certain values.



Example: the number of students in a class

We can't have half a student!

Example: the results of rolling 2 dice

Only has the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12

**Continuous Data**



Continuous Data can take any value (within a range)

Examples:

- A person's height: could be any value (within the range of human heights), not just certain fixed heights,

  Time in a race: you could even measure it to fractions of a second,

Example: A dog's weight,

- The length of a leaf,

## 5.2    Difference between Discrete and Continuous Data

In statistics, data is defined as the facts and figures collected together for the purpose of analysis. It is divided into two broad categories, qualitative data, and quantitative data. Further, the qualitative

**Lovely Professional University**

data is cannot be measured in terms of numbers and it is sub-divided into nominal and ordinal data. On the other hand, quantitative data is one that contains numerical values and uses range. It is sub-classified as discrete and continuous data. Discrete data contains finite values that have nothing in-between

In statistics, data is defined as the facts and figures collected together for the purpose of analysis. It is divided into two broad categories, Qualitative data, and

**Quantitative Data**.

Further, the qualitative data is cannot be measured in terms of numbers and it is sub-divided into nominal and ordinal data. On the other hand, quantitative data is one that contains numerical values and uses range. It is sub-classified as discrete and continuous data. Discrete data contains finite values that have nothing in-betweenAs against, continuous data contains data that can be measured, that includes fractions and decimals. Take a read of the article to know the difference between discrete and continuous data

Task: what is difference between discrete and continuous data?

**Comparison Chart**

| BASIS FOR COMPARISON | DISCRETE DATA | CONTINUOUS DATA |
| --- | --- | --- |
| Meaning | Discrete data is one that has clear spaces between values. | Continuous data is one that falls on a continuous sequence. |
| Nature | Countable | Measurable |
| Values | It can take only distinct or separate values. | It can take any value in some interval. |
| Graphical Representation | Bar Graph | Histogram |
| Tabulation is known as | Ungrouped frequency distribution. | Grouped frequency distribution. |
| Classification | Mutually Inclusive | Mutually Exclusive |

| BASIS FOR COMPARISON | DISCRETE DATA | CONTINUOUS DATA |
|---|---|---|
| Function graph | Shows isolated points | Shows connected points |
| Example | Days of the week | Market price of a product |

### Definition of Discrete Data

The term discrete implies distinct or separate. So, discrete data refers to the type of quantitative data that relies on counts. It contains only finite values, whose subdivision is not possible. It includes only those values that can only be counted in whole numbers or integers and are separate which means the data cannot be broken down into fraction or decimal.

For example, Number of students in the school, the number of cars in the parking lot, the number of computers in a computer lab, the number of animals in a zoo, etc.

### Definition of Continuous Data

Continuous data is described as an unbroken set of observations; that can be measured on a scale. It can take any numeric value, within a finite or infinite range of possible value. Statistically, range refers to the difference between highest and lowest observation. The continuous data can be broken down into fractions and decimal, i.e. it can be meaningfully subdivided into smaller parts according to the measurement precision.

For Example, Age, height or weight of a person, time taken to complete a task, temperature, time, money, etc.

## 5.3   Moments in Statistics

Moments are a set of statistical parameters to measure a distribution. Four moments are commonly used: 1st, Mean: the average. 2d, Variance: Standard deviation is the square root of the variance: an indication of how closely the values are spread about the mean

### Use of Moments in Statistics

Moments are are very useful in statistics because they tell you much about your data. There are four commonly used moments in statistics: the mean, variance, skewness, and kurtosis. The mean gives you a measure of center of the data

### What are raw moments in statistics?

A moment of a probability function taken about 0, (1) (2) The raw moments (sometimes also called "crude moments") can be expressed as terms of the central moments (i.e., those taken about the mean ) using the inverse binomial transform.

### What do you mean by central moments?

**Lovely Professional University**

In probability theory and statistics, a central moment is a moment of a probability distribution of a random variable about the random variable's mean; that is, it is the expected value of a specified integer power of the deviation of the random variable from the mean.

**Moments of a statistical distribution**

The shape of any distribution can be described by its various 'moments'. The first four are:

1) The mean, which indicates the central tendency of a distribution.

2) The second moment is the variance, which indicates the width or deviation.

3) The third moment is the skewness, which indicates any asymmetric 'leaning' to either left or right.

4) The fourth moment is the Kurtosis, which indicates the degree of central 'peakedness' or, equivalently, the 'fatness' of the outer tails.

## 5.4   Scale and Origin

However, sometimes, if we have coordinates that are large numbers, we may need to scale the axis differently. In the graph below, each mark on the axis represents 500 units.
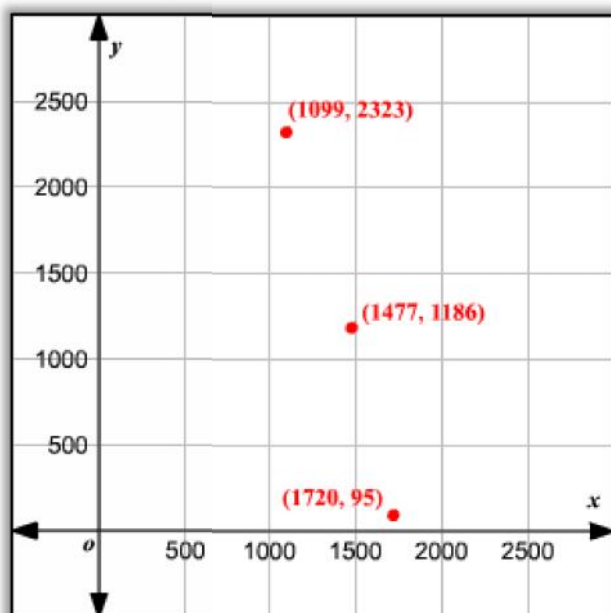


**Change of origin and scale**

- Change of origin and scale make calculations easy as, in origin change, the distribution location is changed, while in scale change, the distribution shape is altered.

- The change of origin means that some value has been either added or subtracted to the observations.

- The change of scale means that some value is either multiplied or divided to the observations.

For example - Effect of change of origin on the mean. If the mean of observations is 7 and 3 is added to all observations, then new mean also increases by 3, and if we subtract 3, the mean decreases by 3.

- If each observation is multiplied by x, then the new mean becomes x times of initial mean. If it is divided divide with x then the new mean = initial mean/x.

- Make calculations easy as, in origin change, the distribution location is changed, while in scale change, the distribution shape is altered.

Task: What is effect of change of origin vs. change of scale?

## 5.5 Effects of change of origin and change of scale

- Following are the effects of change of origin and change of scale on the mean, standard deviation and variance.

- Any constant added or subtracted (Change of origin) than the standard deviation of original data and of change data after addition as subtracted will not change but the mean of new data will change.

- Any constant multiplied are divided (change of scale) then mean, standard deviation and variation will change of the new changed data

### For ease of calculations

Change of scale and change of origin is done either for ease of calculations (in case of grouped data) or to make the distribution look a bit standardized. Change of scale changes standard deviation and variance. Change of origin changes mean, median,mode

- A change in scale would be, for example, measuring something in meters instead of feet. Everything would get scaled down by a factor of about 0.3.

- A change of origin would be if you started counting or measuring at a different point. For example, switching between centigrade and degrees Kelvin is essentially a change in origin (a shift of about 273 degrees).

Example: For calculating the mean, changing the origin for your data means it will affect the mean.

Taking the mean of x,y and z (this is x+y+z/3 ) is different from the

mean of x+A,y+A and z+A (which is x+y+z/3+A ),

unless of course A=0 .

However, the standard deviation does not depend on where the origin is, as long as the scale is the same.

### Effect of Change of origin on mean.

- If mean of some observations is 5

- If we add 3 to all observations then new mean also increases by 3, similarly if we subtract 3 to all observations again mean decreases by 3.

- If we multiply each observations by x.then new mean become x times of initial mean.

- If we divide all observations with x then new mean =initial mean/xShifting (addition and subtraction)

Example: What happens to measures of central tendency and spread when we add a constant value to every value in the data set? To answer this question, let's pretend we have the data set 3,3, 7, 9,13 and let's calculate our measures for the set.

- Mean: (3+3+7+9+13)/5=7

- Median: 7

- Mode: 3

- Range: 13-3=10

- IQR: 11-3=8

Example: If we add 6 to each data point in the set, the new set is 9,9, 13, 15,19, . And our new measures of central tendency and spread are explained as

- Mean: (9+9+13+15+19)/5=13(9+9+13+15+19)/5=13

- Median: 13

- Mode: 9

- Range: 19-9=19−9=10

- IQR: 17-9=17−9=8

- What we see is that adding 6 to the entire data set also adds 6 to the mean, median, and mode, but that the range and IQR stay the same.

Scaling (multiplication and division)

What happens to measures of central tendency and spread when we multiply a constant value to every value in the data set? To answer this question, let's pretend we have the data set 3,3, 7, 9,13 and let's calculate our measures for the set.

- Mean: (3+3+7+9+13)/5=7

- Median: 7

- Mode: 3

- Range: 13-3=10

- IQR: 11-3=8

- Let's multiply the set by 2, making the new set 6,6, 14, 18, 26,. The new measures of central tendency and spread .

New values

- Mean: (6+6+14+18+26)/5=14

- Median: 14

- Mode: 6

- Range: 26-6=20

- IQR: 22-6=16

What we see is that multiplying the entire data set by 2 multiplies all five measures by 2as well. The mean, median, mode, range, and IQR are all doubled when we double the values in the data set.

## 5.6  Skewness

- Skewness is the measure of the shape of a nonsymmetrical distribution

- Two sets of data can have the same mean & SD but different skewness

- Two types of skewness:

- Positive skewness
- Negative skewness

Karl Pearson's Coefficient of Skewness……01

This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

$$SK_P = \frac{Mean - Mode}{\sigma}$$

Where, SKP = Karl Pearson's Coefficient of skewness,

$\sigma$ = standard deviation.

Karl Pearson's Coefficient of Skewness…..02

In case the mode is indeterminate, the coefficient of skewness is:

$$SK_P = \frac{Mean - (3\ Median - 2\ Mean)}{\sigma}$$

Now this formula is equal to

$$SK_P = \frac{3(Mean - Median)}{\sigma}$$

- The value of coefficient of skewness is zero, when the distribution is symmetrical.
- The value of coefficient of skewness is positive, when the distribution is positivelyskewed.
- The value of coefficient of skewness is negative, when the distribution is negativelyskewed.

Example:

**Sample problem:**

Use Pearson's Coefficient #1 and #2 to find the skewness for data with the following characteristics:

Mean = 70.5.

Median = 80.

Mode = 85.

Standard deviation = 19.33.

- Pearson's Coefficient of Skewness #1 (Mode):

- Step 1: Subtract the mode from the mean: 70.5 – 85 = -14.5.

- Step 2: Divide by the standard deviation: -14.5 / 19.33 = -0.75.

- Pearson's Coefficient of Skewness #2 (Median):

- Step 1: Subtract the median from the mean: 70.5 – 80 = -9.5.

- Step 2: Multiply Step 1 by 3: -9.5(3) = -28.5

- Step 2: Divide by the standard deviation: -28.5 / 19.33 = -1.47.

Example: Caution: Pearson's first coefficient of skewness uses the mode. Therefore, if the mode is made up of too few pieces of data it won't be a stable measure of central tendency. For example, the mode in both these sets of data is 9:

- 1 2 3 4 5 6 7 8 9 9

- 1 2 3 4 5 6 7 8 9 9 9 9 9 9 9 9 9 9 9 9 10 12 12 13.

- In the first set of data, the mode only appears twice. This isn't a good measure of central tendency so you would be cautioned not to use Pearson's coefficient of skewness. The second set of data has a more stable set (the mode appears 12 times). Therefore, Pearson's coefficient of skewness will likely give you a reasonable result.

In general:

- The direction of skewness is given by the sign.

- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.

- A value of zero means no skewness at all.

- A large negative value means the distribution is negatively skewed.

- A large positive value means the distribution is positively skewed.

Skewness provides valuable information about the distribution of returns. However, skewness must be viewed in conjunction with the overall level of returns. Skewness by itself isn't very useful. It is entirely possible to have positive skewness (good) but an average annualized return with **a low or negative value (bad)**.

Example: The number of students absent in a class was recorded every day for 60 days and the information is given in the following frequency distribution.

| 0 | 1 | 2 | 3 | 4 | 5 | 6(students absent $x$) |
|---|---|---|---|---|---|---|
| 3 | 6 | 18 | 18 | 8 | 5 | 2 (days $f$) |

Find the Karl Pearson coefficient of skewness.

| $x_i$ | $f_i$ | $f_i * x_i$ | $f_i * x_i^2$ | $cf$ |
|---|---|---|---|---|
| 0 | 3 | 0 | 0 | 3 |
| 1 | 6 | 6 | 6 | 9 |
| 2 | 18 | 36 | 72 | 27 |
| 3 | 18 | 54 | 162 | 45 |
| 4 | 8 | 32 | 128 | 53 |
| 5 | 5 | 25 | 125 | 58 |
| 6 | 2 | 12 | 72 | 60 |
| Total | | 60 | 165 | 565 |

**Sample mean**

The sample mean of $X$ is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i$$
$$= \frac{165}{60}$$
$$= 2.75$$

The average of no. of students absent is 2.75 students.

- The cumulative frequency just greater than or equal to 30 is 45. The corresponding value of  is median. That is, M=3

- Thus, median number of accidents M= 3.

- Thus the standard deviation of no. of students absent is 1.3732 students.

| $x_i$ | $f_i$ | $f_i * x_i$ | $f_i * x_i^2$ | $cf$ |
|---|---|---|---|---|
| 0 | 3 | 0 | 0 | 3 |
| 1 | 6 | 6 | 6 | 9 |
| 2 | 18 | 36 | 72 | 27 |
| 3 | 18 | 54 | 162 | 45 |
| 4 | 8 | 32 | 128 | 53 |
| 5 | 5 | 25 | 125 | 58 |
| 6 | 2 | 12 | 72 | 60 |
| Total | | 60 | 165 | 565 |

**Lovely Professional University**

The Karl Pearson coefficient of skewness can be calculated by

$$s_k = \frac{3(Mean - Median)}{sd}$$
$$= \frac{3 \times (2.75 - 3)}{2.1602}$$
$$= -0.5462$$

As the value of $s_k < 0$, the data is negatively skewed.

## 5.7 Kurtosis Measures

If we have the knowledge of the measures of central tendency, dispersion and skewness, even then we cannot get a complete idea of a distribution. In addition to these measures, we need to know another measure to get the complete idea about the shape of the distribution which can be studied with the help of Kurtosis. Prof. Karl Pearson has called it the "Convexity of a Curve". Kurtosis gives a measure of flatness of distribution.

A measure of whether the curve of a distribution is:

- Bell-shaped -- Mesokurtic

- Peaked -- Leptokurtic

- Flat -- Platykurtic

The degree of kurtosis of a distribution is measured relative to that of a normal curve. The curves with greater peakedness than the normal curve are called "Leptokurtic". The curves which are more flat than the normal curve are called "Platykurtic". The normal curve is called "Mesokurtic.":
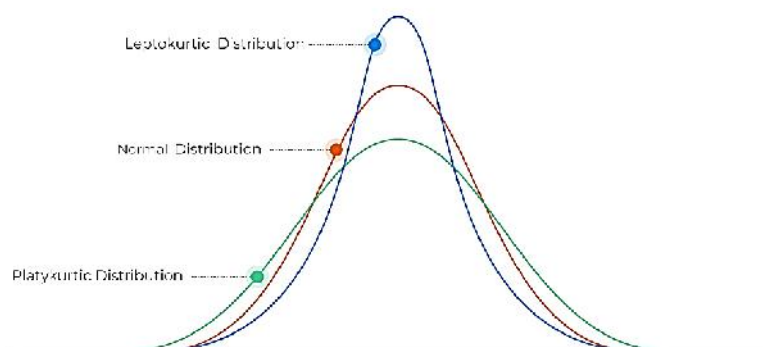
| *Definition of Kurtosis* | For univariate data $Y_1$, $Y_2$, ..., $Y_N$, the formula for kurtosis is: |
|---|---|

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4 / N}{s^4}$$

where $\bar{Y}$ is the mean, $s$ is the standard deviation, and $N$ is the number of data points. Note that in computing the kurtosis, the standard deviation is computed using $N$ in the denominator rather than $N$ - 1.

- Based on the value of kurtosis, the distribution can be classified into three categories.

- The distribution with kurtosis equal to 3 is known as mesokurtic. A random variable which follows normal distribution has kurtosis 3.

If the kurtosis is less than three, the distribution is called as platykurtic. Here, the distribution has shorter and thinner tails than normal distribution. Moreover, the peak is lower and also broader when compared to normal distribution. If the kurtosis is greater than three, the distribution is called as leptokurtic. Here, the distribution has longer and fatter tails than normal distribution. Moreover, the peak is higher and also sharper when compared to normal distribution.

📃Example:

- Suppose we have the following observations:

- {12  13  54  56  25}

- *Using the data from the example above (12  13  54  56  25), determine the type of kurtosis present.*

- *A. Mesokurtic distribution*

- *B. Platykurtic distribution*

- *C. Leptokurtic distribution*

The correct answer is B.

$$X = \frac{(12 + 13 + \ldots + 25)}{5} = \frac{160}{5} = 32$$

$$S^2 = \frac{(12 - 32)^2 + \ldots + (25 - 32)^2}{4} = 467.5$$

*Therefore,*

$$S = 467.5^{\frac{1}{2}} = 21.62$$

$$S_{kr} = \frac{1}{n} \frac{\sum_{i=1}^{n}(X_i - X)^4}{S^4} = \frac{1}{5} \frac{-20^4 + (-19^4) + 22^4 + 24^4 + (-7^4)}{21.62^4} = 0.7861$$

*Next, we subtract 3 from the sample kurtosis and get the excess kurtosis.*

*Thus, excess* kurtosis=0.7861–3=−2.2139excess kurtosis=0.7861–3=−2.2139

*Since the excess kurtosis is negative, we have a platykurtic distribution.*

Like skewness, kurtosis is a statistical measure that is used to describe distribution. Whereas skewness differentiates extreme values in one versus the other tail, kurtosis measures extreme values in either tail. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). Distributions with low kurtosis exhibit tail data that are generally less extreme than the tails of the normal distribution.

For investors, high kurtosis of the return distribution implies the investor will experience occasional extreme returns (either positive or negative), more extreme than the usual + or - three standard deviations from the mean that is predicted by the normal distribution of returns. This phenomenon is known as kurtosis risk.

Breaking Down Kurtosis

- kurtosis is a measure of the combined weight of a distribution's tails relative to the center of the distribution. When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within + or - three standard deviations of the mean. However, when high kurtosis is present, the tails extend farther than the + or - three standard deviations of the normal bell-curved distribution.

- Kurtosis is sometimes confused with a measure of the peakedness of a distribution. However, kurtosis is a measure that describes the shape of a distribution's tails in relation to its overall shape. A distribution can be infinitely peaked with low kurtosis, and a distribution can be perfectly flat-topped with infinite kurtosis. Thus, kurtosis measures "tailedness," not "peakedness."

## What Is Excess Kurtosis?

The term excess kurtosis refers to a metric used in statistics and probability theory comparing the kurtosis coefficient with that of a normal distribution. Kurtosis is a statistical measure that is used to describe the size of the tails on a distribution. Excess kurtosis helps determine how much risk is involved in a specific investment. It signals that the probability of obtaining an extreme outcome or value from the event in question is higher than would be found in a probabilistically normal distribution of outcomes.

## Example of Excess Kurtosis

Let's use a hypothetical example of excess kurtosis. If you track the closing value of stock ABC every day for a year, you will have a record of how often the stock closed at a given value. If you build a graph with the closing values along the X-axis and the number of instances of that closing value that occurred along the Y-axis of a graph, you will create a bell-shaped curve showing the distribution of the stock's closing values.

If there are a high number of occurrences for just a few closing prices, the graph will have a very slender and steep bell-shaped curve. If the closing values vary widely, the bell will have a wider shape with less steep sides. The tails of this bell will show you how often heavily deviated closing prices occurred, as graphs with lots of outliers will have thicker tails coming off each side of the bell.

## Is high kurtosis good or bad?

Kurtosis as a stand-alone metric is not very useful. Kurtosis is only useful when used in conjunction with standard deviation. It is possible that an investment might have a high kurtosis (bad), but the overall standard deviation is low (good). Conversely, one might see an investment with a low kurtosis (good), but the overall standard deviation is high (bad). Kurtosis gives a better understanding of standard deviation, but used in isolation, kurtosis is meaningless

Generally speaking, one would hope to see a low or negative kurtosis. A low or negative kurtosis means that on a period-by-period basis most observations fall within a predictable band. The risk that does occur happens within a moderate range, and there is little risk in the tails. Alternatively,

the higher the kurtosis, the more it indicates that the overall risk of an investment is driven by a few extreme "surprises" in the tails of the distribution.

## 5.8   Why Standard Deviation Is an Important Statistic

The standard deviation is a measure of the spread of scores within a set of data. Usually, we are interested in the standard deviation of a population. However, as we are often presented with data from a sample only, we can estimate the population standard deviation from a sample standard deviation.

These two standard deviations - sample and population standard deviations - are calculated differently.

**When to use the sample or population standard deviation**

We are normally interested in knowing the population standard deviation because our population contains all the values we are interested in. Therefore, you would normally calculate the population standard deviation if: (1) you have the entire population or (2) you have a sample of a larger population, but you are only interested in this sample and do not wish to generalize your findings to the population. However, in statistics, we are usually presented with a sample from which we wish to estimate (generalize to) a population, and the standard deviation is no exception to this. Therefore, if all you have is a sample, but you wish to make a statement about the population standard deviation from which the sample is drawn, you need to use the sample standard deviation. Confusion can often arise as to which standard deviation to use due to the name "sample" standard deviation incorrectly being interpreted as meaning the standard deviation of the sample itself and not the estimate of the population standard deviation based on the sample.

**What type of data should you use when you calculate a standard deviation?**

The standard deviation is used in conjunction with the mean to summarise continuous data, not categorical data. In addition, the standard deviation, like the mean, is normally only appropriate when the continuous data is not significantly skewed or has outliers.

**Examples of when to use the sample or population standard deviation**

Q. A teacher sets an exam for their pupils. The teacher wants to summarize the results the pupils attained as a mean and standard deviation. Which standard deviation should be used?

A. Population standard deviation. Why? Because the teacher is only interested in this class of pupils' scores and nobody else.

Q. A researcher has recruited males aged 45 to 65 years old for an exercise training study to investigate risk markers for heart disease (e.g., cholesterol). Which standard deviation would most likely be used?

A. Sample standard deviation. Although not explicitly stated, a researcher investigating health related issues will not simply be concerned with just the participants of their study; they will want to show how their sample results can be generalised to the whole population (in this case, males aged 45 to 65 years old). Hence, the use of the sample standard deviation.

Q. One of the questions on a national consensus survey asks for respondents' age. Which standard deviation would be used to describe the variation in all ages received from the consensus?

A. Population standard deviation. A national consensus is used to find out information about the nation's citizens. By definition, it includes the whole population. Therefore, a population standard deviation would be used.

What are the formulas for the standard deviation?

The **sample standard deviation formula** is:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$$

where,

s = sample standard deviation

$\sum$ = sum of...

$\bar{X}$ = sample mean

n = number of scores in sample.

The standard deviation is a commonly used statistic, but it doesn't often get the attention it deserves. Although the mean and median are out there in common sight in the everyday media, you rarely see them accompanied by any measure of how diverse that data set was, and so you are getting only part of the story. In fact, you could be missing the most interesting part of the story.

Without **calculating standard deviation**, you can't get a handle on whether the data are close to the average (as are the diameters of car parts that come off of a conveyor belt when everything is operating correctly) or whether the data are spread out over a wide range (as are house prices and income levels in the U.S.).

Without the standard deviation, you can't compare two data sets effectively. Suppose two sets of data have the same average; does that mean that the data sets must be exactly the same? Not at all. For example, the data sets 199, 200, 201 and 0, 200, 400 both have the same average (200) yet they have very different standard deviations. The first data set has a *very* small standard deviation (*s*=1) compared to the second data set (*s*=200).

## Summary

- Central Tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics.
- Change of origin and scale make calculations easy as, in origin change, the distribution location is changed, while in scale change, the distribution shape is altered.
- Any constant added or subtracted(Change of origin) than the standard deviation of original data and of change data after addition as subtracted will not change but the mean of new data will change.
- Any constant multiplied are divided (change of scale) then mean, standard deviation and variation will change of the new changed data

## Keywords

- The direction of skewness is given by the sign.
- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.
- A value of zero means no skewness at all.
- A large negative value means the distribution is negatively skewed.
- A large positive value means the distribution is positively skewed

## Self Assessment

1. The _____means that some value has been either added or subtracted to the observations

A. Change of origin

B. Change of scale

C. Change of Mean

D. None of the above

2. The _____means that some value has been either multiplied or divided to the observations.

A. Change of origin

B. Change of scale

C. Change of Mean

D. None of the above

3. _____ changes standard deviation and variance

A. Change of origin

B. Change of scale

C. Change of Mean

D. None of the above

4. _____ changes mean, median,mode

A. Change of origin

B. Change of scale

C. Change of Mean

D. None of the above

5. No matter what value we add to the set, the mean will shift by that amount but the _____will remain the same

A. Range

B. Mode

C. Median

D. All of the above

6. if we subtract an amount from every data point in the set: the mean will shift to the left but _____will stay the same

A. Interquartile range (IQR),

B. Mode

C. Median

D. All of the above

7. The _____ is the difference between the lowest and highest values.

A. Range
B. Mode
C. Median
D. All of the above

8. In descriptive statistics, the _____ also called the midspread, middle 50%, or H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles

A. Interquartile range (IQR),
B. Mode
C. Median
D. All of the above

9. Most widely used measure of variability is

A. Standard deviation
B. Mean
C. Mode
D. Median

10. Greatly influenced by sample size: the larger the sample, the larger the ____

A. Standard deviation
B. Mean
C. Mode
D. Range

11. A normal distribution (bell curve) exhibits_____ skewness.

A. Positive
B. Negative
C. Zero
D. Range

12. _____distribution  has more than one peak

A. Single modal
B. Multimodal
C. Skewness
D. All of the above

13. In probability theory and statistics, _____ is the expectation of the squared deviation of a random variable from its mean

A. Standard deviation

B. Variance

C. Mode

D. Median

14. _____ measures the degree of peakedness of a frequency distribution.

A. Kurtosis

B. Skewness

C. Mode

D. All of the above

15. The kurtosis reveals a distribution with flat tails

A. Platykurtic

B. Mesokurtic

C. Leptokurtic

D. All of the above

## Answers for Self Assessment

| 1. | A | 2. | B | 3. | B | 4. | B | 5. | A |
|----|---|----|---|----|---|----|---|----|---|
| 6. | A | 7. | A | 8. | A | 9. | A | 10. | D |
| 11. | C | 12. | B | 13. | B | 14. | A | 15. | A |

## Review Questions

1. What is effect of change of origin and scale on median?
2. What is difference between discrete and continuous data?
3. How Standard deviation is useful measure in statistics?
4. What are raw moments in statistics?
5. What are central moments in statistics?
6. What do you say whether high kurtosis good or bad?
7. What is effect of change of origin and scale on standard deviation?
8. How change of origin is different form change in scale
9. What do you do when your data is not normally distributed?
10. How do you interpret skewness and kurtosis?

### Further Readings

An Introduction to Probability and Statistics

Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

First Course in Probability, A

Book by Sheldon M. Ross

Schaums Theory and Problems of Statistics

Book by Murray R. Spiegel

Introduction to Probability, Statistics, and Random ...

Book by Hossein Pishro-Nik

**Web Links**

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability

*Dr. Pritpal Singh, Lovely Professional University*

# Unit 06: Correlation, Regression and Analysis if Variance

## Objectives

After this unit you will be able to:

- Learn the basic concept of correlation.

- Understand the different ways of measuring correlation.

- Understand the concept of Regression.

- Learn the difference between t test and Anova.

## Introduction

The goal of statistical data analysis is to understand a complex, real-world phenomenon from partial and uncertain observations. It is important to make the distinction between the mathematical theory underlying statistical data analysis, and the decisions made after conducting an analysis. Where there is a subjective part in the way statistical analysis yields actual human decisions. Understanding the risk and the uncertainty behind statistical results is critical in the decision-making process.

There are many terms that need introduction before we get started with the actual topics. These notions allow us to classify statistical techniques within multiple axes.

Prediction consists of learning from data, and predicting the outcomes of a random process based on a limited number of observations, the term "predictor" can be misleading if it is interpreted as the ability to predict even beyond the limits of the data. Also, the term "explanatory variable" might give an impression of a causal effect in a situation in which inferences should be limited to

identifying associations. The terms "independent" and "dependent" variable are less subject to these interpretations as they do not strongly imply cause and effect

Observations are independent realizations of the same random process; each observation is made of one or several variables. Mainly variables are either numbers, or elements belonging to a finite set "finite number of values". The first step in an analysis is to understand what your observations and variables are

Study is univariate if you have one variable. It is Bivariate if there are two variables and multivariate if at least two variables. Univariate methods are typically simpler. That being said, univariate methods may be used on multivariate data, using one dimension at a time. Although interactions between variables cannot be explored in that case, it is often an interesting first approach.

## 6.1   **What Are correlation and regression**

Correlation quantifies the degree and direction to which two variables are related. Correlation does not fit a line through the data points. But simply is computing a correlation coefficient that tells how much one variable tends to change when the other one does. When r is 0.0, there is no relationship. When r is positive, there is a trend that one variable goes up as the 2 other one goes up. When r is negative, there is a trend that one variable goes up as the other one goes down.

With correlation, it doesn't have to think about cause and effect. It doesn't matter which of the two variables is call dependent and which is call independent, if the two variables swapped the degree of correlation coefficient will be the same.

The sign (+, -) of the correlation coefficient indicates the direction of the association. The magnitude of the correlation coefficient indicates the strength of the association, e.g. A correlation of r = - 0.8 suggests a strong, negative association (reverse trend) between two variables, whereas a correlation of r = 0.4 suggest a weak, positive association. A correlation close to zero suggests no linear association between two continuous variables.

Linear regression finds the best line that predicts dependent variable from independent variable. The decision of which variable calls dependent and which calls independent is an important matter in regression, as it'll get a different best-fit line if you swap the two. The line that best predicts independent variable from dependent variable is not the same as the line that predicts dependent variable from independent variable in spite of both those lines have the same value for R2 . Linear regression quantifies goodness of fit with R2 , if the same data put into correlation matrix the square of r degree from correlation will equal R2 degree from regression. The sign (+, -) of the regression coefficient indicates the direction of the effect of independent variable(s) into dependent variable, where the degree of the regression coefficient indicates the effect of the each independent variable into dependent variable.

## 6.2   **Test of Significance level**

In linguistic, "significant" means important, while in Statistics "significant" means probably true (not due to chance). A research finding may be true without being important. When statisticians say a result is "highly significant" they mean it is very probably true. They do not (necessarily) mean it is highly important.

Significance levels show you how likely a pattern in your data is due to chance. The most common level, used to mean something is good enough to be believed, is "0.95". This means that the finding has a 95% chance of being true which also means that the finding has a confidence degree 95% of being true. No statistical package will show you "95%" or ".95" to indicate this level. Instead it will show you ".05," meaning that the finding has a five percent (.05) chance of not being true "error", which is the converse of a 95% chance of being true. To find the significance level, subtract the number shown from one. For example, a value of ".01" means that there is a confidence degree 99% (1-.01=.99) chance of it being true.

In other words the significance level "alpha level" for a given hypothesis test is a value for which a P-value "calculated value" less than or equal to is considered statistically significant. Typical value

levels for are 0.1, 0.05, and 0.01. These value levels correspond to the probability of observing such an extreme value by chance. For example, if the P-value is 0.0082, so the probability of observing such a value by chance is less that 0.01, and the result is significant at the 0.01 level.

**Correlation Analysis**

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other. However, correlation does not imply causation. There may be an unknown factor that influences both variables similarly.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data

Correlation is Positive or direct when the values increase together, and

Correlation is Negative when one value decreases as the other increases, and so called inverse or contrary correlation.

Task: How Level of signifance is important in statistics?



If the points plotted were all on a straight line we would have perfect correlation, but it could be positive or negative as shown in the diagrams above,

a. Strong positive correlation between x and y. The points lie close to a straight line with y increasing as x increases.

 b. Weak, positive correlation between x and y. The trend shown is that y increases as x increases but the points are not close to a straight line

 c. No correlation between x and y; the points are distributed randomly on the graph.

d. Weak, negative correlation between x and y. The trend shown is that y decreases as x increases but the points do not lie close to a straight line

e. Strong, negative correlation. The points lie close to a straight line, with y decreasing as x increases.

## 6.3 **Assumption of Correlation**

Employing of correlation rely on some underlying assumptions. The variables are assumed to be independent, assume that they have been randomly selected from the population; the two variables

*Probability and Statistics*

are normal distribution; association of data is homoscedastic (homogeneous), homoscedastic data have the same standard deviation in different groups where data are heteroscedastic have different standard deviations in different groups and assumes that the relationship between the two variables is linear. The correlation coefficient is not satisfactory and difficult to interpret the associations between the variables in case if data have outliers.

An inspection of a scatterplot can give an impression of whether two variables are related and the direction of their relationship. But it alone is not sufficient to determine whether there is an association between two variables. The relationship depicted in the scatterplot needs to be described qualitatively. Descriptive statistics that express the degree of relation between two variables are called correlation coefficients. A commonly employed correlation coefficient are Pearson correlation, Kendall rank correlation and Spearman correlation.

Correlation used to examine the presence of a linear relationship between two variables providing certain assumptions about the data are satisfied. The results of the analysis, however, need to be interpreted with care, particularly when looking for a causal relationship.

## 6.4   **Bivariate Correlation**

Bivariate correlation is a measure of the relationship between the two variables; it measures the strength and direction of their relationship, the strength can range from absolute value 1 to 0. The stronger the relationship, the closer the value is to 1. Direction of The relationship can be positive (direct) or negative (inverse or contrary); correlation generally describes the effect that two or more phenomena occur together and therefore they are linked For example, the positive relationship of .71 can represent positive correlation between the statistics degrees and the science degrees. The student who has high degree in statistics has also high degree in science and vice versa

The Pearson correlation coefficient is given by the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where $\bar{x}$ is the mean of variable $x$ values, and $\bar{y}$ is the mean of variable $y$ values.

**Types of Correlation**
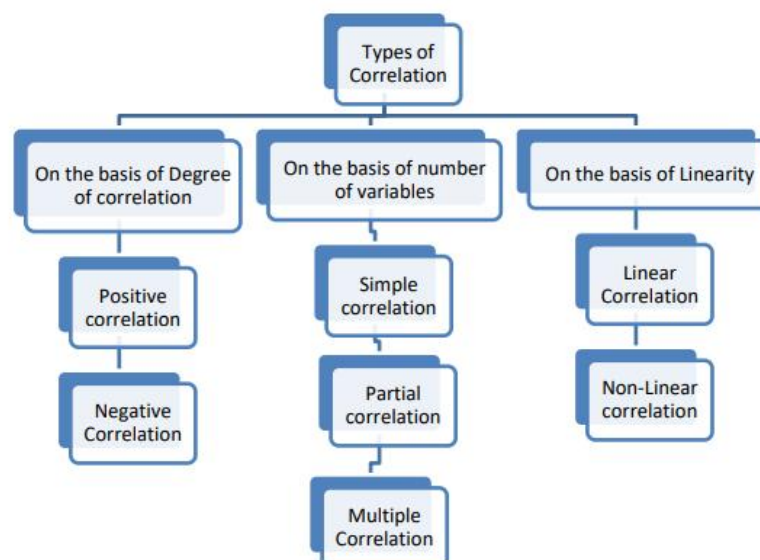
## 6.5 Methods of studying Correlation
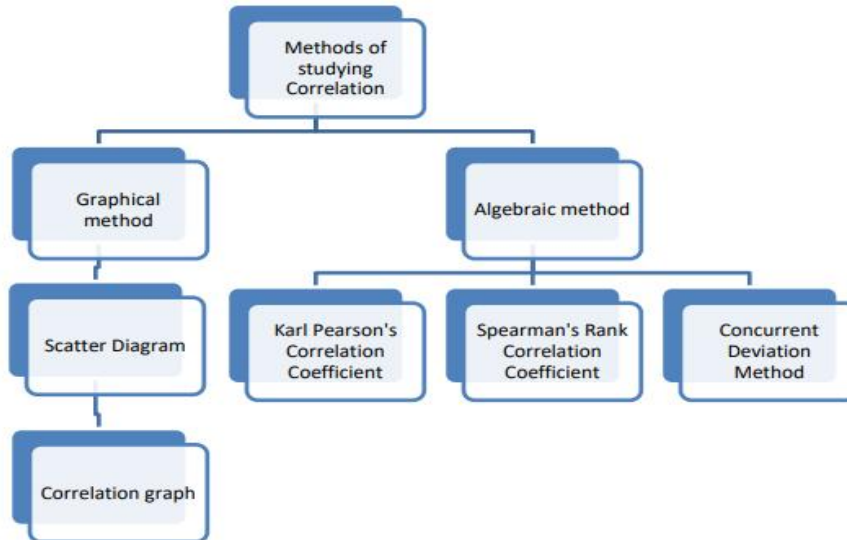


### Scatter Diagram

This is the simplest method of studying correlation between two variables. The two variables x and y are taken on the X and Y axes of a graph paper. Each pair of x and y value we mark a dot and we get as many points as the number of pairs of observation. By looking through the scatter of points, we can form an idea as whether the variables are related or not. If all the plotted points lie on a straight line rising from the lower left hand corner to the upper right hand corner, correlation is said to be perfectly positive. If all the plotted points lie on a straight line falling from the upper left hand corner to the lower right hand corner of the diagram, correlation is said to be perfectly negative. If all the plotted points fall in a narrow line and the points are rising from the lower left hand corner to the upper right hand corner of the diagram, there is degree of positive correlation between variables.

If the plotted points fall in a narrow bank and the points are lying from the upper left hand corner to the right hand corner, there high degree of negative correlation. If the plotted points lie scattered all over the diagram, there is no correlation between the two variables.



## SCATTERPLOTS & CORRELATION

Correlation - indicates a relationship (connection) between two sets of data.

**Strong positive correlation**

**Weak positive correlation**

**Strong negative correlation**

**Weak negative correlation**

**Moderate negative correlation**

**No correlation**

**Lovely Professional University**

*Probability and Statistics*
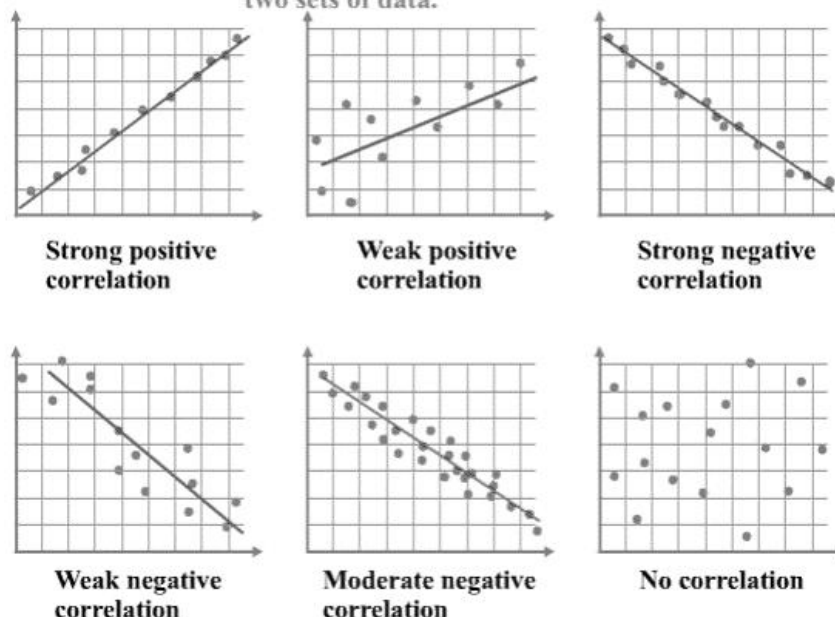
## KARL PEARSON'S CO-EFFICIENT OF CORRELATION

The Karl Pearson's product-moment correlation coefficient (or simply, the Pearson's correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by r or rxy(x and y being the two variables involved).This method of correlation attempts to draw a line of best fit through the data of two variables, and the value of the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.It is a mathematical method for measuring correlation between two variables and was suggested by Karl Pearson, a British Statistician. It is the most widely used method for measuring correlation. It is defined as:

$$r = \frac{Covariance\ (x, y)}{S.D.(x)S.D.(y)}$$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Interpretation of 'r'

The value of the coefficient of correlation will always lie between -1 and +1., i.e., $-1 \le r \le 1$. When r = +1, it means, there is perfect positive correlation between the variables. When r = -1, there is perfect negative correlation between the variables. When r = 0, there is no relationship between the two variables. The coefficient correlation describes not only the magnitude of correlation but also its direction. Thus, +0.8 indicates that correlation is positive because the sign of r is plus and the degree of correlation is high because the numerical value of r(0.8) is close to 1. If r = -0.4, it indicates that there is low degree of negative correlation because the sigh of r is negative e and the numerical value of r is less than 0.5.

Assumptions

While calculating the Pearson's Correlation Coefficient, we make the following assumptions – • There is a linear relationship (or any linear component of the relationship) between the two variables • We keep Outliers either to a minimum or remove them entirely

Properties of the Pearson's Correlation Coefficient

1. r lies between -1 and +1, or $-1 \le r \le 1$, or the numerical value of r cannot exceed one (unity

2. The correlation coefficient is independent of the change of origin and scale.

3. Two independent variables are uncorrelated but the converse is not true.

**Example 1: calculate correlation coefficient for the following data**:

| X | 2 | 4 | 5 | 6 | 8 | 11 |
|---|---|---|---|---|---|----|
| Y | 18 | 12 | 10 | 8 | 7 | 5 |

Solution:

| X | Y | X² | Y² | XY |
|---|---|---|---|---|
| 2 | 18 | 4 | 324 | 36 |
| 4 | 12 | 16 | 144 | 48 |
| 5 | 10 | 25 | 100 | 50 |
| 6 | 8 | 36 | 64 | 48 |
| 8 | 7 | 64 | 49 | 56 |
| 11 | 5 | 121 | 25 | 55 |
| $\sum X = 36$ | $\sum Y = 60$ | $\sum X^2 = 266$ | $\sum Y^2 = 706$ | $\sum(XY) = 293$ |

$$r = \frac{6 \times 293 - 36 \times 60}{\sqrt{6 \times 266 - 36^2}\ \sqrt{6 \times 706 - 60^2}}$$

$$= \frac{1758 - 2160}{\sqrt{1590 - 1296}\ \sqrt{4236 - 3600}}$$

$$= \frac{-402}{17.32 \times 25.22}$$

$$= \frac{-402}{436.81}$$

$$= -0.920$$

(Note: there is high degree of negative correlation)

**SPEARMAN'S RANK CORRELATION COEFFICIENT**

In 1904, C. Spearman introduced a new method of measuring the correlation between two variables. Instead of taking the values of the variables he considered the ranks (or order) of the observations and calculated Pearson's coefficient of correlation for the ranks. The correlation coefficient so obtained is called rank correlation coefficient. This measure is useful in dealing with qualitative characteristics such as intelligence, beauty, morality, honesty etc. The formula for spearman's rank correlation coefficient is:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

*Where, $r_s$ = spearman rank correlation coefficient*
*D = differences in ranks between paired items ($R_1 - R_2$)*
*N = number of pairs of observations*

Example: From the following data, calculate Spearman's rank correlation

| Rank        in | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Economics | | | | | | | | | | |
| Rank        in | 4 | 8 | 2 | 3 | 5 | 7 | 6 | 9 | 10 | 1 |
| Statistics | | | | | | | | | | |

**Solution:**

| $R_1$ | $R_2$ | d | $d^2$ | Steps for solution |
| --- | --- | --- | --- | --- |
| 1 | 4 | -3 | 9 | |
| 2 | 8 | -6 | 36 | $r = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$ |
| 3 | 2 | 1 | 1 | |
| 4 | 3 | 1 | 1 | $= 1 - \dfrac{6(132)}{10(100-1)}$ |
| 5 | 5 | 0 | 0 | |
| 6 | 7 | -1 | 1 | = 1 - 0.8 |
| 7 | 6 | 1 | 1 | = 0.2 |
| 8 | 9 | -1 | 1 | |
| 9 | 10 | -1 | 1 | |
| 10 | 1 | 9 | 81 | **Interpretation: The result indicates that there** |
| Total | -- | | $\sum d^2 = 132$ | **is low positive correlation** |

When ranks are not given:

| X | 17 | 13 | 15 | 16 | 6 | 11 | 14 | 9 | 7 | 12 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Y | 36 | 46 | 35 | 24 | 12 | 18 | 27 | 22 | 2 | 8 |

**Solution:**

| X | Y | Rank X $(R_1)$ | Rank Y $(R_2)$ | d $(R_1 - R_2)$ | $d^2$ | Solving steps |
| --- | --- | --- | --- | --- | --- | --- |
| 17 | 36 | 1 | 2 | -1 | 1 | |
| 13 | 46 | 5 | 1 | 4 | 16 | $r = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$ |
| 15 | 35 | 3 | 3 | 0 | 0 | |
| 16 | 24 | 2 | 5 | -3 | 9 | $= 1 - \dfrac{6(44)}{10(100-1)}$ |
| 6 | 12 | 10 | 8 | 2 | 4 | |
| 11 | 18 | 7 | 7 | 0 | 0 | = 1 - 0.267 |
| 14 | 27 | 4 | 4 | 0 | 0 | = 0.733 |
| 9 | 22 | 8 | 6 | 2 | 4 | |
| 7 | 2 | 9 | 10 | -1 | 1 | |
| 12 | 8 | 6 | 9 | -3 | 9 | **Note: correlation is highly** |
| | | | | | 44 | **positive** |

Task: What are basis to decide level of correlation?
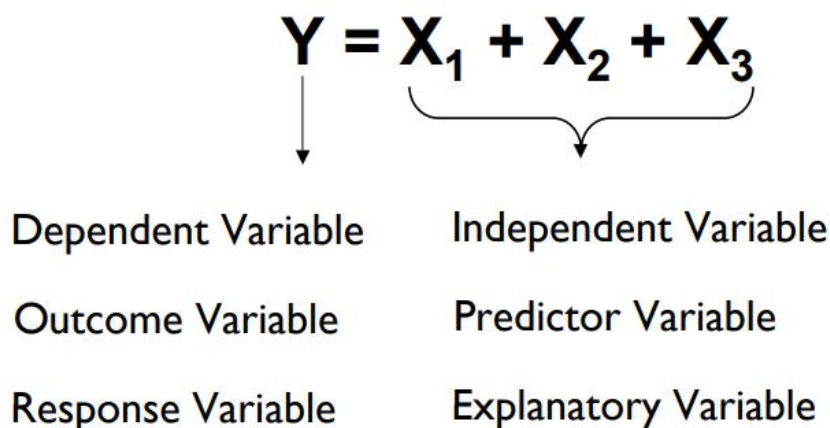
## 6.6 Regression analysis

Regression analysis, like most multivariate statistics, allows you to infer that there is a relationship between two or more variables. These relationships are seldom exact because there is variation caused by many variables, not just the variables being studied.

If you say that students who study more make better grades, you are really hypothesizing that there is a positive relationship between one variable, studying, and another variable, grades. You could then complete your inference and test your hypothesis by gathering a sample of (amount studied, grades) data from some students and use regression to see if the relationship in the sample is strong enough to safely infer that there is a relationship in the population. Notice that even if students who study more make better grades, the relationship in the population would not be perfect; the same amount of studying will not result in the same grades for every student (or for one student every time). Some students are taking harder courses, like chemistry or statistics; some are smarter; some study effectively; and some get lucky and find that the professor has asked them exactly what they understood best. For each level of amount studied, there will be a distribution of grades. If there is a relationship between studying and grades, the location of that distribution of grades will change in an orderly manner as you move from lower to higher levels of studying.

Regression analysis is one of the most used and most powerful multivariate statistical techniques for it infers the existence and form of a functional relationship in a population. Once you learn how to use regression, you will be able to estimate the parameters — the slope and intercept — of the function that links two or more variables. With that estimated function, you will be able to infer or forecast things like unit costs, interest rates, or sales over a wide range of conditions. Though the simplest regression techniques seem limited in their applications, statisticians have developed a number of variations on regression that greatly expand the usefulness of the technique. In this chapter, the basics will be discussed. Once again, the t-distribution and F-distribution will be used to test hypotheses.

Regression • Technique used for the modeling and analysis of numerical data • Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other

Regression can be used for prediction, estimation, hypothesis testing, and modeling causal relationships

$$Y = X_1 + X_2 + X_3$$

| Dependent Variable | Independent Variable |
| Outcome Variable | Predictor Variable |
| Response Variable | Explanatory Variable |

Regression analysis is a technique for using data to identify relationships among variables and use these relationships to make predictions. We will be studying linear regression, in which we assume that the outcome we are predicting depends linearly on the information used to make the

prediction. Linear dependence means constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns).

**What Is Regression?**

Regression analysis is a way of predicting future happenings between a dependent (target) and one or more independent variables (also known as a predictor). For example, it can be used to predict the relationship between reckless driving and the total number of road accidents caused by a driver, or, to use a business example, the effect on sales and spending a certain amount of money on advertising.

Regression is one of the most common models of machine learning. It differs from classification models because it estimates a numerical value, whereas classification models identify which category an observation belongs to.

The main uses of regression analysis are forecasting, time series modeling and finding the cause and effect relationship between variables.

Why Is It Important?

Regression has a wide range of real-life applications. It is essential for any machine learning problem that involves continuous numbers – this includes, but is not limited to, a host of examples, including:

· Financial forecasting (like house price estimates, or stock prices)

· Sales and promotions forecasting

· Testing automobiles

· Weather analysis and prediction

· Time series forecasting

As well as telling you whether a significant relationship exists between two or more variables, regression analysis can give specific details about that relationship. Specifically, it can estimate the strength of impact that multiple variables will have on a dependent variable. If you change the value of one variable (price, say), regression analysis should tell you what effect that will have on the dependent variable (sales).

Businesses can use regression analysis to test the effects of variables as measured on different scales. With it in your toolbox, you can assess the best set of variables to use when building predictive models, greatly increasing the accuracy of your forecasting.

Finally, regression analysis is the best way of solving regression problems in machine learning using data modeling. By plotting data points on a chart and running the best fit line through them, you can predict each data point's likelihood of error: the further away from the line they lie, the higher their error of prediction (this best fit line is also known as a regression line).

## 6.7 **What Are the Different Types of Regression?**

Regression Analysis – Linear Model Assumptions

Linear regression analysis is based on six fundamental assumptions:

The dependent and independent variables show a linear relationship between the slope and the intercept.

The independent variable is not random.

The value of the residual (error) is zero.

The value of the residual (error) is constant across all observations.

The value of the residual (error) is not correlated across all observations.

The residual (error) values follow the normal distribution.

## Regression Analysis – Simple Linear Regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \epsilon$$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope

## Regression Analysis – Multiple Linear Regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + cX_2 + dX_3 + \epsilon$$

Where:

- **Y** – Dependent variable
- **X_1, X_2, X_3** – Independent (explanatory) variables
- **a** – Intercept
- **b, c, d** – Slopes
- **$\epsilon$** – Residual (error)

The analysis is also used to forecast the returns of securities, based on different factors, or to forecast the performance of a business. Learn more forecasting methods in CFI's Budgeting and Forecasting Course!

1. Beta and CAPM

In finance, regression analysis is used to calculate the Beta (volatility of returns relative to the overall market) for a stock. It can be done in Excel using the Slope function.

**Lovely Professional University**

*Probability and Statistics*

**Beta (β) Calculator**

| Individual Stock | | | S&P 500 Index | | |
|---|---|---|---|---|---|
| Date | Price | Return | Date | Price | Return |
| 1/2/2018 | 15.78 | | 1/2/2018 | 2,696 | |
| 1/9/2018 | 16.38 | 3.8% | 1/9/2018 | 2,751 | 2.0% |
| 1/16/2018 | 16.67 | 1.8% | 1/16/2018 | 2,776 | 0.9% |
| 1/23/2018 | 17.17 | 3.0% | 1/23/2018 | 2,839 | 2.3% |
| 1/30/2018 | 17.02 | -0.9% | 1/30/2018 | 2,822 | -0.6% |
| 2/6/2018 | 16.31 | -4.2% | 2/6/2018 | 2,695 | -4.5% |
| 2/13/2018 | 16.00 | -1.9% | 2/13/2018 | 2,663 | -1.2% |
| 2/20/2018 | 16.43 | 2.7% | 2/20/2018 | 2,716 | 2.0% |
| 2/27/2018 | 16.97 | 3.3% | 2/27/2018 | 2,765 | 1.8% |

| Beta (β) | 1.21 |
|---|---|



Beta Chart
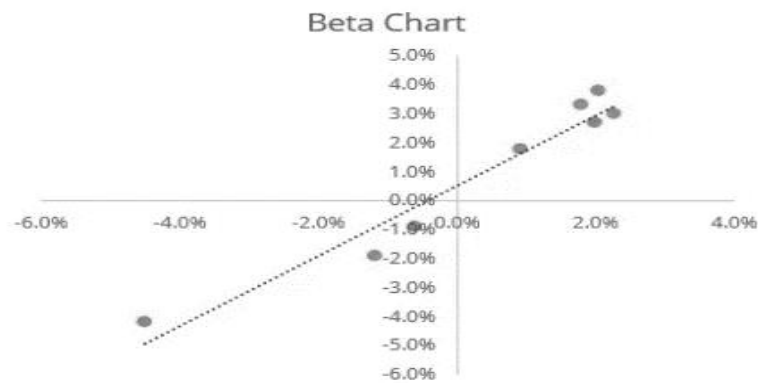
Forecasting Revenues and Expenses

When forecasting financial statements for a company, it may be useful to do a multiple regression analysis to determine how changes in certain assumptions or drivers of the business will impact revenue or expenses in the future. For example, there may be a very high correlation between the number of salespeople employed by a company, the number of stores they operate, and the revenue the business generates.

| Data | Radio ads | Revenue |
|---|---|---|
| Jan | 21 | $8,350.0 |
| Feb | 180 | $22,755.0 |
| Mar | 50 | $13,455.0 |
| Apr | 195 | $21,100.0 |
| May | 96 | $15,000.0 |
| Jun | 44 | $12,500.0 |
| Jul | 171 | $20,700.0 |
| Aug | 135 | $19,722.0 |
| Sep | 120 | $16,115.0 |
| Oct | 75 | $13,100.0 |
| Nov | 106 | $15,670.0 |
| Dec | 198 | $25,300.0 |
| Totals | 1,391 | $203,767.0 |
| Average | 116 | $16,980.6 |



Relationship between ads and revenue

$y = 78.075x + 7930.4$
$R^2 = 0.9327$
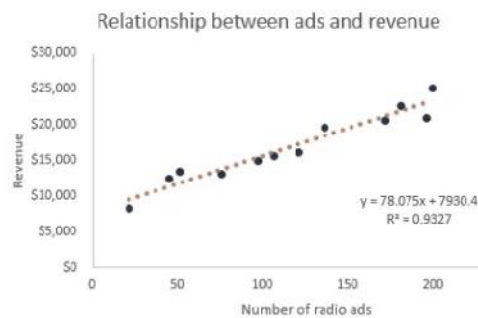
## 6.8 Output of Linear Regression Analysis

Statistics will generate quite a few tables of output for a linear regression. In this section, we show you only the three main tables required to understand your results from the linear regression procedure, assuming that no assumptions have been violated. A complete explanation of the

output you have to interpret when checking your data for the six assumptions required to carry out linear regression is provided in our enhanced guide. This includes relevant scatterplots, histogram (with superimposed normal curve), Normal P-P Plot, case wise diagnostics and the Durbin-Watson statistic. Below, we focus on the results for the linear regression analysis only.

The first table of interest is the **Model Summary** table, as shown below:

## Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .873[a] | .762 | .749 | 874.779 |

a. Predictors: (Constant), Income

This table provides the $R$ and $R^2$ values. The $R$ value represents the simple correlation and is 0.873 (the "**R**" Column), which indicates a high degree of correlation. The $R^2$ value (the "**R Square**" column) indicates how much of the total variation in the dependent variable, Price , can be explained by the independent variable, Income . In this case, 76.2% can be explained, which is very large.

The next table is the **ANOVA** table, which reports how well the regression equation fits the data (i.e., predicts the dependent variable) and is shown below:

### ANOVA[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 44182633.37 | 1 | 44182633.37 | 57.737 | .000[b] |
| | Residual | 13774291.07 | 18 | 765238.393 | | |
| | Total | 57956924.44 | 19 | | | |

a. Dependent Variable: Price

b. Predictors: (Constant), Income

This table indicates that the regression model predicts the dependent variable significantly well. How do we know this? Look at the "**Regression**" row and go to the "**Sig.**" column. This indicates the statistical significance of the regression model that was run. Here, $p < 0.0005$, which is less than 0.05, and indicates that, overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

The **Coefficients** table provides us with the necessary information to predict price from income, as well as determine whether income contributes statistically significantly to the model (by looking at the "**Sig.**" column). Furthermore, we can use the values in the "**B**" column under the "**Unstandardized Coefficients**" column, as shown below:

### Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 8286.786 | 1852.256 | | 4.474 | .000 |
| | Income | .564 | .074 | .873 | 7.598 | .000 |

a. Dependent Variable: Price

to present the regression equation as:

$$Price = 8287 + 0.564(Income)$$

*Probability and Statistics*

## 6.9 <u>Analysis of variance (ANOVA</u>

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyse the differences among means. ANOVA was developed by the statistician Ronald Fisher. ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two mean.

What Does the Analysis of Variance Reveal?

The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

Example of How to Use ANOVA

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

One-Way ANOVA versus Two-Way ANOVA

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

## T-TEST & ANOVA (ANALYSIS OF VARIANCE)

What are they? The t-test is a method that determines whether two populations are statistically different from each other, whereas ANOVA determines whether three or more populations are statistically different from each other. Both of them look at the difference in means and the spread of the distributions (i.e., variance) across groups; however, the ways that they determine the statistical significance are different.



ANOVA (F-test):
$$H_0 : \mu_1 = \mu_2 = ...\mu_t$$
$$df_1 = t-1$$
$$df_2 = N-t$$

One sample T-test:
$$H_0 : \mu = c$$
$$df = n-1$$

Paired sample T-test:
$$H_0 : \mu_d = 0$$
$$df = n-1$$

Independent sample T-test (Pooled variance):
$$H_0 : \mu_1 - \mu_2 = 0$$
$$df = n_1 + n_2 - 2$$

Independent sample T-test (Approximation of d.f.):
$$H_0 : \mu_1 - \mu_2 = 0$$
$$df = approximated$$
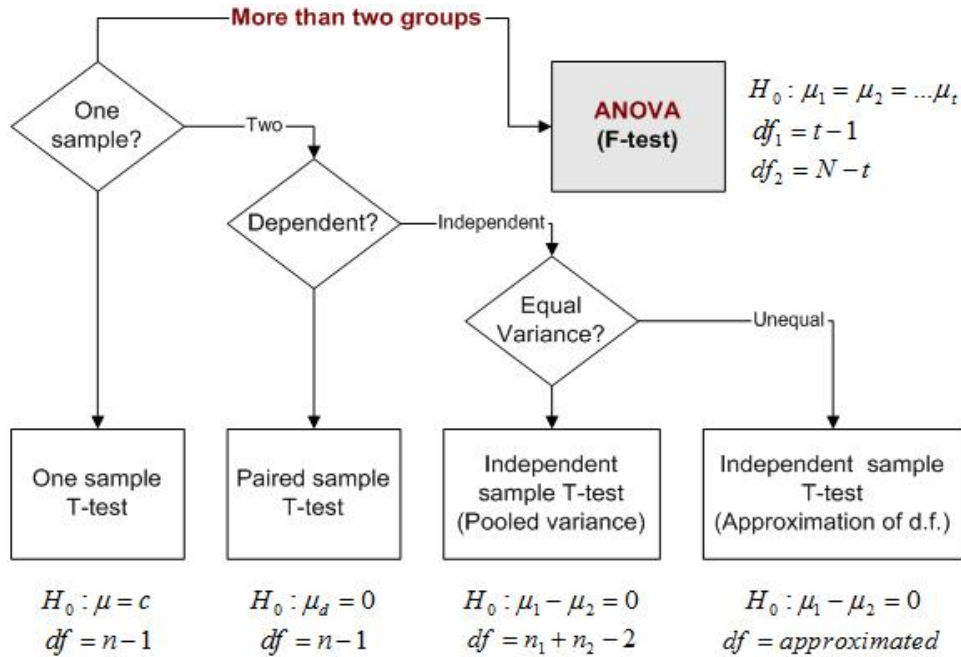
## Summary

Correlation is **a statistical measure that determines the association or co-relationship between two variables**

Regression describes how to numerically relate an independent variable to the dependent variable

Regression indicates the impact of a change of unit on the estimated variable ( y) in the known variable (x)

Analysis of variance (**ANOVA**) is a collection of statistical models and their associated estimation procedures used to analyze the differences among means

A **t-test** is a type of inferential statistic used to determine if there is a significant difference between the means of two groups

## Keywords

The t-test is a **method that determines whether two populations are statistically different from each other**, whereas ANOVA determines whether three or more populations are statistically different from each other

The main difference in correlation vs regression is that the **measures of the degree of a relationship between two variables**

If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

*Probability and Statistics*

## Self Assessment

1. Correlation analysis is a.............................
   A. Univar ate analysis
   B. Bivariate analysis
   C. Multivariate analysis
   D. Both b and c

2. If change in one variable results a corresponding change in the other variable, then the variables are.........................
   A. Correlated
   B. Not correlated
   C. Any of the above
   D. None of the above

3. When the values of two variables move in the same direction, correlation is said to be .............................
   A. Linear
   B. Non-linear
   C. Positive
   D. Negative

4. When the values of two variables move in the opposite directions, correlation is said to be .............................
   A. Linear
   B. Non-linear
   C. Positive
   D. Negative

5. When the amount of change in one variable leads to a constant ratio of change in the other variable, then correlation is said to be .........................
   A. Linear
   B. Non-linear
   C. Positive
   D. Negative

6. ...........................attempts to determine the degree of relationship between variables.
   A. Regression analysis
   B. Correlation analysis
   C. Inferential analysis
   D. None of these

7. Non-linear correlation is also called.....................................
   A. Non-curvy linear correlation

B. Curvy linear correlation

C. Zero correlation

D. None of these


8.  If all the points of a scatter diagram lie on a straight line falling from left upper corner to the right bottom corner, the correlation is called...................

A. Zero correlation

B. High degree of positive correlation

C. Perfect negative correlation

D. Perfect positive correlation


9. The variable whose value is influenced or is to be predicted is called

A. Dependent variable.

B. Independent variable

C. Both of these

D. None of these


10. The variable which influences  the values or is used for prediction is called

A. Dependent variable.

B. Independent variable

C. Both of these

D. None of these


11. In this equation.. $Y = \beta_0 + \beta_1 X$…………..Y is

A. Dependent variable.

B. Independent variable

C. Both of these

D. None of these


12. In this equation.. $Y = \beta_0 + \beta_1 X$…………..X  is

A. Dependent variable.

B. Independent variable

C. Both of these

D. None of these


13. Statistical technique specially designed to test whether the means of more than 2 quantitative populations are equal.

A. Anova

B. Coorelation

C. Regression

D. None of these

**Lovely Professional University**

14. The _____in ANOVA is valid when all the sample means are equal, or they don't have any significant difference.

A. Null hypothesis
B. Alternate hypothesis
C. Both of these
D. None of these

15. _____is a function that allows an analyst to make predictions about one variable based on the information that is known about another one  variable

A. Simple linear regression
B. Multiple linear regression
C. Multilinker
D. None of these

16. _____is a function that allows an analyst to make predictions about one variable based on the information of another three variables.

A. Simple linear regression
B. Multiple linear regression
C. Multilinker
D. None of these

## Answers for Self Assessment

| 1. | D | 2. | A | 3. | C | 4. | D | 5. | A |
|----|---|----|---|----|---|----|---|----|---|
| 6. | C | 7. | B | 8. | C | 9. | A | 10. | B |
| 11. | A | 12. | B | 13. | A | 14. | A | 15. | A |
| 16. | B | | | | | | | | |

## Review Questions

1.  Why Correlation is called as measure of the linear relationship between two quantitative variables?
2.  What is correlation and regression with example?
3.  What types of Research issue can regression analysis answer?
4.  What is simple regression and correlation?
5.  How Anova is different form T test and what are different conditions to use T test and Anova?
6.  What is regression in statistics with example?
7.  How do you write a regression question?
8.  What is difference between Simple linear Regression and Multiple linear regression?
9.  What are different methods to measure correlation and regression?

## Further Readings

- An Introduction to Probability and Statistics

- Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

- First Course in Probability, A

- Book by Sheldon M. Ross

- Schaums Theory and Problems of Statistics

- Book by Murray R. Spiegel

- Introduction to Probability, Statistics, and Random ...

- Book by Hossein Pishro-Nik

## Web Links

- https://www.tutorialspoint.com
- www.webopedia.com
- https://www.britannica.com/science/probability

# Unit 07: Standard Distribution

## Objectives

- Understand basics of Probability distribution
- Learn concepts of Binomial distribution.
- Define basic terms of Normal distribution.
- Understand concept of standard deviation in statistics.
- Solve basic questions related to probability distributions.

## Introduction

Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or a set of any entities. It is a part of probability and statistics.

Random experiments are defined as the result of an experiment, whose outcome cannot be predicted. Suppose, if we toss a coin, we cannot predict, what outcome it will appear either it will come as Head or as Tail. The possible result of a random experiment is called an outcome. And the set of outcomes is called a sample point. With the help of these experiments or events, we can always create a probability pattern table in terms of variables and probabilities.

## 7.1    Probability Distribution of Random Variables

A random variable has a probability distribution, which defines the probability of its unknown values. Random variables can be discrete (not constant) or continuous or both. That means it takes any of a designated finite or countable list of values, provided with a probability mass function feature of the random variable's probability distribution or can take any numerical value in an interval or set of intervals. Through a probability density function that is representative of the random variable's probability distribution or it can be a combination of both discrete and continuous.

Two random variables with equal probability distribution can yet vary with respect to their relationships with other random variables or whether they are independent of these. The recognition of a random variable, which means, the outcomes of randomly choosing values as per the variable's probability distribution function, are called **random variables.**

Probability Distribution Formulas

## Probability Distribution Formulas

| | |
|---|---|
| Binomial Distribution | $P(X) = {}^nC_x a^x b^{n-x}$<br>Where a = probability of success<br>b=probability of failure<br>n= number of trials<br>x=random variable denoting success |
| Cumulative Distribution Function | $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ |
| Discrete Probability Distribution | $P(x) = \frac{n!}{r!(n-r)!} \cdot p^r (1-p)^{n-r}$<br>$P(x) = C(n,r) \cdot p^r (1-p)^{n-r}$ |

## Types of Probability Distribution

There are two types of probability distribution which are used for different purposes and various types of the data generation process.

Normal or Cumulative Probability Distribution

Binomial or Discrete Probability Distribution

Let us discuss now both the types along with their definition, formula and examples.

Cumulative Probability Distribution

The cumulative probability distribution is also known as a continuous probability distribution. In this distribution, the set of possible outcomes can take on values in a continuous range.

For example, a set of real numbers, is a continuous or normal distribution, as it gives all the possible outcomes of real numbers. Similarly, a set of complex numbers, a set of prime numbers, a set of whole numbers etc. are examples of Normal Probability distribution. Also, in real-life scenarios, the temperature of the day is an example of continuous probability. Based on these outcomes we can create a distribution table. A probability density function describes it. The formula for the normal distribution is;

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

Where,

μ = Mean Value

σ = Standard Distribution of probability.

If mean (μ) = 0 and standard deviation (σ) = 1, then this distribution is known to be normal distribution.

x = Normal random variable

Normal Distribution Examples

Since the normal distribution statistics estimates many natural events so well, it has evolved into a standard of recommendation for many probability queries. Some of the examples are:

Height of the Population of the world

Rolling a dice (once or multiple times)

To judge the Intelligent Quotient Level of children in this competitive world

Tossing a coin

Income distribution in countries economy among poor and rich

The sizes of female's shoes

Weight of newly born babies range

Average report of Students based on their performance

Discrete Probability Distribution

A distribution is called a discrete probability distribution, where the set of outcomes are discrete in nature.

For example, if a dice is rolled, then all the possible outcomes are discrete and give a mass of outcomes. It is also known as the probability mass function.

So, the outcomes of binomial distribution consist of n repeated trials and the outcome may or may not occur. The formula for the binomial distribution is;

Where,

n = Total number of events

r = Total number of successful events.

p = Success on a single trial probability.

$^nC_r = [n!/r!(n-r)]!$

1 – p = Failure Probability

## 7.2 <u>Probability Distribution Function</u>

A function which is used to define the distribution of a probability is called a Probability distribution function. Depending upon the types, we can define these functions. Also, these functions are used in terms of probability density functions for any given random variable.

In the case of **Normal distribution**, the function of a real-valued random variable X is the function given by;

$F_X(x) = P(X \le x)$

Where P shows the probability that the random variable X occurs on less than or equal to the value of x.
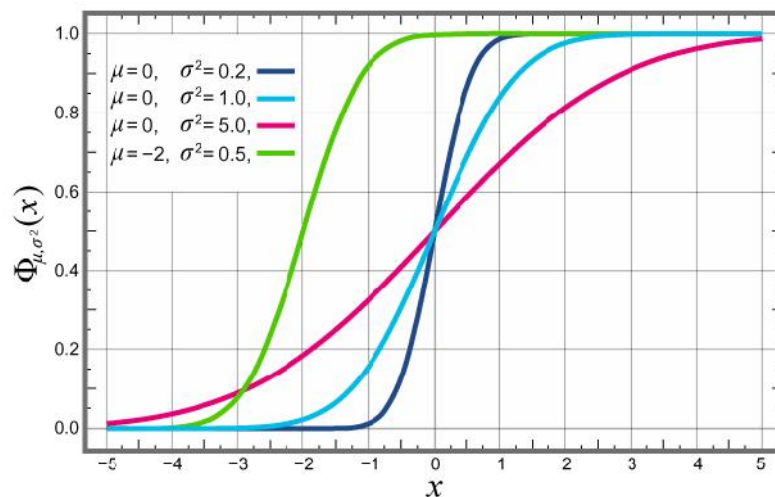
For a closed interval, (a→b), the cumulative probability function can be defined as;

**P (a<X ≤ b) = F$_X$ (b) – F$_X$ (a)**

If we express, the cumulative probability function as integral of its probability density function FX , then,

$$F_x(x) \ = \ \int_{-\infty}^{x} f_x(t)dt$$

In the case of a random variable X=b, we can define cumulative probability function as;

$$P(X=b) = F_x(b) - \lim_{x \to b} f_x(t)$$

**In the case of Binomial distribution**, as we know it is defined as the probability of mass or discrete random variable gives exactly some value. This distribution is also called probability mass distribution and the function associated with it is called a probability mass function.

Probability mass function is basically defined for scalar or multivariate random variables whose domain is variant or discrete. Let us discuss its formula:

Suppose a random variable X and sample space S is defined as;

**X: S → A**

And A ∈ R, where R is a discrete random variable.

Then the probability mass function f$_X$ : A → [0,1] for X can be defined as;

**FX(x) = P$_r$ (X=x) = P ({s ∈S : X(s) = x})**

**What is the Prior Probability?**

**Lovely Professional University**

In Bayesian statistical conclusion, a prior probability distribution, also known as the prior, of an unpredictable quantity is the probability distribution, expressing one's faiths about this quantity before any proof is taken into the record. For instance, the prior probability distribution represents the relative proportions of voters who will vote for some politician in a forthcoming election. The hidden quantity may be a parameter of the design or a possible variable rather than a perceptible variable.

**What is Posterior Probability?**

The posterior probability is the likelihood an event will occur after all data or background information has been brought into account. It is nearly associated with a prior probability, where an event will occur before you take any new data or evidence into consideration. It is an adjustment of prior probability. We can calculate it by using the below formula:

**Posterior Probability = Prior Probability + New Evidence**

It is commonly used in Bayesian hypothesis testing. For instance, old data propose that around 60% of students who begin college will graduate within 4 years. This is the prior probability. Still, if we think the figure is much lower, so we start collecting new data. The data collected implies that the true figure is closer to 50%, which is the posterior probability.

**Example 1**: A coin is tossed twice. X is the random variable of the number of heads obtained. What is the probability distribution of x?

**Solution:**

First write, the value of X= 0, 1 and 2, as the possibility are there that

No head comes

One head and one tail comes

And head comes in both the coins

Now the probability distribution could be written as;

P(X=0) = P(Tail+Tail) = ½ * ½ = ¼

P(X=1) = P(Head+Tail) or P(Tail+Head) = ½ * ½ + ½ *½ = ½

P(X=2) = P(Head+Head) = ½ * ½ = ¼

We can put these values in tabular form;

| X | 0 | 1 | 2 |
|---|---|---|---|
| P(X) | 1/4 | 1/2 | 1/4 |

## 7.3  Binomial Distribution

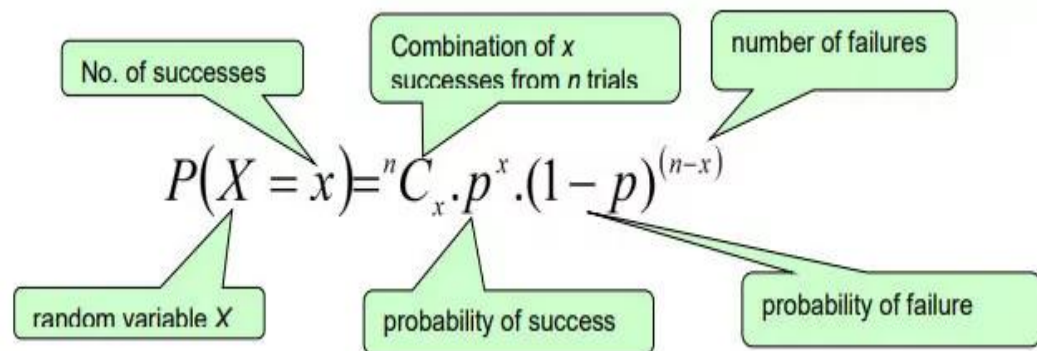In probability theory and statistics, the **binomial distribution** is the discrete probability distribution that gives only two possible results in an experiment, either **Success or Failure**. For example, if we toss a coin, there could be only two possible outcomes: heads or tails, and if any test is taken, then there could be only two results: pass or fail. This distribution is also called a binomial probability distribution.

There are two parameters n and p used here in a binomial distribution. The variable 'n' states the number of times the experiment runs and the variable 'p' tells the probability of any one outcome. Suppose a die is thrown randomly 10 times, then the probability of getting 2 for anyone throw is ⅙. When you throw the dice 10 times, you have a binomial distribution of n = 10 and p = ⅙. Learn the formula to calculate the two outcome distribution among multiple experiments along with solved examples here in this article

## Binomial Probability Distribution

In binomial probability distribution, the number of 'Success' in a sequence of n experiments, where each time a question is asked for yes-no, then the boolean-valued outcome is represented either with success/yes/true/one (probability p) or failure/no/false/zero (probability q = 1 − p). A single success/failure test is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a **Bernoulli process**. For n = 1, i.e. a single experiment, the binomial distribution is a **Bernoulli distribution**.

The binomial distribution is the base for the famous binomial test of statistical importance.



$$P(X = x) = {}^{n}C_{x} \cdot p^{x} \cdot (1 - p)^{(n-x)}$$

### Negative Binomial Distribution

In probability theory and statistics, the number of successes in a series of independent and identically distributed Bernoulli trials before a particularized number of failures happens. It is termed as the negative binomial distribution. Here the number of failures is denoted by 'r'. For instance, if we throw a dice and determine the occurrence of 1 as a failure and all non-1's as successes. Now, if we throw a dice frequently until 1 appears the third time, i.e., r = three failures, then the probability distribution of the number of non-1s that arrived would be the negative binomial distribution.

### Binomial Distribution Examples

As we already know, binomial distribution gives the possibility of a different set of outcomes. In real life, the concept is used for:

Finding the quantity of raw and used materials while making a product.

Taking a survey of positive and negative reviews from the public for any specific product or place.

By using the YES/ NO survey, we can check whether the number of persons views the particular channel.

To find the number of male and female employees in an organisation.

The number of votes collected by a candidate in an election is counted based on 0 or 1 probability.

## Binomial Distribution Formula

The binomial distribution formula is for any random variable X, given by;

$$P(x:n,p) = {}^nC_x \, p^x \, (1-p)^{n-x}$$

Or

$$P(x:n,p) = {}^nC_x \, p^x \, (q)^{n-x}$$

Where,

n = the number of experiments

x = 0, 1, 2, 3, 4, ...

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 − p

The binomial distribution formula can also be written in the form of n-Bernoulli trials, where ${}^nC_x = n!/x!(n-x)!$. Hence,

$$P(x:n,p) = n!/[x!(n-x)!].p^x.(q)^{n-x}$$

### Binomial Distribution Mean and Variance

For a binomial distribution, the mean, variance and standard deviation for the given number of success are represented using the formulas

Mean, $\mu = np$

Variance, $\sigma^2 = npq$

Standard Deviation $\sigma = \sqrt{(npq)}$

Where p is the probability of success

q is the probability of failure,

Where q = 1-p

Binomial Distribution Vs Normal Distribution

The main difference between the binomial distribution and the normal distribution is that binomial distribution is discrete, whereas the normal distribution is continuous. It means that the binomial distribution has a finite amount of events, whereas the normal distribution has an infinite number of events. In case, if the sample size for the binomial distribution is very large, then the distribution curve for the binomial distribution is similar to the normal distribution curve.

Properties of Binomial Distribution

The properties of the binomial distribution are:

There are two possible outcomes: true or false, success or failure, yes or no.

There is 'n' number of independent trials or a fixed number of n times repeated trials.

The probability of success or failure varies for each trial.

Only the number of success is calculated out of n independent trials.

Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

Binomial Distribution Examples And Solutions

*Probability and Statistics*

**Example 1: If a coin is tossed 5 times, find the probability of:**

**(a) Exactly 2 heads**

**(b) At least 4 heads.**

**Solution:**

**(a)** The repeated tossing of the coin is an example of a Bernoulli trial. According to the problem:

Number of trials: n=5

Probability of head: p= 1/2 and hence the probability of tail, q =1/2

For exactly two heads:

x=2

$P(x=2) = {}^5C_2\, p^2\, q^{5-2} = 5!\ /\ 2!\ 3! \times (\tfrac{1}{2})^2 \times (\tfrac{1}{2})^3$

$P(x=2) = 5/16$

**(b)** For at least four heads,

$x \geq 4$, $P(x \geq 4) = P(x = 4) + P(x=5)$

Hence,

$P(x = 4) = {}^5C_4\, p^4\, q^{5-4} = 5!/4!\ 1! \times (\tfrac{1}{2})^4 \times (\tfrac{1}{2})^1 = 5/32$

$P(x = 5) = {}^5C_5\, p^5\, q^{5-5} = (\tfrac{1}{2})^5 = 1/32$

Therefore,

$P(x \geq 4) = 5/32 + 1/32 = 6/32 = 3/16$

**Example 2: For the same question given above, find the probability of:**

a) **Getting at least 2 heads**

Solution: P (at most 2 heads) = $P(X \leq 2) = P(X = 0) + P(X = 1)$

$P(X = 0) = (\tfrac{1}{2})^5 = 1/32$

$P(X=1) = {}^5C_1 (\tfrac{1}{2})^{5.} = 5/32$

Therefore,

$P(X \leq 2) = 1/32 + 5/32 = 3/16$

**Example 3:** A fair coin is tossed 10 times, what are the probability of getting exactly 6 heads and at least six heads.

**Solution:**

Let x denote the number of heads in an experiment.

Here, the number of times the coin tossed is 10. Hence, n=10.

The probability of getting head, p ½

The probability of getting a tail, q = 1-p = 1-(½) = ½.

The binomial distribution is given by the formula:

$P(X= x) = {}^nC_x p^x q^{n-x}$, where = 0, 1, 2, 3, …

Therefore, $P(X = x) = {}^{10}C_x(\tfrac{1}{2})^x(\tfrac{1}{2})^{10-x}$

(i) The probability of getting exactly 6 heads is:

$P(X=6) = {}^{10}C_6(\tfrac{1}{2})^6(\tfrac{1}{2})^{10-6}$

P(X= 6) = $^{10}C_6(½)^{10}$

P(X = 6) = 105/512.

Hence, the probability of getting exactly 6 heads is 105/512.

(ii) The probability of getting at least 6 heads is P(X ≥ 6)

P(X ≥ 6) = P(X=6) + P(X=7) + P(X= 8) + P(X = 9) + P(X=10)

P(X ≥ 6) = $^{10}C_6(½)^{10}$ + $^{10}C_7(½)^{10}$ + $^{10}C_8(½)^{10}$ + $^{10}C_9(½)^{10}$ + $^{10}C_{10}(½)^{10}$

P(X ≥ 6) = 193/512.

## 7.4   Poisson Distribution

In Statistics, Poisson distribution is one of the important topics. It is used for calculating the possibilities for an event with the average rate of value. Poisson distribution is a discrete probability distribution. In this article, we are going to discuss the definition, Poisson distribution formula, table, mean and variance, and examples in detail.

Poisson Distribution Definition

The Poisson distribution is a discrete probability function that means the variable can only take specific values in a given list of numbers, probably infinite. A Poisson distribution measures how many times an event is likely to occur within "x" period of time. In other words, we can define it as the probability distribution that results from the Poisson experiment. A Poisson experiment is a statistical experiment that classifies the experiment into two categories, such as success or failure. Poisson distribution is a limiting process of the binomial distribution.

A Poisson random variable "x" defines the number of successes in the experiment. This distribution occurs when there are events that do not occur as the outcomes of a definite number of outcomes. Poisson distribution is used under certain conditions. They are:

The number of trials "n" tends to infinity

Probability of success "p" tends to zero

np = 1 is finite

Poisson distribution Formula

The formula for the Poisson distribution function is given by:

**f(x) =(e$^{-λ}$ λ$^x$)/x!**

Where,

e is the base of the logarithm

x is a Poisson random variable

λ is an average rate of value

Poisson distribution Table

As with the binomial distribution, there is a table that we can use under certain conditions that will make calculating probabilities a little easier when using the Poisson distribution. The table is showing the values of f(x) = P(X ≥ x), where X has a Poisson distribution with parameter λ. Refer the values from the table and substitute it in the Poisson distribution formula to get the probability value. The table displays the values of the Poisson distribution.

## Poisson Distribution Table

| $\lambda =$ | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| X=0 | 0.6065 | 0.3679 | 0.2231 | 0.1353 | 0.0821 | 0.0498 | 0.0302 | 0.0183 | 0.0111 | 0.0067 |
| 1 | 0.9098 | 0.7358 | 0.5578 | 0.4060 | 0.2873 | 0.1991 | 0.1359 | 0.0916 | 0.0611 | 0.0404 |
| 2 | 0.9856 | 0.9197 | 0.9197 | 0.8088 | 0.6767 | 0.5438 | 0.4232 | 0.3208 | 0.2381 | 0.1247 |
| 3 | 0.9982 | 0.9810 | 0.9344 | 0.8571 | 0.7576 | 0.6472 | 0.5366 | 0.4335 | 0.3423 | 0.2650 |
| 4 | 0.9998 | 0.9963 | 0.9814 | 0.9473 | 0.8912 | 0.8153 | 0.7254 | 0.6288 | 0.5321 | 0.4405 |
| 5 | 1.0000 | 0.9994 | 0.9994 | 0.9955 | 0.9834 | 0.9161 | 0.8576 | 0.7851 | 0.7029 | 0.6160 |
| 6 | 1.0000 | 0.9999 | 0.9991 | 0.9955 | 0.9858 | 0.9665 | 0.9347 | 0.8893 | 0.8311 | 0.7622 |
| 7 | 1.0000 | 1.0000 | 0.9998 | 0.9989 | 0.9958 | 0.9881 | 0.9733 | 0.9489 | 0.9134 | 0.8666 |
| 8 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9989 | 0.9962 | 0.9901 | 0.9786 | 0.9597 | 0.9319 |
| 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9989 | 0.9967 | 0.9919 | 0.9829 | 0.9682 |
| 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9990 | 0.9972 | 0.9933 | 0.9863 |

Poisson distribution Mean and Variance

Assume that, we conduct a Poisson experiment, in which the average number of successes within a given range is taken as $\lambda$. In Poisson distribution, the mean of the distribution is represented by $\lambda$ and e is constant, which is approximately equal to 2.71828. Then, the Poisson probability is:

**$P(x, \lambda ) = (e^{-\lambda} \lambda^x)/x!$**

In Poisson distribution, the mean is represented as **$E(X) = \lambda$.**

For a Poisson Distribution, the mean and the variance are equal. It means that **$E(X) = V(X)$**

Where, V(X) is the variance.

Poisson distribution Expected Value

A random variable is said to have a Poisson distribution with the parameter $\lambda$, where "$\lambda$" is considered as an expected value of the Poisson distribution.

The expected value of the Poisson distribution is given as follows:

$E(x) = \mu = d(e^{\lambda(t-1)})/dt$, at t=1.

$E(x) = \lambda$

Therefore, the expected value (mean) and the variance of the Poisson distribution is equal to $\lambda$.

**Poisson distribution Examples**

An example to find the probability using the Poisson distribution is given below:

**Example 1:**

A random variable X has a Poisson distribution with parameter $\lambda$ such that P (X = 1) = (0.2) P (X = 2). Find P (X = 0).

**Solution:**

For the Poisson distribution, the probability function is defined as:

P (X =x) = $(e^{-\lambda} \lambda^x)/x!$, where $\lambda$ is a parameter.

Given that, P (x = 1) = (0.2) P (X = 2)

$(e^{-\lambda} \lambda^1)/1! = (0.2)(e^{-\lambda} \lambda^2)/2!$

$\Rightarrow \lambda = \lambda^2/ 10$

$\Rightarrow \lambda = 10$

Now, substitute $\lambda = 10$, in the formula, we get:

P (X =0 ) = (e$^{-\lambda}$ $\lambda^0$)/0!

P (X =0) = e$^{-10}$ = 0.0000454

Thus, P (X= 0) = 0.0000454

📧 **Example 2**:

Telephone calls arrive at an exchange according to the Poisson process at a rate λ= 2/min. Calculate the probability that exactly two calls will be received during each of the first 5 minutes of the hour.

**Solution:**

Assume that "N" be the number of calls received during a 1 minute period.

Therefore,

P(N= 2) = (e$^{-2}$. 2$^2$)/2!

P(N=2) = 2e$^{-2}$.

Now, "M" be the number of minutes among 5 minutes considered, during which exactly 2 calls will be received. Thus "M" follows a binomial distribution with parameters n=5 and p= 2e$^{-2}$.

P(M=5) = 32 x e$^{-10}$

P(M =5) = 0.00145, where "e" is a constant, which is approximately equal to 2.718.

## 7.5 Normal Distribution

In probability theory and statistics, the **Normal Distribution**, also called the **Gaussian Distribution**, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other probability distributions, therefore supporting the usage of the word 'normal 'as in about the one, mostly used.

Normal Distribution Definition

The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

$f(x) \geq 0 \ \forall \ x \in (-\infty, +\infty)$

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

Normal Distribution Formula

The probability density function of normal or gaussian distribution is given by;

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

Where,

x is the variable

μ is the mean

σ is the standard deviation

**Normal Distribution Curve**

The random variables following the normal distribution are those whose values can find any unknown value in a given range. For example, finding the height of the students in the school. Here, the distribution can consider any value, but it will be bounded in the range say, 0 to 6ft. This limitation is forced physically in our query.

Whereas, the normal distribution doesn't even bother about the range. The range can also extend to –∞ to + ∞ and still we can find a smooth curve. These random variables are called Continuous Variables, and the Normal Distribution then provides here probability of the value lying in a particular range for a given experiment. Also, use the normal distribution calculator to find the probability density function by just providing the mean and standard deviation value.

### Normal Distribution Standard Deviation

Generally, the normal distribution has any positive standard deviation. We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out. If the standard deviation is smaller, the data are somewhat close to each other and the graph becomes narrower. If the standard deviation is larger, the data are dispersed more, and the graph becomes wider. The standard deviations are used to subdivide the area under the normal curve. Each subdivided section defines the percentage of data, which falls into the specific region of a graph.

Using 1 standard deviation, the Empirical Rule states that,

Approximately 68% of the data falls within one standard deviation of the mean. (i.e., Between Mean- one Standard Deviation and Mean + one standard deviation)

Approximately 95% of the data falls within two standard deviations of the mean. (i.e., Between Mean- two Standard Deviation and Mean + two standard deviations)

Approximately 99.7% of the data fall within three standard deviations of the mean. (i.e., Between Mean- three Standard Deviation and Mean + three standard deviations)
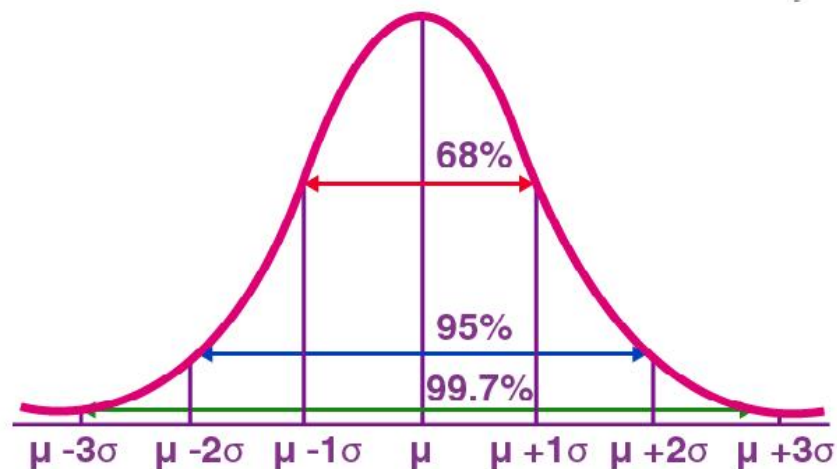


Thus, the empirical rule is also called the 68 – 95 – 99.7 rule.

Normal Distribution Problems and Solutions

**Example:**

**Question 1: Calculate the probability density function of normal distribution using the following data. x = 3, µ = 4 and σ = 2.**

Solution: Given, variable, x = 3

Mean = 4 and

Standard deviation = 2

By the formula of the probability density of normal distribution, we can write;

$$f(3, 4, 2) = \frac{1}{2\sqrt{2\pi}} e^{\frac{-(3-2)^2}{2 \times 2^2}}$$

Hence, f(3,4,2) = 1.106.

**Question 2: If the value of random variable is 2, mean is 5 and the standard deviation is 4, then find the probability density function of the Gaussian distribution.**

Solution: Given,

Variable, x = 2

Mean = 5 and

Standard deviation = 4

By the formula of the probability density of normal distribution, we can write;

$$f(2, 2, 4) = \frac{1}{4\sqrt{2\pi}} e^{\frac{-(2-2)^2}{2 \times 4^2}}$$

f(2,2,4) = 1/(4√2π) e$^0$

f(2,2,4) = 0.0997

There are two main parameters of normal distribution in statistics namely mean and standard deviation.

The location and scale parameters of the given normal distribution can be estimated using these two parameters.

**Normal Distribution Properties**

Some of the important properties of the normal distribution are listed below:
In a normal distribution, the mean, mean and mode are equal.(i.e., Mean = Median= Mode).

The total area under the curve should be equal to 1.

The normally distributed curve should be symmetric at the centre.

There should be exactly half of the values are to the right of the centre and exactly half of the values are to the left of the centre.

The normal distribution should be defined by the mean and standard deviation.

The normal distribution curve must have only one peak. (i.e., Unimodal)

The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.

Applications

The normal distributions are closely associated with many things such as:

Marks scored on the test

Heights of different persons

Size of objects produced by the machine

Blood pressure and so on.

Binomial Distribution is the widely used probability distribution, derived from Bernoulli Process, (a random experiment named after a renowned mathematician Bernoulli). It is also known as biparametric distribution, as it is featured by two parameters n and p. Here, n is the repeated trials and p is the success probability. If the value of these two parameters is known, then it means that the distribution is fully known. The mean and variance of the binomial distribution are denoted by $\mu = np$ and $\sigma2 = npq$.

In the late 1830s, a famous French mathematician Simon Denis Poisson introduced this distribution. It describes the probability of the certain number of events happening in a fixed time interval. It is uniparametric distribution as it is featured by only one parameter $\lambda$ or m.

In Poisson distribution mean is denoted by m i.e. $\mu = m$ or $\lambda$ and variance is labelled as $\sigma2 = m$ or $\lambda$

| BASIS FOR COMPARISON | BINOMIAL DISTRIBUTION | POISSON DISTRIBUTION |
|---|---|---|
| Meaning | Binomial distribution is one in which the probability of repeated number of trials are studied. | Poisson Distribution gives the count of independent events occur randomly with a given period of time. |
| Nature | Biparametric | Uniparametric |
| Number of trials | Fixed | Infinite |
| Success | Constant probability | Infinitesimal chance of success |
| Outcomes | Only two possible outcomes, i.e. success or failure. | Unlimited number of possible outcomes. |
| Mean and Variance | Mean > Variance | Mean = Variance |
| Example | Coin tossing experiment. | Printing mistakes/page of a large book. |

## Summary

- The **binomial distribution** is a common discrete **distribution** used in **statistics**, as opposed to a continuous **distribution**, such as the normal **distribution**. The **binomial distribution**, therefore, represents the probability for x successes in n trials, given a success probability p for each trial.

- The **binomial distribution** is a common discrete **distribution** used in **statistics**, as opposed to a continuous **distribution**, such as the normal **distribution**. The **binomial distribution**, therefore, represents the probability for x successes in n trials, given a success probability p for each trial.

- Each trail can have only two outcomes or the outcomes that can be reduced to two outcomes. These outcomes can be either a success or a failure.

- The main difference between normal distribution and binomial distribution is that while binomial disribution is discrete.

- This means that in binomial distribution there are no data points between any two data points. This is very different from a normal distribution which has continuous data points.

- In other words, there are a finite amount of events in a binomial distribution, but an infinite number in a normal distribution

## Keywords

Binomial distributions must also meet the following criteria:

- **The number of observations or trials is fixed.** In other words, you can only figure out the probability of something happening if you do it a certain number of times. This is common sense—if you toss a coin once, your probability of getting a tails is 50%. If you toss a coin a 20 times, your probability of getting a tails is very, very close to 100%.

- **Each observation or trial is** independent. In other words, none of your trials have an effect on the probability of the next trial.

- Discrete probability functions are also known as probability mass functions and can assume a discrete number of values. For example, coin tosses and counts of events are discrete functions. These are discrete distributions because there are no in-between values

- Normal distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

- The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution.

- The skewness measures the symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set has a skewness less than zero, or negative skewness, then the left tail of the distribution is longer than the right tail; positive skewness implies that the right tail of the distribution is longer than the left.

- The kurtosis statistic measures the thickness of the tail ends of a distribution in relation to the tails of the normal distribution.

- In statistics, a Poisson distribution is a probability distribution that can be used to show how many times an event is likely to occur within a specified period of time. Poisson distributions are often used to understand independent events that occur at a constant rate within a given interval of time

## Self Assessment

1. In a Binomial Distribution, if 'n' is the number of trials and 'p' is the probability of success, then the mean value is given by _____

A. np

B. n

C. p

D. np(1-p)

2. In a Binomial Distribution, if p, q and n are probability of success, failure and number of

trials respectively then variance is given by _____

A. np

B. npq

C. np2q

D. npq2

3. It is suitable to use Binomial Distribution only for _____

A. Large values of 'n'

B. Fractional values of 'n'

C. Small values of 'n'

D. Any value of 'n'

4.. For larger values of 'n', Binomial Distribution _____

A. loses its discreteness

B. tends to Poisson Distribution

C. stays as it is

D. gives oscillatory values

5. Binomial Distribution is a _____

A. Continuous distribution

B. Discrete distribution

C. Irregular distribution

D. Not a Probability distribution

6. Poison Distribution is a _____

A. Continuous distribution

B. Discrete distribution

C. Irregular distribution

D. Not a Probability distribution

7. Normal Distribution is a _____

A. Continuous distribution

B. Discrete distribution

C.  Irregular distribution

D.  Not a Probability distribution

8. A _____ is one in which the data can only take on certain values, for example

integers.

A.  Continuous distribution

B.  Discrete distribution

C.  Irregular distribution

D.  Not a Probability distribution

9. A _____ is one in which data can take on any value within a specified range (which

may be infinite).

A.  Continuous distribution

B.  Discrete distribution

C.  Irregular distribution

D.  Not a Probability distribution

10. Height, weight, temperature and length are all examples of continuous data are example

of

A.  Continuous data
B.  Discrete data
C.  Irregular data
D.  None of the above

11. Mean median and mode are equal in

A.  Continuous distribution
B.  Normal distribution
C.  Irregular distribution
D.  Not a Probability distribution

12. In a Poisson distribution, if 'n' is the number of trials and 'p' is the probability of success,

then the mean value is given by?

A.  m = np

B.  m = (np)2

C.  m = np(1-p)

D.  m = p

13. . If 'm' is the mean of a Poisson Distribution, then variance is given by _____

A.  m2

B. m1/2

C. m

D. m2

14. Poisson distribution is applied for _____

A. Continuous Random Variable

B. Discrete Random Variable

C. Irregular Random Variable

D. Uncertain Random Variable

15. The _____ is used to describe the distribution of rare events in a large population

A. Continuous distribution
B. Normal distribution
C. Poison distribution
D. Not a Probability distribution

## Answers for Self Assessment

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. | A | 2. | B | 3. | C | 4. | B | 5. | B |
| 6. | B | 7. | A | 8. | B | 9. | A | 10. | A |
| 11. | B | 12. | A | 13. | C | 14. | B | 15. | C |

## Review Questions

1. What does binomial distribution mean?
2. What is an example of a binomial probability distribution?
3. How to Tell When a Random Variable Doesn't Have a Binomial Distribution
4. What is the Poisson distribution in statistics?
5. When should Poisson distribution be used?
6. What is the difference between Poisson and binomial distribution?
7. What is the skewness of Poisson distribution?
8. What is the standard deviation of a Poisson distribution?
9. What is measure of kurtosis?
10. What are some real world examples of normal distribution?

## Further Readings

An Introduction to Probability and Statistics

Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

First Course in Probability, A

Book by Sheldon M. Ross

Schaums Theory and Problems of Statistics

Book by Murray R. Spiegel

Introduction to Probability, Statistics, and Random ...

Book by Hossein Pishro-Nik

**Web Links**

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability

# Unit 08: Statistical Quality Control

**CONTENTS**

Objectives

Introduction

8.1      Statistical quality control techniques

8.2      SQC vs. SPC

8.3      Control Charts

8.4      X Bar S Control Chart Definitions

8.5      P-chart

8.6      Np-chart

8.7      c-chart

8.8      Importance of Quality Management

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

## Objectives

After this unit Students will able to

- Understand basics of Statistical quality control
- Learn concepts of control charts.
- Define basic terms of X bar and R chart.  .
- Understand concept of X bar and S chart in statistics.
- Solve basic questions related to control charts.

## Introduction

**Statistics:** Statistics means data, a good amount of data. Or simply, the collaborative study of accumulation, analysis, interpretation and presentation of massive volumes of data.

**Statistical tools:** Applications of statistical methods in order to visualize, interpret and anticipate outcomes over collected data.

**Quality** "a characteristic of fitness for purpose at lowest cost", or "degree of perfection that suffices the customer requirements". Quality can be defined as "the entirety of features and characteristics for products and services satisfying implicit and explicit demands of customers.

**Control:** An approach of measuring and inspecting a certain phenomenon for a product or a service, control suggests when to inspect, and how much to inspect. The system includes feedback to understand the causes for poor quality and necessary corrective steps. The control system basically determines the quality characteristics of an item, correlates the same with predefined quality standards and distinguishes between defective items from non-defectives ones.

**Quality control:** Quality control is one of the most important tools deployed to check the definite level of quality of products, or services. In todays' highly competitive business environment, quality control has evolved as a prominent tool and a critical factor via any successful industry to ensure standard quality.

Making use of statistical tools and techniques in order to monitor and manage product quality across various industries including food, pharmaceutical and manufacturing units, and the process is named as Statistical Quality Control. The method can be conducted as

A part of production process,

A part of last-minute quality control check

A part of eventual check by quality control department

## 8.1   Statistical quality control techniques

Statistical quality control techniques are extremely important for operating the estimable variations embedded in almost all manufacturing processes. Such variations arise due to raw material, consistency of product elements, processing machines, techniques deployed and packaging applications. Moreover, any of these factors or combination of two can impact the eventual quality of finished product.

The method incorporates legislation allowing manufacturing units to make sure that the finished product must contain the net quantity mentioned in packaging. Any overfilled quantity can lead to financial loss for the manufacturer and therefore must be avoided. Fill control, validating weight and weight variation are hugely deployed statistical quality control techniques that make use of weights of individual products in the statistical data analysis.

In case of pharmaceutical goods, such as tablets, pills, capsules, syrups etc, the standard weight must not be exceeded the upper limit that saves consumers from taking high doses of active ingredients that might result in severe consequences. At the same time, the weight shouldn't be too less, if not the drug might not be effective. In this case, the weight variation based statistical quality control test is used to ensure the consistency of the dosage unit, and also to support product identity, reliability and quality.

Another example would be, in the production of food and beverages, it is required to inspect the weight of packages rendering quick confirmation such that filled quantities fulfil the legal necessities. Any deviation from standard value signifies errors in the production process, imprecise ingredient-quantities leading to impactful consequences.

In addition to this, while confirming consumer satisfaction, safety and compliance with regulations, SQC with weight determination is highly important. Though, it is recommended to employ actual balances or measuring scales and software suitable for particular applications.

Advantages of Statistical Quality Control

One of the excellent scientific tools, SQC has the following advantages

Cost reduction: In this method, only a fragmentary output is inspected to ensure the quality of product, therefore probe cost would be reduced greatly.

Huge efficiency: Inspection of a fractional portion requires lesser time and tedium in comparison to holistic investigation leading to huge escalation in efficiency and production.

Easier to use: Pitching SQC not only reduces process variability but also makes the process of production-in-control. Even, it is much to apply by an individual without having such extensive specialized guidance.

Authentic anticipation: SQC is the most preeminent approach that can accurately predict future production. To ensure the degree of perfection and product performance, SQC provides a great predictability.

Prior fault detection: Any deviation from standard control limits depicts signs of danger in the underlying production process that invites necessary corrective measurement to be taken earlier. SQC is helpful in early detection of faults.

## 8.2  SQC vs. SPC

Both SQC and SPC support smooth operations in order to escalate efficiencies, desired output and optimized performance while playing a key role in overall success in operations, but in different ways. Lets' understand the difference;

SPC: is the procedure of collecting and computing parameters of a process such as speed, pressure, vernier caliper etc with respect to standard values using various statistical methods validating values must reside within limits while aiming to minimize variation and execute to achieve desired/optimum targets.

SQC: is the process of compiling and determining data on the subject of particular specifications regarding a product and to meet requirements, for example, size, weight, texture etc. while aiming at validating process outcomes to meet the user requirements or the next stage of the manufacturing process.

SPC is responsible for reduction of variation in processes and run efficiently, in contrast to this, SQC facilitates manufacturers to accomplish user requirements.

For example, in food and beverage manufacturing there are various numbers of different products being produced, SPC monitors that operations are executing effectively at their entirety, SQC controls measurable quality characteristics used during production so that finished products must live up with customer requirements/expectations.

Statistical process control uses sampling and statistical methods to monitor the quality of an ongoing process such as a production operation. A graphical display referred to as a control chart provides a basis for deciding whether the variation in the output of a process is due to common causes (randomly occurring variations) or due to out-of-the-ordinary assignable causes. Whenever assignable causes are identified, a decision can be made to adjust the process in order to bring the output back to acceptable quality levels.

Control charts can be classified by the type of data they contain. For instance, an $\bar{x}$-chart is employed in situations where a sample mean is used to measure the quality of the output. Quantitative data such as length, weight, and temperature can be monitored with an $\bar{x}$-chart. Process variability can be monitored using a range or R-chart. In cases in which the quality of output is measured in terms of the number of defectives or the proportion of defectives in the sample, an np-chart or a p-chart can be used.

All control charts are constructed in a similar fashion. For example, the centre line of an $\bar{x}$-chart corresponds to the mean of the process when the process is in control and producing output of acceptable quality. The vertical axis of the control chart identifies the scale of measurement for the variable of interest. The upper horizontal line of the control chart, referred to as the upper control limit, and the lower horizontal line, referred to as the lower control limit, are chosen so that when the process is in control there will be a high probability that the value of a sample mean will fall between the two control limits. Standard practice is to set the control limits at three standard deviations above and below the process mean. The process can be sampled periodically. As each sample is selected, the value of the sample mean is plotted on the control chart. If the value of a sample mean is within the control limits, the process can be continued under the assumption that the quality standards are being maintained. If the value of the sample mean is outside the control limits, an out-of-control conclusion points to the need for corrective action in order to return the process to acceptable quality levels.

*Probability and Statistics*

## 8.3    Control Charts

**X-bar and range chart**

What is it?

An X-bar and R (range) chart is a pair of control charts used with processes that have a subgroup size of two or more. The standard chart for variables data, X-bar and R charts help determine if a process is stable and predictable. The X-bar chart shows how the mean or average changes over time and the R chart shows how the range of the subgroups changes over time. It is also used to monitor the effects of process improvement theories. As the standard, the X-bar and R chart will work in place of the X-bar and s or median and R chart.

What does it look like?

The X-bar chart, on top, shows the mean or average of each subgroup. It is used to analyze central location . The range chart, on the bottom, shows how the data is spread . It is used to study system variability



From "Improving Healthcare with Control Charts", Raymond G. Carey, PhD.

When is it used?

You can use X-bar and R charts for any process with a subgroup size greater than one. Typically, it is used when the subgroup size falls between two and ten, and X-bar and s charts are used with subgroups of eleven or more.

Use X-bar and R charts when you can answer yes to these questions:

Do you need to assess system stability?

Is the data in variables form?

Is the data collected in subgroups larger than one but less than eleven?

Is the time order of subgroups preserved?

Getting the most

Collect as many subgroups as possible before calculating control limits. With smaller amounts of data, the X-bar and R chart may not represent variability of the entire system. The more subgroups you use in control limit calculations, the more reliable the analysis. Typically, twenty to twenty-five subgroups will be used in control limit calculations.

X-bar and R charts have several applications. When you begin improving a system, use them to assess the system's stability.

After the stability has been assessed, determine if you need to stratify the data. You may find entirely different results between shifts, among workers, among different machines, among lots of materials, etc. To see if variability on the X-bar and R chart is caused by these factors, collect and enter data in a way that lets you stratify by time, location, symptom, operator, and lots.

You can also use X-bar and R charts to analyze the results of process improvements. Here you would consider how the process is running and compare it to how it ran in the past. Do process changes produce the desired improvement?

Example: Finally, use X-bar and R charts for standardization. This means you should continue collecting and analyzing data throughout the process operation. If you made changes to the system and stopped collecting data, you would have only perception and opinion to tell you whether the changes actually improved the system. Without a control chart, there is no way to know if the process has changed or to identify sources of process variability.

What are X Bar S Control Charts?

X Bar S charts often used control chart to examine the process mean and standard deviation over the time. These charts are used when the subgroups have large sample size and S chart provides better understanding of the spread of subgroup data than range.

X bar S charts are also similar to X Bar R Control chart, the basic difference is that X bar S charts plots the subgroup standard deviation whereas R charts plots the subgroup range

Selection of appropriate control chart is very important in control charts mapping, otherwise ended up with inaccurate control limits for the data.

**Lovely Professional University**
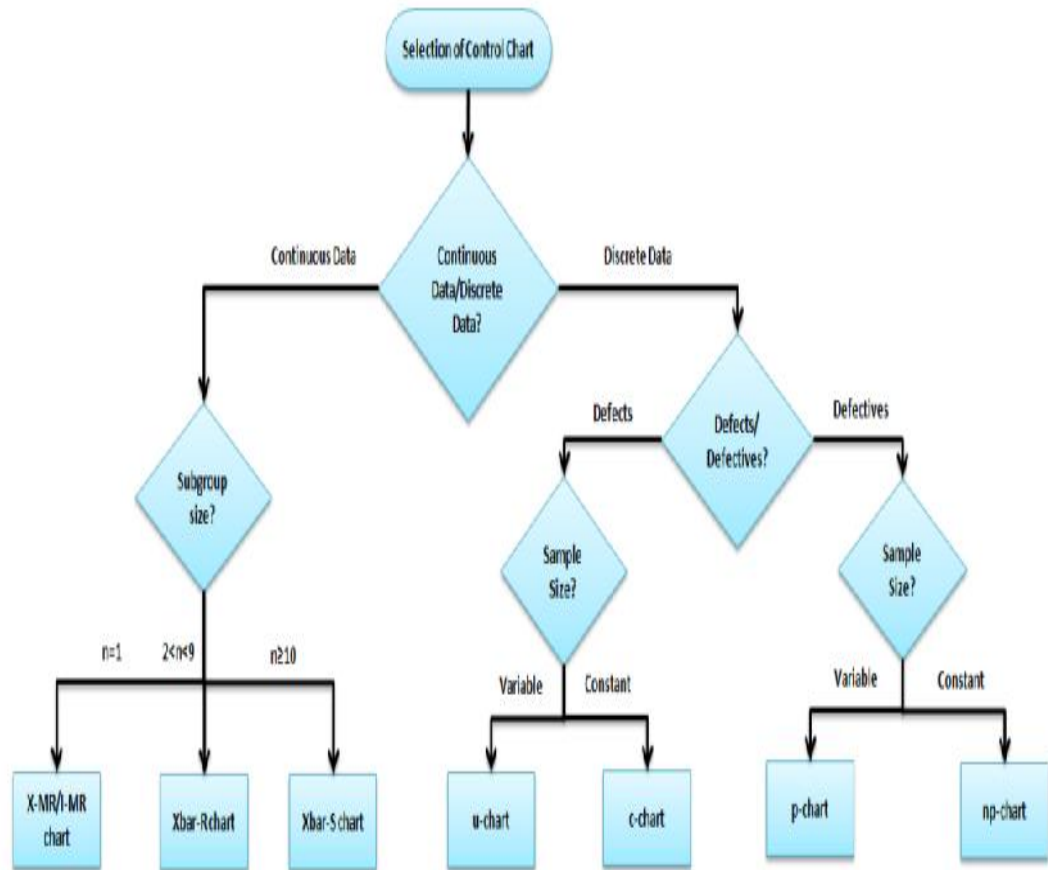
Manually it is very easy to compute X Bar R Control chart,whereas sigma chart may be difficult due to tedious calculations and large sample size. With large sample size in the subgroup, the standard deviation is better measure of variation than the range because it considers all the data not just minimum and maximum values.

It is actually a two plots to monitor the process mean and the process range (as described by standard deviation) over time and is an example of statistical process control. These combination charts helps to understand the stability of processes and also detects the presence of special cause variation.

The cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) charts are also monitors the mean of the process, but the basic difference is unlike X bar chart they consider the previous value means at each point. Moreover these charts are considered as a reliable estimate when correct standard deviation exists.

Task: What are conditions to use X bar R chart ?

## 8.4   X Bar S Control Chart Definitions

**X-bar chart:** The mean or average change in process over time from subgroup values. The control limits on the X-Bar brings the sample's mean and center into consideration.

**S-chart:** The standard deviation of the process over the time from subgroups values. This monitors the process standard deviation (as approximated by the sample moving range)

**Use X Bar S Control Charts When:**

The sampling procedure is same for each sample and is carried out consistently.

When the data is assumed to be normally distributed.

**Lovely Professional University**

The X bar S chart to be used when rationally collect measurements in subgroup size is more than 10.

X Bar R chart is to be considered if the subgroup size is between two and 10 observations (for I-MR chart the subgroup size is one only).

When the collected data is in continuous (ie Length, Weight) etc. and captures in time order

**How to Interpret the X Bar S Control Charts**

To correctly interpret X bar S chart, always examine the S chart first.

The X bar chart control limits are derived from the S bar (average standard deviation) values, if the values are out of control in S chart that means the X bar chart control limits are not accurate.

If the points are out of control in S chart, then stop the process. Identify the special cause and address the issue. Remove those subgroups from the calculations.

Once the S chart is in control, then review X bar chart and interpret the points against the control limits.

All the points to be interpret against the control limits but not specification limits.

If any point out of control in X bar chat. Identify the special cause and address the issue.

## Compute the process standard deviation, if the S chart is in statistical control $\hat{\sigma} = \dfrac{\bar{s}}{c_4}$

**Steps to follow for X bar S chart**

**Objective of the chart and subgroup size**

Determine the objective of the chart and choose the important variables

Choose the appropriate subgroup size and the sampling frequency

Shewhart suggested collecting 20 to 25 sets of samples with a subgroup size of 10 and above

**Note:** To demonstrate an example, we just took subgroup size 4 in the below example, but it is always recommended to take 10 and above for X bar S chart.

**Example:** A packing organization monitoring the performance of a packing machine, each container should weigh 35 lb, during Measure phase, project team performed the process capability study and identified that the process is not capable(less than one sigma). In Analyze phase collected 12 sets of container weights with a subgroup size of 4.

| Sample | Measured values | | | |
|--------|------|------|------|------|
|        | 1 | 2 | 3 | 4 |
| 1  | 65 | 63 | 55 | 54 |
| 2  | 54 | 55 | 55 | 55 |
| 3  | 66 | 14 | 54 | 34 |
| 4  | 37 | 54 | 36 | 35 |
| 5  | 67 | 10 | 12 | 65 |
| 6  | 36 | 36 | 37 | 37 |
| 7  | 10 | 12 | 12 | 14 |
| 8  | 36 | 36 | 35 | 24 |
| 9  | 36 | 46 | 35 | 36 |
| 10 | 38 | 46 | 36 | 34 |
| 11 | 55 | 12 | 67 | 55 |
| 12 | 22 | 22 | 33 | 12 |

**Compute X bar and S values**

Measure the average of each subgroup i.e X bar, then compute grand average of all X bar value, this will be center line for X bar chart

Compute the standard deviation of each subgroup, then measure grand averages of all standard values ie S bar and this will be the center line for S chart

| | Measured values | | | | Sample Average($\bar{X}$) | ($\bar{X}$-X$_i$) | | | | ($\bar{X}$-X$_i$)$^2$ | | | | $\bar{s} = \sqrt{\frac{\sum(X-Xi)^2}{n-1}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | 1 | 2 | 3 | 4 | | | | | | | | | | |
| 1 | 65 | 63 | 55 | 54 | 59.3 | -5.75 | -3.75 | 4.25 | 5.25 | 33.06 | 14.06 | 18.06 | 27.56 | 5.56 |
| 2 | 54 | 55 | 55 | 55 | 54.8 | 0.75 | -0.25 | -0.25 | -0.25 | 0.56 | 0.06 | 0.06 | 0.06 | 0.50 |
| 3 | 66 | 14 | 54 | 34 | 42.0 | 24.00 | 28.00 | 12.00 | 8.00 | 576.00 | 784.00 | 144.00 | 64.00 | 22.86 |
| 4 | 37 | 54 | 36 | 35 | 40.5 | 3.50 | 13.50 | 4.50 | 5.50 | 12.25 | 182.25 | 20.25 | 30.25 | 9.04 |
| 5 | 67 | 10 | 12 | 65 | 38.5 | 28.50 | 28.50 | 26.50 | 26.50 | 812.25 | 812.25 | 702.25 | 702.25 | 31.78 |
| 6 | 36 | 36 | 37 | 37 | 36.5 | 0.50 | 0.50 | -0.50 | -0.50 | 0.25 | 0.25 | 0.25 | 0.25 | 0.58 |
| 7 | 10 | 12 | 12 | 14 | 12.0 | 2.00 | 0.00 | 0.00 | -2.00 | 4.00 | 0.00 | 0.00 | 4.00 | 1.63 |
| 8 | 36 | 36 | 35 | 24 | 32.8 | -3.25 | -3.25 | -2.25 | 8.75 | 10.56 | 10.56 | 5.06 | 76.56 | 5.85 |
| 9 | 36 | 46 | 35 | 36 | 38.3 | 2.25 | -7.75 | 3.25 | 2.25 | 5.06 | 60.06 | 10.56 | 5.06 | 5.19 |
| 10 | 38 | 46 | 36 | 34 | 38.5 | 0.50 | -7.50 | 2.50 | 4.50 | 0.25 | 56.25 | 6.25 | 20.25 | 5.26 |
| 11 | 55 | 12 | 67 | 55 | 47.3 | -7.75 | 35.25 | 19.75 | -7.75 | 60.06 | 1242.56 | 390.06 | 60.06 | 24.17 |
| 12 | 22 | 22 | 33 | 12 | 22.3 | 0.25 | 0.25 | 10.75 | 10.25 | 0.06 | 0.06 | 115.56 | 105.06 | 8.58 |
| | | | | Total | 462.5 | | | | | | | | | 120.99 |
| | | | | $\bar{X}$ | 38.54 | | | | | | | | | |
| | | | | $\bar{s}$ | | | | | | | | | | 10.08 |

**Determine the Control Limits**

The first set of subgroups are to determine the process mean and standard deviation, these values are to be consider for creation of control limits for both standard deviation and mean of each subgroup

- If the individual values (X) assumed to be normal, then the distribution of average of reading in a sample (X-bar) will be normal
- The standard deviation among the sample means is smaller by a factor of $\frac{1}{\sqrt{n}}$

  Hence
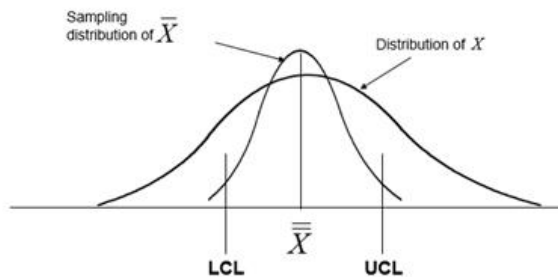


- $\hat{\sigma}$ = estimated standard deviation of X
- $\frac{\hat{\sigma}}{\sqrt{n}}$ = *estimated standard error of $\bar{X}$*

- Walter Shewhart mentioned that control limits should be 3 times standard deviation from the center line in order to reduce the probability of error happening in detecting the assignable causes of variation.

- X bar chart : $UCL_{\bar{x}} = \bar{\bar{x}} + 3\frac{\hat{\sigma}}{\sqrt{n}} = \bar{\bar{x}} + A_3 \bar{S}$

- $\phantom{X bar chart :} LCL_{\bar{x}} = \bar{\bar{x}} - 3\frac{\hat{\sigma}}{\sqrt{n}} = \bar{\bar{x}} - A_3 \bar{S}$

- S chart   : $UCL_{\bar{S}} = B_4 \bar{S}$

- $\phantom{S chart : } LCL_{\bar{S}} = B_3 \bar{S}$

Where

X is the individual value (data)

n is the sample size

X bar is the average of reading in a sample

S is the standard deviation

S bar is the average of all the standard deviation.

UCL is Upper control limit

LCL is Lower control limit

The below control chart constants are approximate values to measure the control limits for X bar S chart and other control charts based on subgroup size

| Subgroup | X bar chart | | Sigma estimate | R chart | | S chart | |
|---|---|---|---|---|---|---|---|
| | A₂ | A₃ | d₂ | D₃ | D₄ | B₃ | B₄ |
| 2 | 1.880 | 2.659 | 1.128 | - | 3.267 | - | 3.267 |
| 3 | 1.023 | 1.954 | 1.693 | - | 2.574 | - | 2.568 |
| 4 | 0.729 | 1.628 | 2.059 | - | 2.282 | - | 2.266 |
| 5 | 0.577 | 1.427 | 2.326 | - | 2.114 | - | 2.089 |
| 6 | 0.483 | 1.287 | 2.534 | - | 2.004 | 0.030 | 1.970 |
| 7 | 0.419 | 1.182 | 2.704 | 0.076 | 1.924 | 0.118 | 1.882 |
| 8 | 0.373 | 1.099 | 2.847 | 0.136 | 1.864 | 0.185 | 1.815 |
| 9 | 0.337 | 1.032 | 2.970 | 0.184 | 1.816 | 0.239 | 1.761 |
| 10 | 0.308 | 0.975 | 3.078 | 0.223 | 1.777 | 0.284 | 1.716 |
| 11 | 0.285 | 0.927 | 3.173 | 0.256 | 1.744 | 0.321 | 1.679 |
| 12 | 0.266 | 0.886 | 3.258 | 0.283 | 1.717 | 0.354 | 1.646 |
| 13 | 0.249 | 0.850 | 3.336 | 0.307 | 1.693 | 0.382 | 1.618 |
| 14 | 0.235 | 0.817 | 3.407 | 0.328 | 1.672 | 0.406 | 1.594 |
| 15 | 0.223 | 0.789 | 3.472 | 0.347 | 1.653 | 0.428 | 1.572 |
| 16 | 0.212 | 0.763 | 3.532 | 0.363 | 1.637 | 0.448 | 1.552 |
| 17 | 0.203 | 0.739 | 3.588 | 0.378 | 1.622 | 0.466 | 1.534 |
| 18 | 0.194 | 0.718 | 3.640 | 0.391 | 1.608 | 0.482 | 1.518 |
| 19 | 0.187 | 0.698 | 3.689 | 0.403 | 1.597 | 0.497 | 1.503 |
| 20 | 0.180 | 0.680 | 3.735 | 0.415 | 1.585 | 0.510 | 1.490 |
| 21 | 0.173 | 0.663 | 3.778 | 0.425 | 1.575 | 0.523 | 1.477 |
| 22 | 0.167 | 0.647 | 3.819 | 0.434 | 1.566 | 0.534 | 1.466 |
| 23 | 0.162 | 0.633 | 3.858 | 0.443 | 1.557 | 0.545 | 1.455 |
| 24 | 0.157 | 0.619 | 3.895 | 0.451 | 1.548 | 0.555 | 1.445 |
| 25 | 0.153 | 0.606 | 3.931 | 0.459 | 1.541 | 0.565 | 1.435 |

Refer common factors for various control charts

*Probability and Statistics*

**Example cont.:** In the above example n=4

$$X \text{ bar chart}: UCL_{\bar{x}} = \bar{\bar{x}} + A_3 \bar{S} = 38.54 + 1.628 * 10.08 = 54.96$$

$$LCL_{\bar{x}} = \bar{\bar{x}} - A_3 \bar{S} = 38.54 - 1.628 * 10.08 = 22.13$$

$$S \text{ chart} : UCL_{\bar{S}} = B_4 \bar{S} = 2.266 * 10.08 = 22.85$$
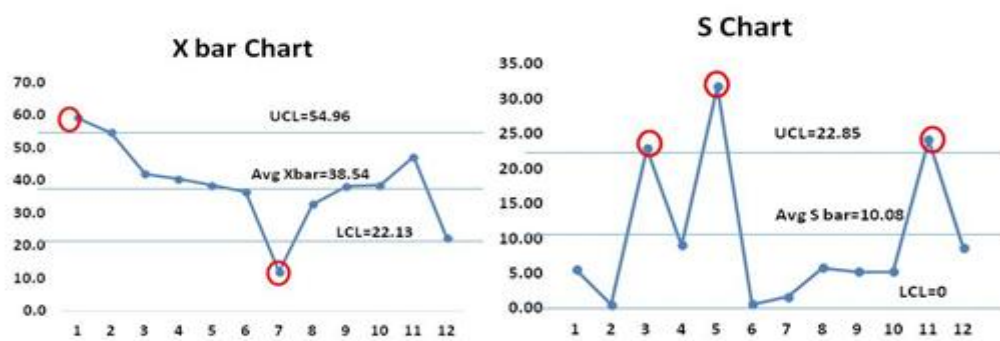
$$LCL_{\bar{S}} = B_3 \bar{S} = 0$$

**Interpret X bar and S chart**

Plot both X bar and S chart and identify the assignable causes

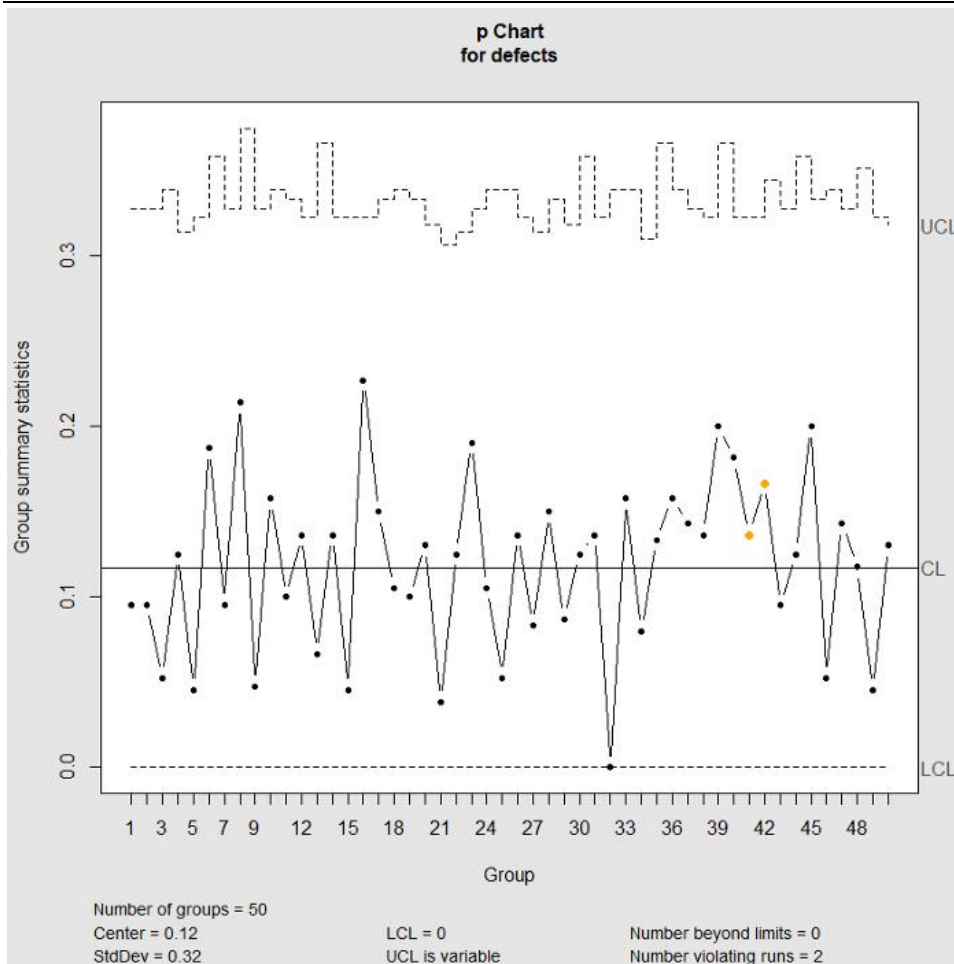**Example Cont:** Use the above values and plot the X bar and Sigma chart

Task: What is difference between X bar R vs. X bar S chart?



From the both X bar and S charts it is clearly evident that most of the values are out of control, hence the process is not stable
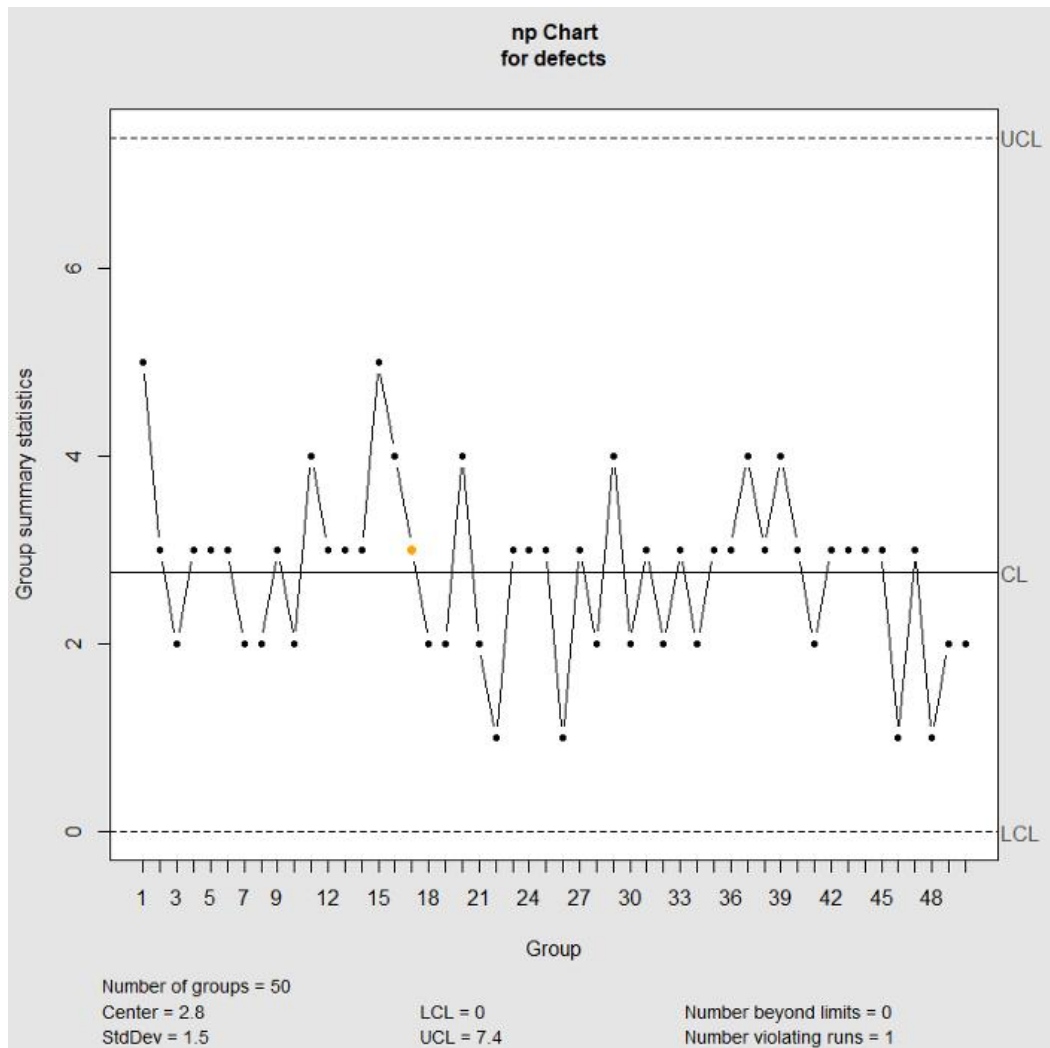
## 8.5 P-chart

The p-chart is a quality control chart used to monitor the **proportion** of nonconforming units in different samples of size $n$; it is based on the binomial distribution where each unit has only two possibilities (i.e. defective or not defective). The y-axis shows the proportion of nonconforming units while the x-axis shows the sample group. Let's take a look at the *R* code using the *qcc* package to generate a p-chart

p Chart
for defects

Number of groups = 50
Center = 0.12          LCL = 0                    Number beyond limits = 0
StdDev = 0.32          UCL is variable            Number violating runs = 2

The p-chart generated by *R* provides significant information for its interpretation, including the samples (Number of groups), both control limits (UCL and LCL), the overall proportion mean (Center) the standard deviation (StdDev), and most importantly, the points beyond the control limits and the violating runs. Engineers must take a special look at these points in order to identify and assign causes attributed to changes in the system that led to nonconforming units.

## 8.6    Np-chart

The np-chart is a quality control chart used to monitor the count of nonconforming units in fixed samples of size n. The y-axis shows the total count of nonconforming units while the x-axis shows the sample group. Let's take a look at the R code using the qcc package to generate a np-chart

The np-chart generated by *R* also provides significant information for its interpretation, just as the p-chart generated above. In the same way, engineers must take a special look to points beyond the control limits and to violating runs in order to identify and assign causes attributed to changes on the system that led to nonconforming units.

NP charts are used to monitor the number of nonconforming units of a process based on samples taken from the process at given times (hours, shifts, days, weeks, months, etc.). Typically, an initial series of samples is used to estimate the average number of nonconforming units per sample. The estimated average is then used to produce control limits for the number of nonconforming units. During this initial phase, the process should be in control. If points are out-of-control during the initial (estimation) phase, the assignable cause should be determined, and the sample should be removed from estimation. Once the control limits have been established for the NP chart, these limits may be used to monitor the number nonconforming going forward. When a point is outside these established control limits it indicates that the number of nonconforming units of the process is out-of-control. An assignable cause is suspected whenever the control chart indicates an out-of-control process

The NP Chart versus the P Chart the NP chart is very similar to the P chart. Rather than focusing on the proportion of nonconforming units, as does the P chart, the NP chart focuses on the average number of non-conforming units. As such, choice of the P or NP chart is simply a matter of preference, as each is a scaled version of the other. One case in which the P chart may be recommended over the NP chart is the case where the sample size varies across samples, since the P chart is easier to interpret for this scenario.

## 8.7 c-chart

A c-chart is an attributes control chart used with data collected in subgroups that are the same size. C-charts show how the process, measured by the number of nonconformities per item or group of items, changes over time. Nonconformities are defects or occurrences found in the sampled subgroup. They can be described as any characteristic that is present but should not be, or any characteristic that is not present but should be. For example a scratch, dent, bubble, blemish, missing button, and a tear would all be nonconformities. C-charts are used to determine if the process is stable and predictable, as well as to monitor the effects of process improvement theories.

What does it look like?

The c-chart shows the number of nonconformities in subgroups of equal size.
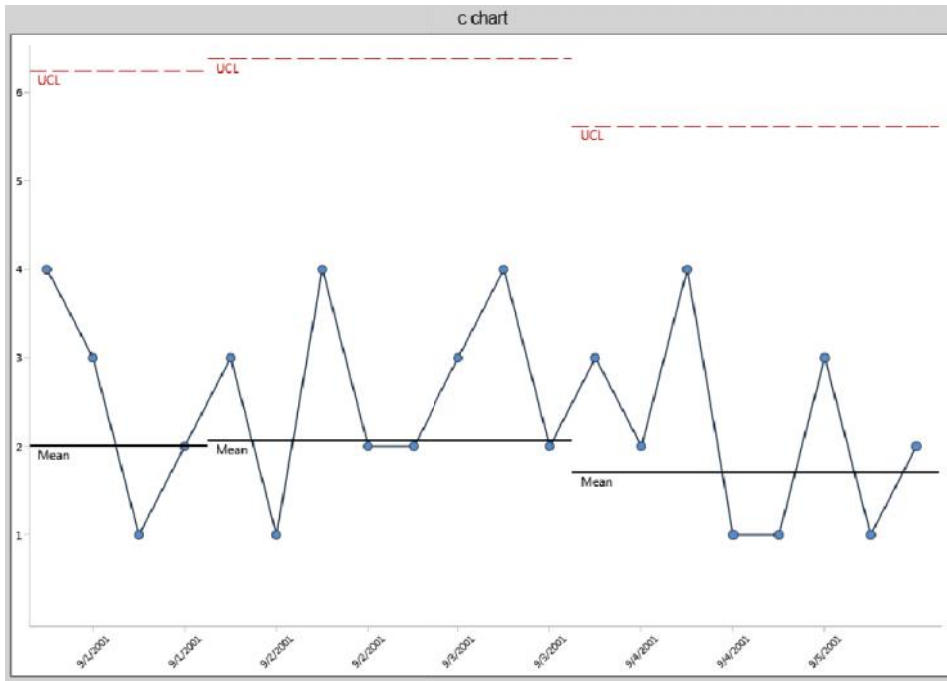


Attribute chart: c chart is also known as the control chart for defects (counting of the number of defects). It is generally used to monitor the number of defects in constant size units. There may be a single type of defect or several different types, but the c chart tracks the total number of defects in each unit and it assumes the underlying data approximate the Poisson distribution. The unit may be a single item or a specified section of items—for example, scratches on plated metal, number of insufficient soldering in a printed circuit board.

c chart takes into account the number of defects in each defective unit or in a given sample. While p chart analyzes the proportions of non-conforming or defective items in a process.

c chart, the number of defects is plotting on the y-axis and the number of units on the x-axis. The centerline of the c chart ($\bar{c}$) is the total number of defects divided by the number of samples

$$\bar{c} = \frac{Total\ number\ of\ defects}{Number\ of\ samples}$$

**Why and when do you use a c Chart?**

C chart is one of the quality control charts used to track the number of defects in a product of constant size, while u chart is used for a varying size.

c-chart is used to determine if the process is stable and predictable and also to monitor the effects of before and after process improvements. c chart is especially used when there are high opportunities for defects in the subgroup, but the actual number of defects is less.

c chart requires that each subgroup's sample size be the same and compute control limits based on the Poisson distribution.

Four types of control charts exist for attribute data. p chart plots the proportion of defective items, and np chart is for the number of defectives. u chart is for the average number of defects per unit and c chart is for the number of defects.
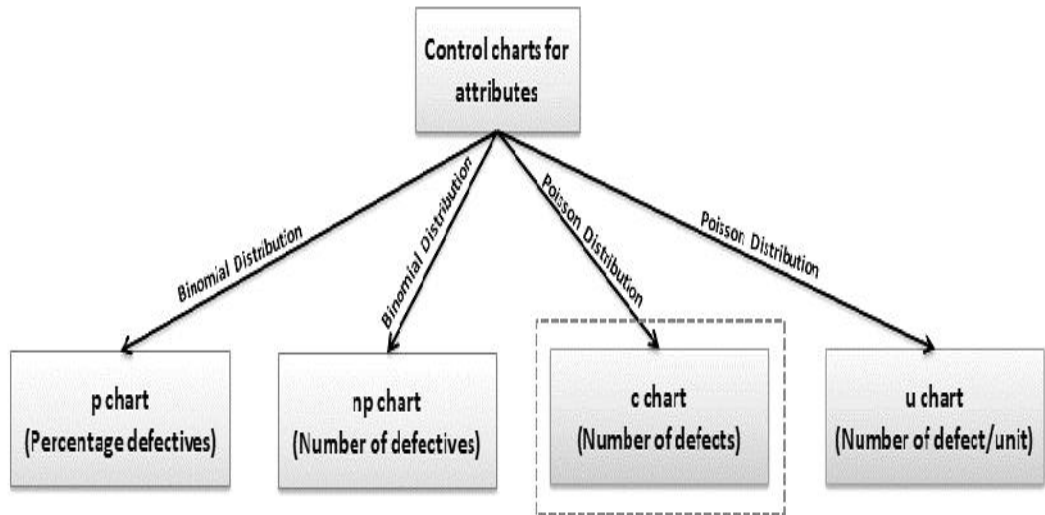


**Assumptions of Attribute charts: c chart**

The probability of defect is the same for each item

Each unit is independent of the other

The testing procedure should be the same for each lot

## 8.8 Importance of Quality Management

**"Quality management" ensures superior quality products and services**. Quality of a product can be measured in terms of performance, reliability and durability. Quality is a crucial parameter which differentiates an organization from its competitors. Quality management tools ensure changes in the systems and processes which eventually result in superior quality products and services. Quality management methods such as Total Quality management or Six Sigma have a common goal - to deliver a high quality product. Quality management is essential to create superior quality products which not only meet but also exceed customer satisfaction. Customers need to be satisfied with your brand. Business marketers are successful only when they emphasize on quality rather than quantity. Quality products ensure that you survive the cut throat competition with a smile.

**Quality management is essential for customer satisfaction which eventually leads to customer loyalty**. How do you think businesses run? Do businesses thrive only on new customers? It is important for every business to have some loyal customers. You need to have some customers who would come back to your organization no matter what.

Would you buy a Nokia mobile again if the previous handset was defective? The answer is NO.

Customers would return to your organization only if they are satisfied with your products and services. Make sure the end-user is happy with your product. Remember, a customer would be happy and satisfied only when your product meets his expectations and fulfills his needs. Understand what the customer expects from you? Find out what actually his need is? Collect relevant data which would give you more insight into customer's needs and demands. Customer feedbacks should be collected on a regular basis and carefully monitored. Quality management ensures high quality products and services by eliminating defects and incorporating continuous changes and improvements in the system. High quality products in turn lead to loyal and satisfied customers who bring ten new customers along with them. Do not forget that you might save some money by ignoring quality management processes but ultimately lose out on your major customers, thus incurring huge losses. Quality management ensures that you deliver products as per promises made to the customers through various modes of promotions. **Quality management tools help an organization to design and create a product which the customer actually wants and desires**.

**Quality Management ensures increased revenues and higher productivity for the organization**. Remember, if an organization is earning, employees are also earning. Employees are frustrated only when their salaries or other payments are not released on time. Yes, money is a strong motivating factor. Would you feel like working if your organization does not give you salary on time? Ask yourself. Salaries are released on time only when there is free cash flow. Implementing Quality management tools ensure high customer loyalty, thus better business, increased cash flow, satisfied employees, healthy workplace and so on. Quality management processes make the organization a better place to work.

Remove unnecessary processes which merely waste employee's time and do not contribute much to the organization's productivity. Quality management enables employees to deliver more work in less time.

**Quality management helps organizations to reduce waste and inventory**. It enables employees to work closely with suppliers and incorporate "Just in Time" Philosophy.

Quality management ensures close coordination between employees of an organization. It inculcates a strong feeling of team work in the employees.

Importance of Quality Management

1) Consistent quality and make of the products

It is highly imperative for the firms to plan, design, execute, and manufacture the product offerings for the target market realizing the Importance of Quality Management and maintaining the parameters of total quality management at every facet.

It helps to maintain the realms of quality on a consistent and continuous basis. Plus the firm is able to conduct market research and study on a regular basis having a drive to offering the products that stand as a testimony to the quality and its principles.

2) Ensures long lasting efficiency



When we come to talk about the factor of efficiency whilst discussing the Importance of Quality Management, it is not only confined to the working efficiency of the staff that is into the manufacturing of the products but also to the each and every employee of the firm and even the types of machinery.

When all the employees of the firm right from the engineers to the sales managers understand and follow the Importance of Quality Management, it improves their efficiency as they know that the product that they are manufacturing and selling is best in class.

The confidence and agility that is gained understanding the overall process, elevate their efficiency in manifolds. And all of it has a cascading effect on the overall sales and profits of the firm.

*Probability and Statistics*

3) Higher productivity levels

When the firm realizes and follows the Importance of Quality Management in each of its business operations, there is a rise in the productivity of the employees

They know and understand that they working on something that is unique and high on quality plus due to the high parameters of quality, obstacles and bottlenecks are ironed out automatically, thus, increasing their productivity levels.

4) Attracts a loyal set of customers

It is the thumb rule of every business and industry domain that the business can successfully survive and thrive in the ever competitive market only if it is able to retain the long list of loyal customers.

Customer nowadays is presented with lot many options and alternatives on a silver platter and is much more aware of the quality standards with the help of social media and digital marketing. Hence, it is very crucial for the firms to follow the Importance of Quality Management in order to attract and retain the loyal set of customers and set their cash registers ringing.

5) Beat the competition in the market

For successfully survive and thrive in the market that is all-time high on the competition for the new as well existing brands, it is vital for the firms to understand the Importance of Quality Management and make it as an integral part of its objectives and work culture.

There are many brands in the market that have to shut their stores and business operations in a short period of time as they are unable to adhere to the standards of quality. As mentioned above, the customer nowadays is much more aware and agile plus there are lot many other brands in the market waiting for you to leave the market.

Hence, TQM is one of the sure shot ways and means to beat the competition and carve a distinctive identity for your brand in the market.

6) Enhanced brand value



In continuation of the above-mentioned point, every brand needs a higher market share and an enhanced brand value. And it is the aspect of following and astutely understanding the Importance of Quality Management that helps the firm make its brand value and equity soar amongst other prominent players of the market.

7) Customer Satisfaction:

Customer Satisfaction and following the Importance of Quality Management go hand in hand. Realizing its importance at each and every level of your business operation ensures the higher level of customer satisfaction and happiness.

Majority of customers today wish to go for the products that are high on quality and they don't mind paying an extra amount of money for the same. And if there is any sort of glitch in the quality of the products, the customer realizes the same at the very same moment and perceives the brand in a negative light.

With the power of social media and various industry-specific forums on the digital space, it takes no time for any customer to spread it and make it viral that deteriorates the brand value of the firm.

8) Reduced risks

Yet another aspect that helps the firm to enhance and maintain its brand value in the market is the reduced amount of risks. And risks only occur in the business operations when the firm does not adhere to the parameters of quality.

Risk mainly occur during the manufacturing process of the products and whilst dealing with the customers during the before and after sales procedures. Hence, it is of the vital significance for the firms to understand the Importance of Quality Management especially in these two aspects of the business.

9) Less human errors

When the firm follows the Importance of Quality Management, it also follows a set of guidelines and principles that have been framed for each of the business operations. And right from the top management to the management trainees of the firm, all of them have to follow the same.

This result in the less amount of human errors enhancing the productivity and work efficiency levels. Plus with less human errors there is a very low chance of risks

10) Increased revenues and profits



In today's dynamic market that is ever high on competition, it is very difficult for the firm to generate the desired revenues and profits meeting their long term and short term objectives. And following the Importance of Quality Management is one of the assured ways to accomplish all the business aims and objectives.

It ensures a high level of customer satisfaction, high brand value, higher market share, loyal customers, and a competitive edge. But many a time, firm fail to understand this simple and one of the most crucial fundamentals making them incur losses.

## Summary

An X-bar and R (range) chart is a pair of control charts used with processes that have a subgroup size of two or more

X Bar S charts often used control chart to examine the process mean and standard deviation over the time.

**Quality management" ensures superior quality products and services**. Quality of a product can be measured in terms of performance, reliability and durability

## Keywords

**Statistical tools:** Applications of statistical methods in order to visualize, interpret and anticipate outcomes over collected data.

**Quality** "a characteristic of fitness for purpose at lowest cost", or "degree of perfection that suffices the customer requirements". Quality can be defined as "the entirety of features and characteristics for products and services satisfying implicit and explicit demands of customers.

**Control:** An approach of measuring and inspecting a certain phenomenon for a product or a service, control suggests when to inspect, and how much to inspect

## Self Assessment

1. _____, the use of statistical methods in the monitoring and maintaining of the

   quality of products and services

   A. Statistical quality control

   B. Standard quality control

   C. Symmetry quality control

D. None of these

2. _____uses sampling and statistical methods to monitor the quality of an ongoing

   process such as a production operation.

A. Statistical process control

B. Statistical quality control

C. Standard quality control

D. Symmetry quality control

3. The _____ is a graph used to study how a process changes over time

A. Control chart

B. Carrier chart

C. Histogram

D. Bar chart

4. _____is not suitable for continuously monitoring the process.

A. Control chart
B. Histogram
C. Both of these
D. None of these

5. The chart is particularly advantageous when your sample size is relatively small and

   constant.

A. xbar and r chart

B. xbar and s chart

C. X chart

D. Y chart

6. _____ chart is a pair of control charts used with processes that have a subgroup

   size of two or more

A. xbar and r chart

B. xbar and s chart

C. X chart

D. Y chart

7. X Bar S charts often used control chart to examine the process mean and standard deviation

   over the time

A. xbar and r chart

B. xbar and s chart

C. X chart

D. Y chart


8. These charts are used when the subgroups have large sample size

A. xbar and r chart
B. xbar and s chart
C. X chart
D. Y chart

9. _____charts plots the subgroup standard deviation

A. xbar and r chart
B. xbar and s chart
C. X chart
D. Y chart

10. _____charts plots the subgroup Range

a) xbar and r chart

b) xbar and s chart

c)X chart

d) Y chart


11. Which chart to use when your subgroup sizes are 9 or greater

A. xbar and r chart
B. xbar and s chart
C. X chart
D. Y chart

12. _____is a quality control chart used to monitor the proportion of nonconforming units

   in different samples of size *n.*

A. xbar and r chart
B. xbar and s chart
C. P chart
D. Y chart

13.  The _____is a quality control chart used to monitor the count of nonconforming units

   in fixed samples of size *n.*

A. xbar and r chart
B. xbar and s chart
C. NP chart
D. Y chart

14.  A _____is an attributes control chart used with data collected in subgroups that are the

   same size

A. xbar and r chart
B. xbar and s chart
C. C chart
D. Y chart


**Lovely Professional University**

*Probability and Statistics*

15. _____requires that each subgroup's sample size be the same and compute control limits

based on the Poisson distribution.

A. xbar and r chart
B. xbar and s chart
C. C chart
D. Y chart

## Answers for Self Assessment

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. | A | 2. | A | 3. | A | 4. | B | 5. | A |
| 6. | A | 7. | B | 8. | B | 9. | B | 10. | A |
| 11. | B | 12. | C | 13. | C | 14. | C | 15. | C |

## Review Questions

1) What is difference between SPC and SQC?
2) What are some of the benefits of SQC?
3) What does an X bar R chart tell you?
4) Why are X bar and R charts used together?
5) What is p-chart and NP chart?
6) Create a flow chart explaining conditions for different flow charts?
7) Why statistical process control is important in business
8) What are the main objectives of quality control?
9) What is difference between X bar R chart and X bar S chart?
10) In what scenario p and Np chart is used?

## Further Readings

An Introduction to Probability and Statistics

Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

First Course in Probability, A

Book by Sheldon M. Ross

Schaums Theory and Problems of Statistics

Book by Murray R. Spiegel

Introduction to Probability, Statistics, and Random ...

Book by Hossein Pishro-Nik

## Web Links

- https://www.tutorialspoint.com
- www.webopedia.com
- https://www.britannica.com/science/probability

# Unit 09: Index Numbers

## Objectives

After this unit Students will able to

- Understand basics of Index Numbers.
- Learn about features of Index Numbers.
- Understand construction of Index Numbers in statistics.
- Understand Consumer Price Index.
- Solve basic questions related to Index Numbers.

## Introduction

Meaning of Index Numbers:

The value of money does not remain constant over time. It rises or falls and is inversely related to the changes in the price level. A rise in the price level means a fall in the value of money and a fall in the price level means a rise in the value of money. Thus, changes in the value of money are reflected by the changes in the general level of prices over a period of time. Changes in the general level of prices can be measured by a statistical device known as 'index number.'

Index number is a technique of measuring changes in a variable or group of variables with respect to time, geographical location or other characteristics. There can be various types of index numbers, but, in the present context, we are concerned with price index numbers, which measures changes in the general price level (or in the value of money) over a period of time.

Price index number indicates the average of changes in the prices of representative commodities at one time in comparison with that at some other time taken as the base period. According to L.V. Lester, "An index number of prices is a figure showing the height of average prices at one time relative to their height at some other time which is taken as the base period."

Index number in statistics is the measurement of change in a variable or variables across a determined period. It will show general relative change and not a directly measurable figure. An index number is expressed in percentage form.

---

Importance of Index Number

Index numbers occupy an important place due to its efficacy in measuring the extent of economic changes across a stipulated period. It helps to study such changes' effects due to factors that cannot be directly measured.

---

## 9.1  Characteristics of Index Numbers



The main features of index numbers are –

It is a special category of average for measuring relative changes in such instances where absolute measurement cannot be undertaken

Example: Index number only shows the tentative changes in factors that may not be directly measured. It gives a general idea of the relative changes

The method of index number measure alters from one variable to another related variable

It helps in the comparison of the levels of a phenomenon concerning a specific date and to that of a previous date

It is representative of a special case of averages especially for a weighted average

Index numbers have universal utility. The index that is used to ascertain the changes in price can also be used for industrial and agricultural production.

## 9.2   Types of Index Numbers
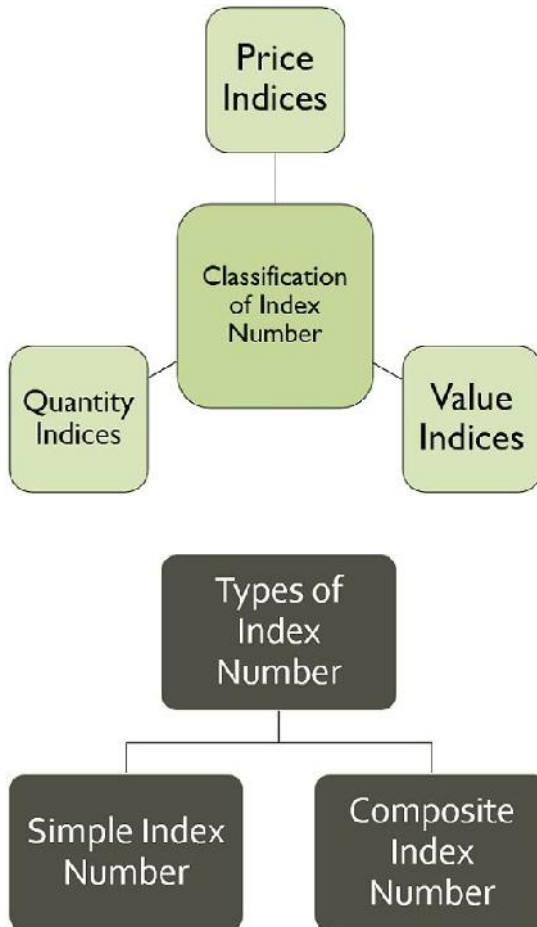




📋 Task:  What are different features of  Index numbers ?

**Its major types are –**

**Value Index:** A value index number is formed from the ratio of the aggregate value for a particular period with that of the aggregate value that is found in the base period. The value index is utilised in for inventories, sales and foreign trade, among others.

**Quantity Index:** A quantity index number is used to measure changes in the volume or quantity of goods that are produced, consumed and sold within a stipulated period. It shows the relative change across a period for particular quantities of goods. Index of Industrial Production (IIP) is an example of Quantity Index.

**Price Index:** A price index number is used to measure how price alters across a period. It will indicate the relative value and not the absolute value. The Consumer Price Index (CPI) and Wholesale Price Index (WPI) are major examples of a price index.

## 9.3   Uses of Index Number in Statistics

It helps in measuring changes in the standard of living as well as the price level.

Wage rate regulation is consistent with the changes in the price level. With the determination of price levels, wage rates may be revised.

*Probability and Statistics*

Government policies are framed following the index number of prices. This price stability inherent to fiscal and economic policies is based on index numbers.

It gives a pointer for international comparison concerning different economic variables—for instance, living standards between two countries.

## 9.4   Advantages of Index Number

It adjusts primary data at varying costs, which is useful for deflating. It facilitates the transformation from nominal wage to real wage.

Index numbers find extensive usage in economics and help in the framing of appropriate policies. Such findings help with the establishment of researches as well.

It helps in case of trends such as drawing outcomes for irregular forces and cyclical forces.

Index number can be leveraged in case of future development of activities in the economic sphere. This time series analysis is utilized for the determination trends and cyclical developments.

The number is useful in measuring the changes that take place in the standard of living in different countries over an established period.

## 9.5   Limitations and Features  of Index Number

There are chances for errors given that index numbers come as a result of samples. These samples are put together after deliberation, which creates chances for errors. It can also be found in weights or base periods etc.

It is always calculated based on items. Items that are so selected may not exactly be in trend, which in turn creates an inaccurate analysis.

Multiple methods can be used to formulate index numbers. Due to this multiplicity of methods, outcomes may bring forward a different set of values which may further lead to confusion.

The index numbers show the approximate indications of the relative changes that occur. Moreover, the changes in variables that are compared over a prolonged time may fall short on reliability.

The selection of representative commodities may be skewed. It is since these commodities are based on samples.

## 9.6   Features of Index Numbers

The following are the main features of index numbers:

(i) Index numbers are a special type of average. Whereas mean, median and mode measure the absolute changes and are used to compare only those series which are expressed in the same units, the technique of index numbers is used to measure the relative changes in the level of a phenomenon where the measurement of absolute change is not possible and the series are expressed in different types of items.

(ii) Index numbers are meant to study the changes in the effects of such factors which cannot be measured directly. For example, the general price level is an imaginary concept and is not capable of direct measurement. But, through the technique of index numbers, it is possible to have an idea of relative changes in the general level of prices by measuring relative changes in the price level of different commodities.

iii) The technique of index numbers measures changes in one variable or group of related variables. For example, one variable can be the price of wheat, and group of variables can be the price of sugar, the price of milk and the price of rice.

(iv) The technique of index numbers is used to compare the levels of a phenomenon on a certain date with its level on some previous date (e.g., the price level in 1980 as compared to that in 1960 taken as the base year) or the levels of a phenomenon at different places on the same date (e.g., the price level in India in 1980 in comparison with that in other countries in 1980).

Steps or Problems in the Construction of Price Index Numbers:

The construction of the price index numbers involves the following steps or problems:

1. Selection of Base Year: The first step or the problem in preparing the index numbers is the selection of the base year. The base year is defined as that year with reference to which the price changes in other years are compared and expressed as percentages. The base year should be a normal year.

Notes: What is Base Period?

When you have two given periods, the period with which a comparison is made, is the base period.

An index number begins in a specific year called the base year or reference year, whose value is 100. And in the following years, the percentage increases tend to shift the index number above or below its base value, i.e. 100. This means that, if the index number for a year is 105, it reflects a 5% rise from the base year, whereas when the index number is 95, it signifies a 5% fall from its base year value

In other words, it should be free from abnormal conditions like wars, famines, floods, political instability, etc. Base year can be selected in two ways-

(a) through fixed base method in which the base year remains fixed; and

(b) through chain base method in which the base year goes on changing, e.g., for 1980 the base year will be 1979, for 1979 it will be 1978, and so on.

2. Selection of Commodities: The second problem in the construction of index numbers is the selection of the commodities. Since all commodities cannot be included, only representative commodities should be selected keeping in view the purpose and type of the index number.

In selecting items, the following points are to be kept in mind:

(a) The items should be representative of the tastes, habits and customs of the people.

(b) Items should be recognizable,

(c) Items should be stable in quality over two different periods and places.

(d) The economic and social importance of various items should be considered

(e) The items should be fairly large in number.

(f) All those varieties of a commodity which are in common use and are stable in character should be included.

3. Collection of Prices: After selecting the commodities, the next problem is regarding the collection of their prices:

(a) From where the prices to be collected;

(b) Whether to choose wholesale prices or retail prices.

(c) Whether to include taxes in the prices or not etc.

While collecting prices, the following points are to be noted:

a) Prices are to be collected from those places where a particular commodity is traded in large quantities.

(b) Published information regarding the prices should also be utilized,

(c) In selecting individuals and institutions who would supply price quotations, care should be taken that they are not biased.

*Probability and Statistics*

(d) Selection of wholesale or retail prices depends upon the type of index number to be prepared. Wholesale prices are used in the construction of general price index and retail prices are used in the construction of cost-of-living index number.

(e) Prices collected from various places should be averaged.

4. Selection of Average: Since the index numbers are, a specialized average, the fourth problem is to choose a suitable average. Theoretically, geometric mean is the best for this purpose. But, in practice, arithmetic mean is used because it is easier to follow.

5. Selection of Weights: Generally, all the commodities included in the construction' of index numbers are not of equal importance. Therefore, if the index numbers are to be representative, proper weights should be assigned to the commodities according to their relative importance.

For example, the prices of books will be given more weightage while preparing the cost-of-living index for teachers than while preparing the cost-of-living index for the workers. Weights should be unbiased and be rationally and not arbitrarily selected.

6. Purpose of Index Numbers:The most important consideration in the construction of the index numbers is the objective of the index numbers. All other problems or steps are to be viewed in the light of the purpose for which a particular index number is to be prepared. Since, different index numbers are prepared with specific purposes and no single index number is 'all purpose' index number, it is important to be clear about the purpose of the index number before its construction.

7. Selection of Method: The selection of a suitable method for the construction of index numbers is the final step.

There are two methods of computing the index numbers:

(a) Simple index number and

(b) Weighted index number.

Simple index number again can be constructed either by – (i) Simple aggregate method, or by (ii) simple average of price relative's method. Similarly, weighted index number can be constructed either by (i) weighted aggregative method, or by (ii) weighted average of price relative's method. The choice of method depends upon the availability of data, degree of accuracy required and the purpose of the study.

## 9.7 Construction of Price Index Numbers (Formula and Examples):

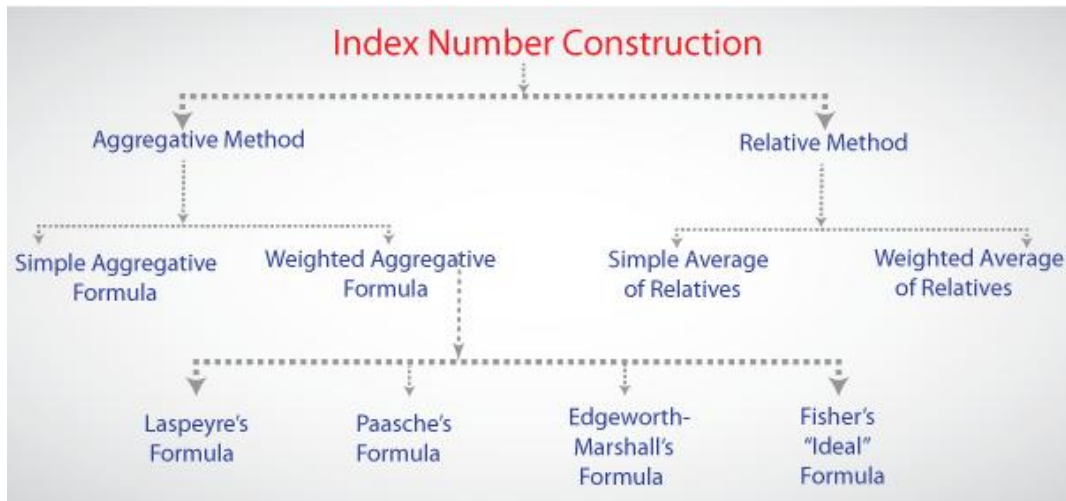Construction of price index numbers through various methods can be understood with the help of the following examples:

1. Simple Aggregative Method: In this method, the index number is equal to the sum of prices for the year for which index number is to be found divided by the sum of actual prices for the base year.

The formula for finding the index number through this method is as follows:

Task: What is Purpose of Index Numbers?

## Index Number Construction

```
                    Index Number Construction
            ┌──────────────────────────┴──────────────────────────┐
      Aggregative Method                                    Relative Method
    ┌───────────┴───────────┐                       ┌───────────┴───────────┐
Simple Aggregative   Weighted Aggregative     Simple Average        Weighted Average
   Formula               Formula              of Relatives          of Relatives
           ┌───────────┬──────────────────────────┬───────────┐
      Laspeyre's    Paasche's          Edgeworth-        Fisher's
       Formula       Formula           Marshall's        "Ideal"
                                        Formula          Formula
```

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

Where $P_{01}$ Stands for the index number

$\Sigma P_1$ Stands for the sum of the prices for the year for which index number is to be found :

$\Sigma P_0$ Stands for the sum of prices for the base year.

| Commodity | Prices in Base Year 1980 (in Rs.) $P_0$ | Prices in current Year 1988 (in Rs.) $P_1$ |
|---|---|---|
| A | 10 | 20 |
| B | 15 | 25 |
| C | 40 | 60 |
| D | 25 | 40 |
| Total | $\Sigma P_0 = 90$ | $\Sigma P_1 = 145$ |

Index Number $(P_{01}) = \dfrac{\Sigma P_1}{\Sigma P_0} \times 100$ ; $P_{01} = \dfrac{145}{90} \times 100$ ; $P_{01} = 161.11$

2. Simple Average of Price Relatives Method:In this method, the index number is equal to the sum of price relatives divided by the number of items and is calculated by using the following formula:

$$P_{01} = \frac{\Sigma R}{N}$$

Where $\Sigma R$ stands for the sum of price relatives i. e. $R = \dfrac{P_1}{P_0} \times 100$ and

N stands for the number of items.

**Example**

| Commodity | Base Year Prices (in Rs.) $P_0$ | Current year Prices (in Rs.) $P_1$ | Price Relatives $R = \dfrac{P_1}{P_0} \times 100$ |
|---|---|---|---|
| A | 10 | 20 | $\dfrac{20}{10} \times 100 = 200.0$ |
| B | 15 | 25 | $\dfrac{25}{15} \times 100 = 166.7$ |
| C | 40 | 60 | $\dfrac{60}{40} \times 100 = 150.00$ |
| D | 25 | 40 | $\dfrac{40}{25} \times 100 = 160.0$ |
| N = 4 | | | $\Sigma R = 676.7$ |

*Probability and Statistics*

3. Weighted Aggregative Method:In this method, different weights are assigned to the items according to their relative importance. Weights used are the quantity weights.

Many formulae have been developed to estimate index numbers on the basis of quantity weights.

(i) **Laspeyre's Formula.** In this formula, the quantities of base year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

Where $P_1$ is the price in the current year ; $P_0$ is the price in the base year ; and $q_0$ is the quantity in the base year.

(ii) **Paasche's Formula.** In this formula, the quantities of the current year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

Where $q_1$ is the quantity in the current year.

(iii) **Dorbish and Bowley's Formula.** Dorbish and Bowley's formula for estimating weighted index number is as follows :

$$P_{01} = \frac{\dfrac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \dfrac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 \quad \text{or} \quad P_{01} = \frac{L + P}{2}$$

Where L is Laspeyre's index and P is paasche's Index.

(iv) **Fisher's Ideal Formula.** In this formula, the geometric mean of two indices (i.e., Laspeyre's Index and paasche's Index) is taken :

$$P_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P} \times 100$$

where L is Lespeyre's Index and P is paasche's Index.

**Example**

| Comm- odity | Base Year | | Current Year | | $P_0 q_0$ | $P_1 q_0$ | $P_0 q_1$ | $P_1 q_1$ |
|---|---|---|---|---|---|---|---|---|
| | $P_0$ | $q_0$ | $P_1$ | $q_1$ | | | | |
| A | 10 | 5 | 20 | 2 | 50 | 100 | 20 | 40 |
| B | 15 | 4 | 25 | 8 | 60 | 100 | 120 | 200 |
| C | 40 | 2 | 60 | 6 | 80 | 120 | 240 | 360 |
| D | 25 | 3 | 40 | 4 | 75 | 120 | 100 | 160 |
| Total | | | | | 265 $\Sigma P_0 q_0$ | 440 $\Sigma P_1 q_0$ | 480 $\Sigma P_0 q_1$ | 760 $\Sigma P_1 q_1$ |

(i) Laspeyre's Formula :

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

$$P_{01} = \frac{440}{265} \times 100 = 166.04$$

Example:

(ii) Paasche' Formula :

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

$$P_{01} = \frac{700}{480} \times 100 = 158.3$$

(iii) Dorbish and Bowley's Formula :

$$P_{01} = \frac{\dfrac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \dfrac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 = 162.2$$

$$P_{01} = \frac{\dfrac{440}{265} + \dfrac{760}{480}}{2} \times 100 = 162$$

(iv) Fisher's Ideal Formula :

$$P_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100$$

$$P_{01} = \sqrt{\frac{440}{265} \times \frac{760}{480}} \times 100 = 162.1$$

4. Weighted Average of Relatives Method:In this method also different weights are used for the items according to their relative importance.

The price index number is found out with the help of the following formula:

$$P_{01} = \frac{\Sigma RW}{\Sigma W}$$

where $\Sigma W$ stands for the sum of weights of different commodities :
and $\Sigma R$ stands for the sum of price relatives.

| Commodity | Weights W | Base Prices Year $P_0$ | Current Year Prices $P_1$ | Price Relatives $R = \frac{P_1}{P_0} \times 100$ | RW |
|---|---|---|---|---|---|
| A | 5 | 10 | 20 | $20/10 \times 100 = 200.0$ | 1000.0 |
| B | 4 | 15 | 25 | $25/15 \times 100 = 166.7$ | 666.8 |
| C | 2 | 40 | 60 | $60/40 \times 100 = 150.0$ | 300.0 |
| D | 3 | 25 | 40 | $40/25 \times 100 = 160.0$ | 480.0 |
| Total | $\Sigma W=14$ | | | | $\Sigma RW = 2446.8$ |

Index Number $(P_{01}) = \frac{\Sigma RW}{\Sigma W}$

$$P_{01} = \frac{2446.8}{14} = 174.8$$

## 9.8 Difficulties in Measuring Changes in Value of Money:

Measurement of changes in the value of money through price index number is not an easy and reliable technique. There are a number of theoretical as well as practical difficulties in the construction of price index numbers. Moreover, the index number technique itself has many limitations.

(A) Conceptual Difficulties: The following are the conceptual difficulties during the construction of price index numbers:

1. Vague Concept of Value of Money: The concept of money is vague, abstract and cannot be clearly defined. The value of money is a relative concept which changes from person to person depending upon the type of goods on which the money is spent.

2. Inaccurate Measurement: Price index numbers do not measure the changes in the value of money accurately and reliably. A rise or fall in the general level of prices as indicated by the price index numbers does not mean that the price of every commodity has risen or fallen to the same extent.

3. Reflect General Changes: Price index numbers are averages and measure general changes in the value of money on the average. Therefore, they are not of much significance for the particular individuals who may be affected by the changes in the actual prices quite differently from that indicated by the index numbers.

4. Limitations of Wholesale Price Index: The wholesale price index numbers, which are generally used to measure changes in the value of money, suffer from certain limitations:

(a) They do not reflect the changes in the cost of living because retail prices are generally higher than the wholesale prices.

(b) They ignore some of the important items concerning the urban population, such as, expenditure on education, transport, house rent, etc.

(c) They do not take into consideration the changes in the consumers' preferences.

(B) Practical Difficulties: The practical difficulties in the way of constructing price index numbers, and therefore, in measuring changes in the value of money are as follows:

*Probability and Statistics*

1. Selection of Base Year:While preparing the index number, first difficulty arises regarding the selection of base year. The base year should be a normal year. But, it is very difficult to find out a fully normal year free from any unusual happening. There is every possibility that the selected base year may be an abnormal year, or a distant year, or may be selected by an immature or biased person.

2. Selection of Items: The selection of the representative commodities is the second difficulty in the construction of index numbers:

(a) With the passage of time the quality of the product may change ; if the quality of a product changes in the year of enquiry from what it was in the base year, the product becomes irrelevant,

(b) The relative importance of certain commodities may change due to a change in the consumption pattern of the people in the course of time; for example, Vanaspati Ghee was not an important item of consumption in India in the pre-war period, but today it has become an item of necessity. Under such conditions, it is not easy to select the appropriate commodities.

3. Collection of Prices: It is also difficult to obtain correct, adequate and representative data regarding prices. It is not an easy job to select representative places from which the information about prices to be collected and to select the experienced and unbiased individuals or institutions who will supply price quotations. Moreover, there is the problem of deciding which prices (wholesale or retail) are to be taken into consideration. It is comparatively easy to get information about wholesale prices which vary considerably.

4. Assigning Weights: Another important difficulty that arises in preparing the index numbers is that of assigning proper weights to different items in order to arrive at correct and unbiased conclusions. As there are no hard and fast rules to weights for the commodities according to their relative importance, there is very likelihood that the weights are decided arbitrarily on the basis of personal judgment and involve biasness.

5. Selection of Averages: Another major problem is that which average should be employed to find out the price relatives. There are many types of averages such as arithmetic average, geometric average, mean, median, mode, etc. The use of different averages gives different results. Therefore, it is essential to select the method with great care. Dr. Marshall has advocated the use of chain index number to solve the problem of averaging and weighing.

6. Problem of Dynamic Changes: In the dynamic world, the consumption pattern of the individuals and the number and varieties of goods undergo continuous changes.

They create difficulties for preparing index numbers and making temporal comparisons:

(a) Since, in the course of time, old commodities may disappear and many new ones come into existence, the long-run comparison may become difficult,

(b) The quantity and quality of commodities may also change over the period of time, thus making the choice of commodities for constructing index numbers difficult,

(c) A number of factors, like income, education, fashion, etc., bring changes in the consumption pattern of the people which render the index numbers incomparable.

More Types of Index Numbers:

Index numbers are of different types.

Important types of index numbers are discussed below:

1. Wholesale Price Index Numbers:Wholesale price index numbers are constructed on the basis of the wholesale prices of certain important commodities. The commodities included in preparing these index numbers are mainly raw-materials and semi-finished goods. Only the most important

and most price-sensitive and semi- finished goods which are bought and sold in the wholesale market are selected and weights are assigned in accordance with their relative importance.

The wholesale price index numbers are generally used to measure changes in the value of money. The main problem with these index numbers is that they include only the wholesale prices of raw materials and semi-finished goods and do not take into consideration the retail prices of goods and services generally consumed by the common man. Hence, the wholesale price index numbers do not reflect true and accurate changes in the value of money.

2. Retail Price Index Numbers:These index numbers are prepared to measure the changes.in the value of money on the basis of the retail prices of final consumption goods. The main difficulty with this index number is that the retail price for the same goods and for continuous periods is not available. The retail prices represent larger and more frequent fluctuations as compared to the wholesale prices.

3. Cost-of-Living Index Numbers:These index numbers are constructed with reference to the important goods and services which are consumed by common people. Since the number of these goods and services is very large, only representative items which form the consumption pattern of the people are included. These index numbers are used to measure changes in the cost of living of the general public.

4. Working Class Cost-of-Living Index Numbers: The working class cost-of-living index numbers aim at measuring changes in the cost of living of workers. These index numbers are consumed on the basis of only those goods and services which are generally consumed by the working class. The prices of these goods and index numbers are of great importance to the workers because their wages are adjusted according to these indices.

5. Wage Index Numbers:The purpose of these index numbers is to measure time to time changes in money wages. These index numbers, when compared with the working class cost-of-living index numbers, provide information regarding the changes in the real wages of the workers.

6. Industrial Index Numbers:Industrial index numbers are constructed with an objective of measuring changes in the industrial production. The production data of various industries are included in preparing these index numbers.

## 9.9 Importance of Index Numbers

Index numbers are used to measure all types of quantitative changes in different fields.

Various advantages of index numbers are given below:

1. General Importance:In general, index numbers are very useful in a number of ways:

(a) They measure changes in one variable or in a group of variables.

(b) They are useful in making comparisons with respect to different places or different periods of time,

(c) They are helpful in simplifying the complex facts.

(d) They are helpful in forecasting about the future,

(e) They are very useful in academic as well as practical research.

2. Measurement of Value of Money:Index numbers are used to measure changes in the value of money or the price level from time to time. Changes in the price level generally influence production and employment of the country as well as various sections of the society. The price index numbers also forewarn about the future inflationary tendencies and in this way, enable the government to take appropriate anti- inflationary measures.

*Probability and Statistics*

3. Changes in Cost of Living: Index numbers highlight changes in the cost of living in the country. They indicate whether the cost of living of the people is rising or falling. On the basis of this information, the wages of the workers can be adjusted accordingly to save the wage earners from the hardships of inflation.

4. Changes in Production:Index numbers are also useful in providing information regarding production trends in different sectors of the economy. They help in assessing the actual condition of different industries, i.e., whether production in a particular industry is increasing or decreasing or is constant.

5. Importance in Trade:Importance in trade with the help of index numbers, knowledge about the trade conditions and trade trends can be obtained. The import and export indices show whether foreign trade of the country is increasing or decreasing and whether the balance of trade is favorable or unfavorable.

6. Formation of Economic Policy: Index numbers prove very useful to the government in formulating as well as evaluating economic policies. Index numbers measure changes in the economic conditions and, with this information, help the planners to formulate appropriate economic policies. Further, whether particular economic policy is good or bad is also judged by index numbers.

7. Useful in All Fields:Index numbers are useful in almost all the fields. They are especially important in economic field.

Some of the specific uses of index numbers in the economic field are:

(a) They are useful in analyzing markets for specific commodities.

(b) In the share market, the index numbers can provide data about the trends in the share prices,

(c) With the help of index numbers, the Railways can get information about the changes in goods traffic.

(d) The bankers can get information about the changes in deposits by means of index numbers.

## 9.10  Limitations of Index Numbers

Index number technique itself has certain limitations which have greatly reduced its usefulness:

(i) Because of the various practical difficulties involved in their computation, the index numbers are never cent per cent correct.

(ii) There are no all-purpose index numbers. The index numbers prepared for one purpose cannot be used for another purpose. For example, the cost-of-living index numbers of factory workers cannot be used to measure changes in the value of money of the middle income group.

(iii) Index numbers cannot be reliably used to make international comparisons. Different countries include different items with different qualities and use different base years in constructing index numbers.

(iv) Index numbers measure only average change and indicate only broad trends. They do not provide accurate information.

(v) While preparing index numbers, quality of items is not considered. It may be possible that a general rise in the index is due to an improvement in the quality of a product and not because of a rise in its price.

The Criteria of a Good Index Number

- A number of mathematical test discussed below have been suggested for comparing various index numbers.

- **Unit Test:** This test requires the index numbers to be independent of the units in which prices and quantities are quoted. This test is satisfied by all the formulas.

- **Time Reversal Test:** This is one of the two very important test proposed by Irving Fisher as tests of consistency for a good index number. According to this 'the formula for calculating an index number should be such that it will give the same ratio between one point of comparison and the other no matter which of the two is taken as base

**Factor Reversal Test**

- This is the second test of consistency suggested by I. Fisher.

- In his words: "Just as our formula should permit the interchange of two items without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results – i.e. the two results multiplied together should give the true value ratio, except for a constant of proportionality".

**Consumer Price Index**

- Price index number is the measure of relative changes in the prices of some commodity over a period of time. In practice these indices are used for various different purposes. One very important use of the theory of index number is in obtaining consumer price index or alternatively also called cost of living index number.

- It is a well-known fact that the prices of the commodities required for day to day living go on increasing, e.g. prices of food items like wheat, rice, oil etc. are different in different years. This increase (or decrease, if there is any) in prices of commodities directly hit the purchasing power of consumer. A consumer price index is, therefore, devised to measure the overall changes in the purchasing power of the consumer.

- Aconsumer price index or cost of living index, is a measure which indicates the relative changes in the prices of a group of items, necessary for the living for a selected group of consumers. In a way, it tells us about what should be the increase in the wages of consumer so that they are able to maintain some standard of living in two time periods. For this purpose, the total expenditure of a household are categorized like food, clothing, rent, electricity, entertainment, education, medicines, miscellaneous etc.

Various kinds of indexes or index numbers are used by the scientific and academic communities, and the popular media. In fact, some indexes are so commonly used that one does not even recognize that they are index numbers, not absolute values of the variable or the item of interest. Changes in stock prices associated with several stock markets, for example, are quoted in terms of index numbers. Thus, when the media reports that the Dow fell by 200 points on a particular day, the reference is to the fall in the value of the index representing 30 industrial stocks included in the Dow Jones Industrial Average index, sometimes called the Dow 30. Suppose that the Dow Jones Industrial Average was at the 8,000 mark before it fell by 200 points. This implies that the average price of the 30 stocks included in the Dow index fell by 2.5 percent in one day ([200/8,000] * 100 = 2.5 percent). In order to understand index numbers, it is useful to have some idea of the methods of constructing indexes, and to be able to relate these methods to some common examples of index numbers.

The widespread use of indexes in

Business and economics

A wide variety of sources generate the various indexes in the fields of business and economics. Many government agencies in the United States regularly produce information in index number form on a variety of variables. The Bureau of Labor Statistics under the U.S. Department of Labor reports data on various price indexes that are used to measure the inflation rate in the economy. The U.S. Department of Commerce also reports economic data on a regular basis. In addition to government sources, economic data are also reported (many of them in index number form) by

**Lovely Professional University**

*Probability and Statistics*

private sources and by partisan and nonpartisan research organizations. Most business-related indexes can be categorized within two broad categories—those associated with the financial markets and those that describe the state of the economy.

Indexes associated with the financial markets.

There are many indexes associated with financial markets. These indexes measure different attributes of the financial markets at various degrees of depth and rigor. The majority of them track two major components of the financial markets: stock markets and bond markets.

Several index numbers measure the changes in stock prices at different levels of aggregation (that is, with respect to the number of stocks included in an index). The Dow Jones Industrial Average (DJIA), an index of 30 industrial stocks, is one of the most commonly followed stock price indexes. The DJIA includes major U.S. industrial stocks (such as Coca-Cola, IBM, General Motors, du Pont, Eastman Kodak, Disney, Sears, Goodyear, Merck, and AT&T) listed on the New York Stock Exchange (NYSE). There are also index numbers that represent stock price changes at a particular stock exchange or a stock-trading network. Thus, pertaining to the three trading mediums NYSE based in New York, the American Stock Exchange also based in New York (AMEX), and the computer-linked stock trading network of the National Securities Dealers Association (NASDAQ)—there are separate indexes called NYSE, AMEX, and NASDAQ, respectively. In addition to these so-called exchange-based indexes, there are several other stock price indexes that include an increasing number of stocks. For example, the S&P100 index, maintained by Standard & Poor's, measures stock price movements of 100 important stocks. Similarly, the S&P500 index measures the change in the aggregate price level of 500 selected stocks. The Wilshire 5000 index measures the change in the aggregate price level of the 5,000 stocks it monitors—a large number of stocks are included in this index to give a better picture of the overall stock market, rather than a narrow group of stocks.

By contrast, the bond market has far fewer indexes. The most widely used bond price index is the Lehman Brothers' bond price index; it measures changes in prices of long-term bonds included in the index. Changes in bond prices convey information regarding changes in interest rates. Thus, the bond price index is also closely watched by financial market participants.

Indexes associated with the economy.

Information and data on a large number of economic variables are regularly reported. Some of these data are reported in index number form. All price data in the economy are reported in index number form—as will be explained later, this is out of necessity, rather than choice. Three major price indexes are: Consumer Price Index, Producer Price Index, and Implicit Gross Domestic Product Price Deflator. Similarly, the Federal Reserve Bank computes the Index of Industrial Production on a monthly basis to gauge the pace of industrial production.

Other regularly reported indexes that describe the state of the economy (current or future) include: Index of Leading Indicators (LEI), which indicates the pace of economic activity in the economy in the near future; Consumer Confidence Index, which captures consumer sentiment and thus suggests consumers' willingness to spend; the Housing Affordability Index, which tracks the cost of being able to afford a home. Some indexes are not reported regularly, but occasionally crop up. For example, George Bush during his reelection campaign kept referring to the Misery Index, which captures the combined effects of inflation and unemployment.

## 9.11  The need for an Index

Most economic variables are measured in absolute terms. For example, 12 million cars were produced in 1994 in the United States, or the gross output of goods and services in the United States was estimated at $8.7 trillion during 1998. It is not possible, however, to measure the price associated with a group of commodities in absolute terms; it can only be done so long as we refer to one commodity. We thus can say that the average price of a loaf of bread in the United States was $1.75 in 1998. But when dealing with a number of commodities together (as in the calculation of an individual's cost of living from year-to-year, one cannot simply compute the average price of all the goods and services bought in each year. Since most people buy a different set of commodities each year, the resulting averages would not be comparable. An index helps us out of this quandary. In

the most basic terms, an index usually attaches weights to the prices of items in order to track the resulting collective price movement.

## Summary

- The value of money does not remain constant over time. It rises or falls and is inversely related to the changes in the price level. A rise in the price level means a fall in the value of money and a fall in the price level means a rise in the value of money.
- Index number is a technique of measuring changes in a variable or group of variables with respect to time, geographical location or other characteristics.
- Price index number indicates the average of changes in the prices of representative commodities at one time in comparison with that at some other time taken as the base period
- Index number in statistics is the measurement of change in a variable or variables across a determined period. It will show general relative change and not a directly measurable figure. An index number is expressed in percentage form
- It is representative of a special case of averages especially for a weighted average
- Index numbers have universal utility. The index that is used to ascertain the changes in price can also be used for industrial and agricultural production.

## Keywords

- It is a special category of average for measuring relative changes in such instances where absolute measurement cannot be undertaken
- Index number only shows the tentative changes in factors that may not be directly measured. It gives a general idea of the relative changes
- The method of index number measure alters from one variable to another related variable
- It helps in the comparison of the levels of a phenomenon concerning a specific date and to that of a previous date.
- A value index number is formed from the ratio of the aggregate value for a particular period with that of the aggregate value that is found in the base period. The value index is utilised in for inventories, sales and foreign trade, among others.
- A quantity index number is used to measure changes in the volume or quantity of goods that are produced, consumed and sold within a stipulated period. It shows the relative

## Self Assessment

1. _____ are a special type of average

A. Whole Numbers
B. Index Number
C. Natural Numbers
D. None of these

2. _____ indicates the average of changes in the prices of representative commodities at one time in comparison with that at some other time taken as the base period.

A. Price index number
B. Index Number

C. Natural Numbers

D. None of these

3. A _____number is formed from the ratio of the aggregate value for a particular period with that of the aggregate value that is found in the base.

A. Price index number

B. Value Index Number

C. Natural Numbers

D. None of these

4. A _____number is used to measure changes in the volume or quantity of goods.

A. Quantity index Numbers

B. Value Index Number

C. Natural Numbers

D. None of these

5. CPI full form is

A. Consumer Price Index

B. Customer Price Index

C. Consumer Proper Index

D. None of these

6. WPI full form is

A. Wholesale Price Index

B. While Price Index

C. Wright  Price Index

D. None of These

7. Under this type of index, the quantities in the base year are the values of weights.

A. Laspeyres Index

B. . Passche's Index

C. Marshall-Edgeworth Index

D. Fisher's Ideal Price Index

8. Under this type of Index, the quantities in the current year are the values of weights.

A. Laspeyres Index

B. Passche's Index

C. Marshall-Edgeworth Index

D. Fisher's Ideal Price Index

9. Under this type of index, we take both i.e. the current year as well as the base year into consideration for specifying the methods.

A. Laspeyres Index

B. Passche's Index

C. Marshall-Edgeworth Index

D. Fisher's Ideal Price Index

10. It gives a pointer for international comparison concerning different economic variables.

A. Variance

B. Standard deviation

C. Index Number

D. Mean

11. Index number for base year is always considered as------

A. 100

B. 101

C. 201

D. 1000

12. Index number is also called as-----------

A. Economic barometer

B. Parameter

C. Constant

D. None of the above

13. The geometric mean of Laspeyres' and Paasche's is

a. Laspeyres Index

b. . Passche's Index

c. Marshall-Edgeworth Index

d. Fisher's Ideal Price Index

14. This test requires the index numbers to be independent of the units in which prices and quantities are quoted.

a. Unit Test

b. Time reversal Test

c. Factor reversal test

d. None of these

15. According to this test, formula for calculating an index number should be such that it will give the same ratio between one point of comparison and the other no matter which of the two is taken as base

a. Unit Test

b. Time reversal Test

c. Integration test

d. None of these

## Answers for Self Assessment

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. | B | 2. | A | 3. | B | 4. | A | 5. | A |
| 6. | A | 7. | A | 8. | B | 9. | C | 10. | C |
| 11. | C | 12. | A | 13. | D | 14. | A | 15. | B |

## Review Questions

1) What do you mean by index number?
2) What is index number and its types?
3) Which is the ideal method to find index number?
4) What is the most commonly used index number?
5) What is index number what is its formula?
6) What is the index number for base year?
7) What are use of Index number in business and applications?
8) What is difference between Consumer Price index vs. Quantity index?

## Further Readings

- An Introduction to Probability and Statistics

- Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

- First Course in Probability, ABook by Sheldon M. Ross
- Schaums Theory and Problems of Statistics Book by Murray R. Spiegel
- Introduction to Probability, Statistics, and Random ...Book by Hossein Pishro-Nik

## Web Links

- https://www.tutorialspoint.com
- www.webopedia.com
- https://www.britannica.com/science/probability

# Unit 10: Time Series

## Objective

**Students will able to:**

- Understand concept of time series data
- UnderstandMeasurement of Time series.
- Solve problems related to time series data
- Compare time series data with cross-sectional data

## Introduction

**What is a Time Series?**

A time series is a sequence of data points that occur in successive order over some period of time. This can be contrasted with cross-sectional data, which captures a point-in-time.

In investing, a time series tracks the movement of the chosen data points, such as a security's price, over a specified period of time with data points recorded at regular intervals. There is no minimum or maximum amount of time that must be included, allowing the data to be gathered in a way that provides the information being sought by the investor or analyst examining the activity.

In investing, a time series tracks the movement of the chosen data points, such as a security's price, over a specified period of time with data points recorded at regular intervals. There is no minimum or maximum amount of time that must be included, allowing the data to be gathered in a way that provides the information being sought by the investor or analyst examining the activity.

## 10.1  What is Time Series Analysis?

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time. What sets time series data apart from other data is that the analysis can show how variables change over time. In other words, time

is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data. Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting—predicting future data based on historical data.

Why organizations use time series data analysis

Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time. Using data visualizations, business users can see seasonal trends and dig deeper into why these trends occur. With modern analytics platforms, these visualizations can go far beyond line graphs. When organizations analyze data over consistent intervals, they can also use time series forecasting to predict the likelihood of future events. Time series forecasting is part of predictive analytics. It can show likely changes in the data, like seasonality or cyclic behavior, which provides a better understanding of data variables and helps forecast better. For example, Des Moines Public Schools analyzed five years of student achievement data to identify at-risk students and track progress over time. Today's technology allows us to collect massive amounts of data every day and it's easier than ever to gather enough consistent data for comprehensive analysis.

When time series analysis is used and when it isn't

Time series analysis is not a new study, despite technology making it easier to access. Many of the recommended texts teaching the subject's fundamental theories and practices have been around for several decades. And the method itself is even older than that. We have been using time series analysis for thousands of years, all the way back to the ancient studies of planetary movement and navigation. Time series analysis is used for non-stationary data—things that are constantly fluctuating over time or are affected by time. Industries like finance, retail, and economics frequently use time series analysis because currency and sales are always changing. Stock market analysis is an excellent example of time series analysis in action, especially with automated trading algorithms. Likewise, time series analysis is ideal for forecasting weather changes, helping meteorologists predict everything from tomorrow's weather report to future years of climate change. Examples of time series analysis in action include:

- Weather data
- Rainfall measurements
- Temperature readings
- Heart rate monitoring (EKG)
- Brain monitoring (EEG)
- Quarterly sales
- Stock prices
- Automated stock trading
- Industry forecasts
- Interest rates

Because time series analysis includes many categories or variations of data, analysts sometimes must make complex models. However, analysts can't account for all variances, and they can't generalize a specific model to every sample. Models that are too complex or that try to do too many things can lead to lack of fit. Lack of fit or over fitting models lead to those models not distinguishing between random error and true relationships, leaving analysis skewed and forecasts incorrect.

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

An observed time series can be decomposed into three components: the trend (long term direction),

the seasonal (systematic, calendar related movements) and the irregular (unsystematic, short term fluctuations).

## 10.2 What are Stock and Flow Series?

Time series can be classified into two different types: stock and flow.

A stock series is a measure of certain attributes at a point in time and can be thought of as "stock takes". For example, the **Monthly Lab our Force Survey** is a stock measure because it takes stock of whether a person was employed in the reference week.

Flow series are series which are a measure of activity over a given period. For example, surveys of **Retail Trade** activity. Manufacturing is also a flow measure because a certain amount is produced each day, and then these amounts are summed to give a total value for production for a given reporting period.

The main difference between a stock and a flow series is that flow series can contain effects related to the calendar (trading day effects). Both types of series can still be seasonally adjusted using the same seasonal adjustment process.

## 10.3 What Are Seasonal Effects?

A seasonal effect is a systematic and calendar related effect. Some examples include the sharp escalation in most Retail series which occurs around December in response to the Christmas period, or an increase in water consumption in summer due to warmer weather.

Other seasonal effects include trading day effects (the number of working or trading days in a given month differs from year to year which will impact upon the level of activity in that month) and moving holiday (the timing of holidays such as Easter varies, so the effects of the holiday will be experienced in different periods each year).

What Is Seasonal Adjustment And Why Do We Need It?

Seasonal adjustment is the process of estimating and then removing from a time series influences that are systematic and calendar related. Observed data needs to be seasonally adjusted as seasonal effects can conceal both the true underlying movement in the series, as well as certain non-seasonal characteristics which may be of interest to analysts.

Why Can't We Just Compare Original Data From The Same Period In Each Year?
A comparison of original data from the same period in each year does not completely remove all seasonal effects. Certain holidays such as Easter and Chinese New Year fall in different periods in each year, hence they will distort observations. Also, year to year values will be biased by any changes in seasonal patterns that occur over time. For example, consider a comparison between two consecutive March months i.e. compare the level of the original series observed in March for 2000 and 2001. This comparison ignores the moving holiday effect of Easter. Easter occurs in April for most years but if Easter falls in March, the level of activity can vary greatly for that month for some series. This distorts the original estimates. A comparison of these two months will not reflect the underlying pattern of the data. The comparison also ignores trading day effects. If the two consecutive months of March have different composition of trading days, it might reflect different levels of activity in original terms even though the underlying level of activity is unchanged. In a similar way, any changes to seasonal patterns might also be ignored.

Original estimates also contains the influence of the irregular component

If the magnitude of the irregular component of a series is strong compared with the magnitude of the trend component, the underlying direction of the series can be distorted.

However, the major disadvantage of comparing year to year original data, is lack of precision and

time delays in the identification of turning points in a series. Turning points occur when the direction of underlying level of the series changes, for example when a consistently decreasing series begins to rise steadily. If we compare year apart data in the original series, we may miss turning points occurring during the year.

📋 Task: What is difference between time series Vs. Crossectional data

💬 For example, if March 2001 has a higher original estimate than March 2000, by comparing these year apart values, we might conclude that the level of activity has increased during the year. However, the series might have increased up to September 2000 and then started to decrease steadily.

When is Seasonal Adjustment Inappropriate?

When a time series is dominated by the trend or irregular components, it is nearly impossible to identify and remove what little seasonality is present. Hence seasonally adjusting a non-seasonal series is impractical and will often introduce an artificial seasonal element.

Difference between Time Series and Cross Sectional Data

The **key difference** between time series and cross sectional data is that the time series data focuses on the same **variable** over a period of time while the cross sectional data focuses on several variables at the same point of time. Furthermore, the time series data consist of observations of a single subject at multiple time intervals whereas, the cross sectional data consist of observations of many subjects at the same point in time.

Fields such as Statistics, Econometrics gathers data and analyze them. Data is a vital aspect of activities such as for research, predictions and proving theories. There are various types of data. Two of them are time series and cross sectional data.
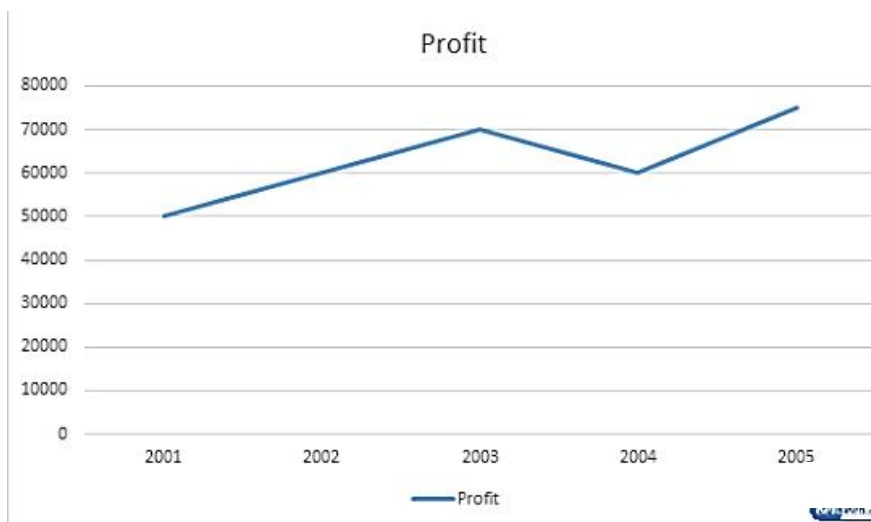
What is Time Series Data?

Time series data focuses on observations of a single individual at different times usually at uniform intervals. It is the data of the same variable over a period of time such as months, quarters, years etc. The time series data takes the form of Xt. The t represents the time. Below is an example of the profit of an organization over a period of 5 years' time. Profit is the variable that changes each year.

| Year | Profit |
|------|--------|
| 2001 | 50000  |
| 2002 | 60000  |
| 2003 | 70000  |
| 2004 | 60000  |
| 2005 | 75000  |

📋 Task: What is Seasonal Adjustment?

Usually, time series data is useful in business applications. Time measurement can be months, quarters or years but it can also be any time interval. Generally, the time has uniform intervals.

What is Cross Sectional Data?

In cross sectional data, there are several variables at the same point in time. Data set with maximum temperature, humidity, wind speed of few cities on a single day is an example of a cross sectional data.

| City | Maximum Temperature | Humidity | Wind Speed |
|---|---|---|---|
| City A | 29 | 60% | 20mph |
| City B | 27 | 65% | 26mph |
| City C | 30 | 60% | 21mph |

Another example is the sales revenue, sales volume, number of customers and expenses of an organization in the past month. Cross sectional data takes the form of Xi. Expanding the data from several months will convert the cross sectional data to time series data.

## 10.4 <u>What is the Difference between Time Series and Cross Sectional Data?</u>

Time series data consist of observations of a single subject at multiple time intervals. Cross sectional data consist of observations of many subjects at the same point in time. Time series data focuses on the same variable over a period of time. On the other hand, cross sectional data focuses on several variables at the same point in time. This is the main difference between time series and cross sectional data.

Profit of an organization over a period of 5 years' time is an example for a time series data while maximum temperature of several cities on a single day is an example for a cross sectional data.

**Lovely Professional University**

*Probability and Statistics*



Examples of time series data include:



## What is Seasonality?

The seasonal component consists of effects that are reasonably stable with respect to timing, direction and magnitude. It arises from systematic, calendar related influences such as:

Natural Conditions

Weather fluctuations that are representative of the season (uncharacteristic weather patterns such as snow in summer would be considered irregular influences)

Business and Administrative procedures

Start and end of the school term

Social and Cultural behavior

Christmas

It also includes calendar related systematic effects that are not stable in their annual timing or are caused by variations in the calendar from year to year, such as:

Trading Day Effects

The number of occurrences of each of the day of the week in a given month will differ from year to year there were 4 weekends in March in 2000, but 5 weekends in March of 2002

Moving Holiday Effects

Holidays which occur each year, but whose exact timing shifts Easter, Chinese New Year

How do We Identify Seasonality?

Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend. The following diagram depicts a strongly seasonal series. There is an obvious large seasonal increase in December retail sales in New South Wales due to Christmas shopping. In this example, the magnitude of the seasonal component increases over time, as does the trend.

Figure: Monthly Retail Sales in New South Wales (NSW) Retail Department Stores
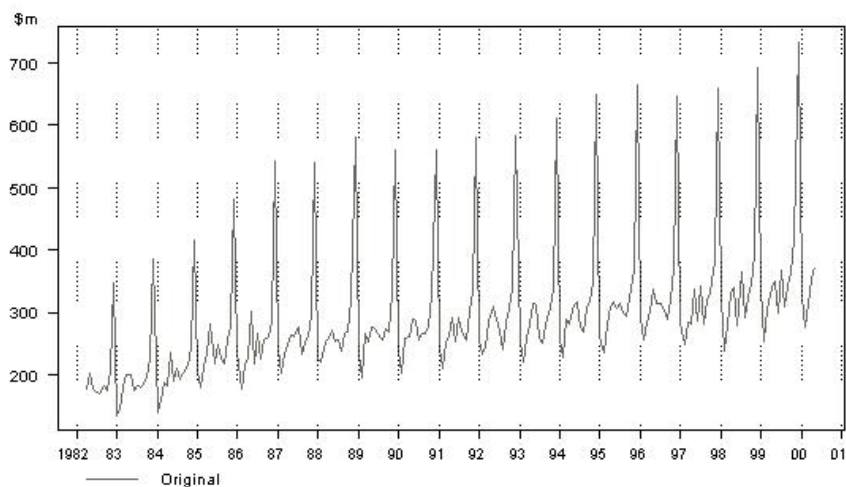


What is an Irregular?

The irregular component (sometimes also known as the residual) is what remains after the seasonal and trend components of a time series have been estimated and removed. It results from short term fluctuations in the series which are neither systematic nor predictable. In a highly irregular series, these fluctuations can dominate movements, which will mask the trend and seasonality. The following graph is of a highly irregular time series:

Figure Monthly Value of Building Approvals, Australian Capital Territory (ACT)
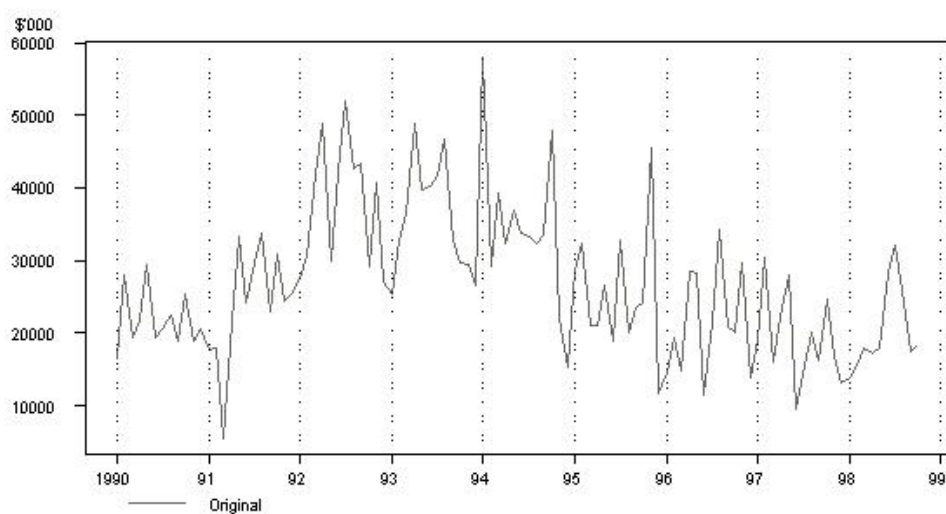
What is The Trend?

The ABS trend is defined as the 'long term' movement in a time series without calendar related and irregular effects, and is a reflection of the underlying level. It is the result of influences such as population growth, price inflation and general economic changes. The following graph depicts a series in which there is an obvious upward trend over time:

Figure: Quarterly Gross Domestic Product
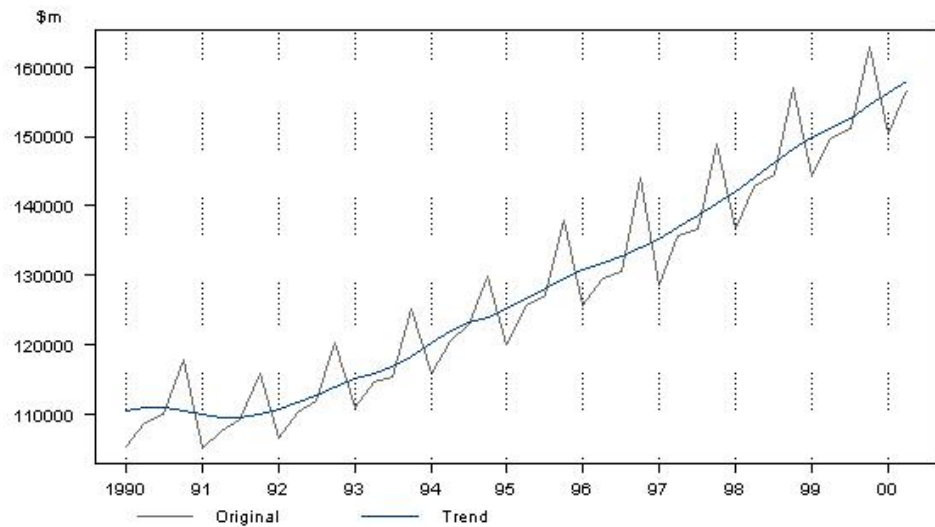


## 10.5 Components for Time Series Analysis

The various reasons or the forces which affect the values of an observation in a time series are the components of a time series. The four categories of the components of time series are:

- Trend
- Seasonal Variations
- Cyclic Variations

Random or Irregular movements

Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations.
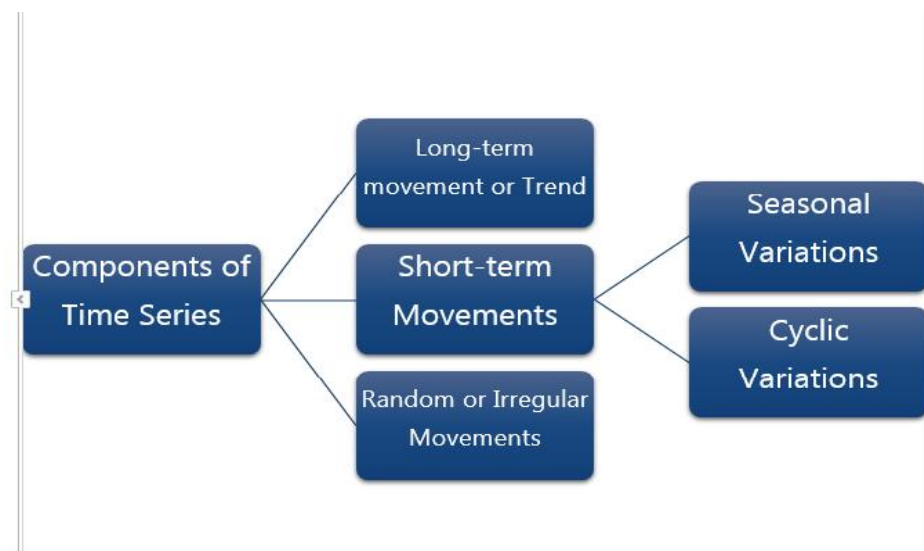
Trend: The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.

It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable. The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.

Linear and Non-Linear Trend: If we plot the time series values on a graph in accordance with time t. The pattern of the data clustering shows the type of trend. If the set of data cluster more or less round a straight line, then the trend is linear otherwise it is non-linear (Curvilinear)

Seasonal Variations: These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

## 10.6  Cyclic Variations

The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.

It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them

Random or Irregular Movements: There is another factor which causes the variation in the variable under study. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

Cyclic and seasonal time series: A **seasonal** pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a fixed and known period. Hence, seasonal time series are sometimes called **periodic** time series.

A **cyclic** pattern exists when data exhibit rises and falls that are not of fixed period. The duration of these fluctuations is usually of at least 2 years. Think of business cycles which usually last several years, but where the length of the current cycle is unknown beforehand.

Many people confuse cyclic behavior with seasonal behavior, but they are really quite different. If the fluctuations are not of fixed period then they are cyclic; if the period is unchanging and associated with some aspect of the calendar, then the pattern is seasonal. In general, the average length of cycles is longer than the length of a seasonal pattern, and the magnitude of cycles tends to be more variable than the magnitude of seasonal patterns.

*Probability and Statistics*

Trend: A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend as "changing direction," when it might go from an increasing trend to a decreasing trend. There is a trend in the ant diabetic drug sales data shown in Figure

Seasonal

A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency. The monthly sales of ant diabetic drugs above shows seasonality which is induced partly by the change in the cost of the drugs at the end of the calendar year.

Cyclic: A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. These fluctuations are usually due to economic conditions, and are often related to the "business cycle." The duration of these fluctuations is usually at least 2 years.

Many people confuse cyclic behavior with seasonal behavior, but they are really quite different. If the fluctuations are not of a fixed frequency then they are cyclic; if the frequency is unchanging and associated with some aspect of the calendar, then the pattern is seasonal. In general, the average length of cycles is longer than the length of a seasonal pattern, and the magnitudes of cycles tend to be more variable than the magnitudes of seasonal patterns.

Many time series include trend, cycles and seasonality. When choosing a forecasting method, we will first need to identify the time series patterns in the data, and then choose a method that is able to capture the patterns properly.

The examples in Figure show different combinations of the above components
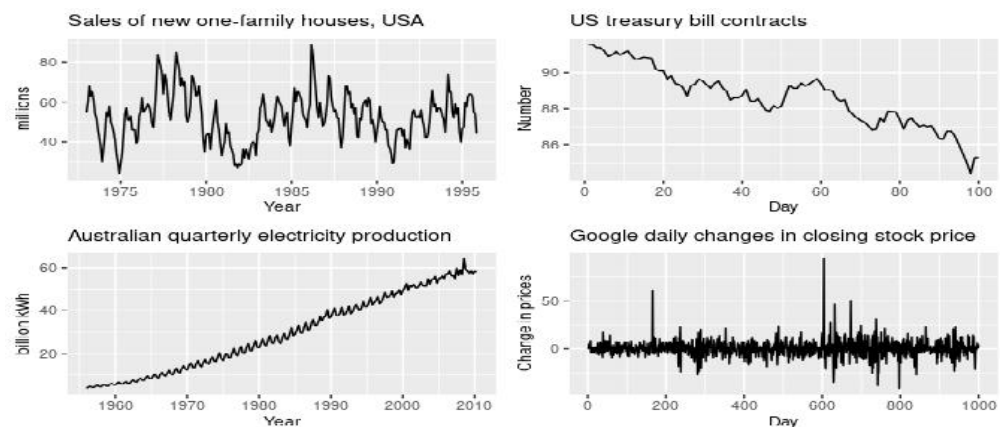


Figure: Four examples of time series showing different patterns.

The monthly housing sales (top left) show strong seasonality within each year, as well as some strong cyclic behavior with a period of about 6–10 years. There is no apparent trend in the data over this period.

The US Treasury bill contracts (top right) show results from the Chicago market for 100 consecutive trading days in 1981. Here there is no seasonality, but an obvious downward trend. Possibly, if we had a much longer series, we would see that this downward trend is actually part of a long cycle, but when viewed over only 100 days it appears to be a trend.

The Australian quarterly electricity production (bottom left) shows a strong increasing trend, with strong seasonality. There is no evidence of any cyclic behavior here.

The daily change in the Google closing stock price (bottom right) has no trend, seasonality or cyclic behavior. There are random fluctuations which do not appear to be very predictable, and no strong patterns that would help with developing a forecasting model.

The advantages of time series analysis are as follows:

Reliability: Time series analysis uses historical data to represent conditions along with a progressive linear chart. The information or data used is collected over a period of time say, weekly, monthly, quarterly or annually. This makes the data and forecasts reliable.

Seasonal Patterns: As the data related to a series of periods, it helps us to understand and predict the seasonal pattern. For example, the time series may reveal that the demand for ethnic clothes not only increases during Diwali but also during the wedding season.

Estimation of trends: The time series analysis helps in the identification of trends. The data tendencies are useful to managers as they show an increase or decrease in sales, production, share prices, etc.

Growth: Time series analysis helps in the measurement of financial growth. It also helps in measuring the internal growth of an organization that leads to economic growth.

Measurements of Trends

Following are the methods by which we can measure the trend.

(i) Freehand or Graphic Method.

(ii) Method of Semi-Averages.

(iii) Method of Moving Averages.

(iv) Method of Least Squares.

Freehand or Graphic Method.

It is the simplest and most flexible method for estimating a trend. We will see the working procedure of this method.

Procedure:

(a) Plot the time series data on a graph.

(b) Draw a freehand smooth curve joining the plotted points.

(C)Examine the direction of the trend based on the plotted points.

(d) Draw a straight line which will pass through the maximum number of plotted points.

## Fit a trend line by the method of freehand method for the given data.

| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|------|------|------|------|------|------|------|------|------|
| Sales | 30 | 46 | 25 | 59 | 40 | 60 | 38 | 65 |
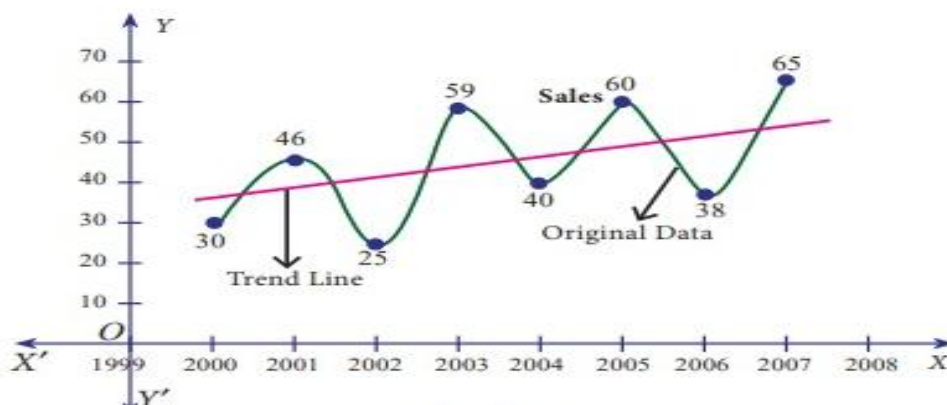
*Probability and Statistics*

The trend drawn by the freehand method can be extended to predict the future values of the given data. However, this method is subjective in nature, predictions obtained by this method depends on the personal bias and judgment of the investigator handling the data.

Method of Semi-Averages

In this method, the semi-averages are calculated to find out the trend values. Now, we will see the working procedure of this method.

Procedure:

(i) The data is divided into two equal parts. In case of odd number of data, two equal parts can be made simply by omitting the middle year.

(ii) The average of each part is calculated, thus we get two points.

iii) Each point is plotted at the mid-point (year) of each half.

(iv) Join the two points by a straight line.

(v) The straight line can be extended on either side.

(vi) This line is the trend line by the methods of semi-averages.

## Fit a trend line by the method of semi-averages for the given data

| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|------|------|------|------|------|------|------|------|
| Production | 105 | 115 | 120 | 100 | 110 | 125 | 135 |

**Since the number of years is odd(seven), we will leave the middle year's production value and obtain the averages of first three years and last three years.**

| Year | Production | Average |
|------|------------|---------|
| 2000 | 105 | |
| 2001 | 115 | $\dfrac{105+115+120}{3}=113.33$ |
| 2002 | 120 | |
| 2003 | 100 (left out) | |
| 2004 | 110 | |
| 2005 | 125 | $\dfrac{110+125+135}{3}=123.33$ |
| 2006 | 135 | |

Table
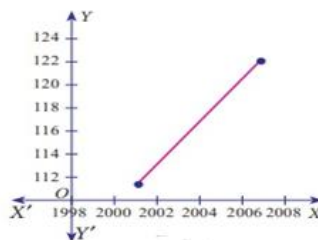


Method of Moving Averages

Moving Averages Method gives a trend with a fair degree of accuracy. In this method, we take arithmetic mean of the values for a certain time span. The time span can be three-years, four -years, five- years and so on depending on the data set and our interest. We will see the working procedure of this method.

Procedure:

(i) Decide the period of moving averages (three- years, four -years).

(ii) In case of odd years, averages can be obtained by calculating,

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}, \frac{d+e+f}{3}, \ldots\ldots$$

(iii) If the moving average is an odd number, there is no problem of centering it, the average value will be centered besides the second year for every three years.

In case of even years, averages can be obtained by calculating,

$$\frac{a+b+c+d}{4}, \frac{b+c+d+e}{4}, \frac{c+d+e+f}{4}, \frac{d+e+f+g}{4}, \ldots$$

If the moving average is an even number, the average of first four values will be placed between 2 nd and 3rd year, similarly the average of the second four values will be placed between 3rd and 4th year. These two averages will be again averaged and placed in the 3rd year. This continues for rest of the values in the problem. This process is called as centering of the averages

Method of Least Squares

The line of best fit is a line from which the sum of the deviations of various points is zero. This is the best method for obtaining the trend values. It gives a convenient basis for calculating the line of best fit for the time series. It is a mathematical method for measuring trend. Further the sum of the squares of these deviations would be least when compared with other fitting methods

## Summary

Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations.

The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency

It is not always necessary that the increase or decrease is in the same direction throughout the given period of time

Seasonal variations are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year.

Cyclic variations are time series which operate themselves over a span of more than one year are the cyclic variations.

The most important use of studying time series is that it helps us to predict the future behaviour of the variable based on past experience

It is helpful for business planning as it helps in comparing the actual current performance with the expected one

## Keywords

Methods by which we can measure the trend.

(i) Freehand or Graphic Method.

(ii) Method of Semi-Averages.

(iii) Method of Moving Averages.

(iv) Method of Least Squares

Forecasting is a common statistical task in business, where it helps to inform decisions about the scheduling of production, transportation and personnel, and provides a guide to long-term strategic planning.

A time series is a sequence of data points that occur in successive order over some period of time. This can be contrasted with cross-sectional data, which captures a point-in-time

Forecasting methods using time series are used in both fundamental and technical analysis.

Although cross-sectional data is seen as the opposite of time series, the two are often used together in practice.

## Self Assessment

1. Which of the following is an example of time series problem?

   1. Estimating number of hotel rooms booking in next 6 months.
   2. Estimating the total sales in next 3 years of an insurance company.
   3. Estimating the number of calls for the next one week.

   A. 1and 2
   B. 2 and 3
   C. 1 and 3
   D. 1, 2 and 3

2. _____are used in strategic planning

   A. Long-term forecasts
   B. Short term forecasts
   C. Medium term forecasts
   D. None of the above

3. _____are needed for the scheduling of personnel, production and transportation

   A. Long-term forecasts
   B. Short term forecasts
   C. Medium term forecasts
   D. None of the above

4. _____are needed to determine future resource requirements, in order to purchase raw materials, hire personnel, or buy machinery and equipment.

   A. Long-term forecasts
   B. Short term forecasts
   C. Medium term forecasts
   D. None of the above

5. A _____ is a sequence of data points that occur in successive order over some period of time.

A.  Time series

B.  Forecasting

C.  Planning

D.  None of these


6. _____methods using time series are used in both fundamental and technical analysis.

A.  Time series

B.  Forecasting

C.  Planning

D.  None of these


7. _____data is seen as the opposite of time series.

A.  Cross-sectional

B.  Planning data

C.  Short term data

D.  None of these


8. _____analysis looks at data collected at a single point in time, rather than over a period of time.

A.  Cross-sectional

B.  Planning data

C.  Short term data

D.  None of these


9. A _____ when there is a long-term increase or decrease in the data

A.  Trend

B.  Cyclic

C.  Seasonal

D.  All of these


10. These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year

A.  Trend

B.  Cyclic

C.  Seasonal variations

D.  All of these


11.Oscillatory movement has a period of oscillation of more than a year.

A.  Trend

B.  Cyclic

C.  Seasonal variations

D.  All of these

12. Seasonal and Cyclic Variations are _____

A. Short-term fluctuations.
B. Long term fluctuations.
C. Both of these
D. None of these

13. A _____is a smooth, general, long-term, average tendency

A. Trend
B. Cyclic
C. Seasonal variations
D. All of these

14 .It is the simplest and most flexible method for estimating a trend.

A. Freehand method
B. Method of Semi averages
C. Method of least squares
D. None of these

15. The _____is a standard approach in analysis to approximate the solution of over determined systems by minimizing the sum.

A. Freehand method
B. Method of Semi averages
C. Method of least squares
D. None of these

## Answers for Self Assessment

| 1. | D | 2. | A | 3. | B | 4. | C | 5. | A |
|----|---|----|---|----|---|----|---|----|---|
| 6. | B | 7. | A | 8. | A | 9. | A | 10. | C |
| 11. | B | 12. | A | 13. | A | 14. | A | 15. | C |

## Review Question

1) What is time series analysis with example?
2) How do you analyze time series?
3) What are the 4 components of time series?
4) What are the types of time series analysis?
5) What is the purpose of time series analysis?
6) How time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time?
7) How many elements are there in time series?
8) How do you know if a time series is multiplicative or additive?

## Further Readings

An Introduction to Probability and Statistics Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

First Course in Probability, Book by Sheldon M. Ross

Schaums Theory and Problems of Statistics Book by Murray R. Spiegel

Introduction to Probability, Statistics, and Random .Book by Hossein Pishro-Nik

## Web Links

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability

**Lovely Professional University**

# Unit 11: Sampling Theory

| CONTENTS |
| --- |
| Objectives |
| Introduction |
| 11.1    Population and Sample |
| 11.2    Types of Sampling: Sampling Methods |
| 11.3    What is Non-Probability Sampling? |
| 11.4    Uses of Probability Sampling |
| 11.5    Uses of Non-Probability Sampling |
| 11.6    What is a sampling error? |
| 11.7    Categories of Sampling Errors |
| 11.8    Sampling with Replacement and Sampling without Replacement |
| 11.9    Definition of Sampling Theory |
| 11.10   Data Collection Methods |
| Summary |
| Keywords |
| Self Assessment |
| Answers for Self Assessment |
| Review Questions |
| Further Readings |

## Objectives

- Understand basics of sampling techniques.

- Learn concepts of probability sampling.

- Define basic terms ofNon Probability Sampling.

- Understand concept of Sampling with replacement.

- Solve basic questions related to Sampling without replacement

## Introduction

What are the sampling methods or Sampling Techniques?

In Statistics, the sampling method or sampling technique is the process of studying the population by gathering information and analyzing that data. It is the basis of the data where the sample space is enormous.

There are several different sampling techniques available, and they can be subdivided into two groups.

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country's population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug's behavior

## 11.1 Population and Sample

In statistics as well as in quantitative methodology, the set of data are collected and selected from a statistical population with the help of some defined procedures. There are two different types of data sets namely, **population and sample**. So basically when we calculate the mean deviation, variance and standard deviation, it is necessary for us to know if we are referring to the entire population or to only sample data. Suppose the size of the population is denoted by 'n' then the sample size of that population is denoted by n -1. Let us take a look of population data sets and sample data sets in detail.

Population

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a **parameter**.

For example, all people living in India indicates the population of India.

There are different types of population. They are:

Finite Population

Infinite Population

Existent Population

Hypothetical Population

Let us discuss all the types one by one.

**Finite Population**: The finite population is also known as a countable population in which the population can be counted. In other words, it is defined as the population of all the individuals or objects that are finite. For statistical analysis, the finite population is more advantageous than the infinite population. Examples of finite populations are employees of a company, potential consumer in a market.

**Infinite Population**: The infinite population is also known as an uncountable population in which the counting of units in the population is not possible. Example of an infinite population is the number of germs in the patient's body is uncountable.

**Existent Population**: The existing population is defined as the population of concrete individuals. In other words, the population whose unit is available in solid form is known as existent population. Examples are books, students etc.

**Hypothetical Population**: The population in which whose unit is not available in solid form is known as the hypothetical population. A population consists of sets of observations, objects etc that are all something in common. In some situations, the populations

**Sample**: It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic. Sampling is the process of selecting the sample from the population. For example, some people living in India is the sample of the population.

Some of the key differences between population and sample are clearly given below:

| Comparison | Population | Sample |
|---|---|---|
| Meaning | Collection of all the units or elements that possess common characteristics | A subgroup of the members of the population |

| Includes | Each and every element of a group | Only includes a handful of units of population |
|---|---|---|
| Characteristics | Parameter | Statistic |
| Data Collection | Complete enumeration or census | Sampling or sample survey |
| Focus on | Identification of the characteristics | Making inferences about the population |

## 11.2  Types of Sampling: Sampling Methods

Sampling in market research is of two types – probability sampling and non-probability sampling. Let's take a closer look at these two methods of sampling.

Probability sampling: Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.
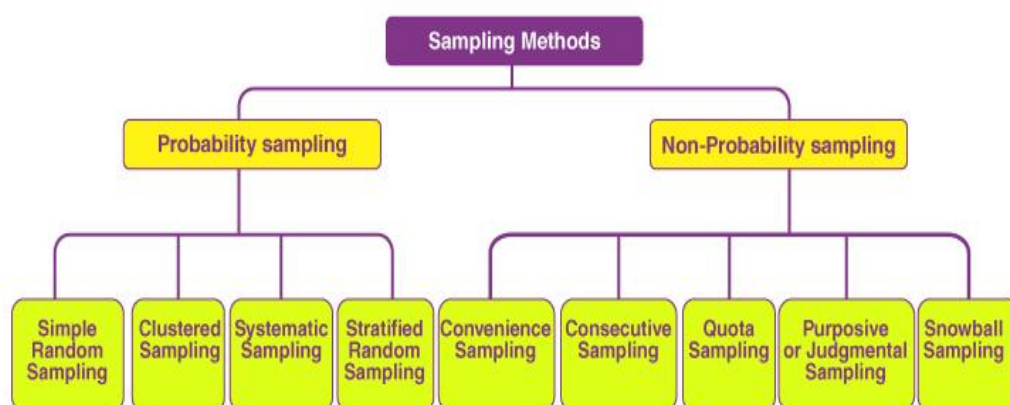
Non-probability sampling: In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

Types of Sampling Method

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are::

Probability Sampling

Non-probability Sampling

*Probability And Statistics*

## What is Probability Sampling?

The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of the population.

**Probability Sampling Types**: Probability Sampling methods are further classified into different types, such as simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Let us discuss the different types of probability sampling methods along with illustrative examples here in detail.

**Simple Random Sampling**: In simple random sampling technique, every item in the population has an equal and likely chance of being selected in the sample. Since the item selection entirely depends on the chance, this method is known as "**Method of chance Selection**". As the sample size is large, and the item is chosen randomly, it is known as "**Representative Sampling**".

Example:Suppose we want to select a simple random sample of 200 students from a school. Here, we can assign a number to every student in the school database from 1 to 500 and use a random number generator to select a sample of 200 numbers.

**Systematic Sampling**: In the systematic sampling method, the items are selected from the target population by selecting the random selection point and selecting the other methods after a fixed sample interval. It is calculated by dividing the total population size by the desired population size.

Example:Suppose the names of 300 students of a school are sorted in the reverse alphabetical order. To select a sample in a systematic sampling method, we have to choose some 15 students by randomly selecting a starting number, say 5. From number 5 onwards, will select every 15th person from the sorted list. Finally, we can end up with a sample of some students.

**Stratified Sampling**: In a stratified sampling method, the total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group, the statisticians randomly select the sample.

For example, there are three bags (A, B and C), each with different balls. Bag A has 50 balls, bag B has 100 balls, and bag C has 200 balls. We have to choose a sample of balls from each bag proportionally. Suppose 5 balls from bag A, 10 balls from bag B and 20 balls from bag C.

**Clustered Sampling**: In the clustered sampling method, the cluster or group of people are formed from the population set. The group has similar significatory characteristics. Also, they have an equal chance of being a part of the sample. This method uses simple random sampling for the cluster of population.

Example:An educational institution has ten branches across the country with almost the number of students. If we want to collect some data regarding facilities and other things, we can't travel to every unit to collect the required data. Hence, we can use random sampling to select three or four branches as clusters.

All these four methods can be understood in a better manner with the help of the figure given below.
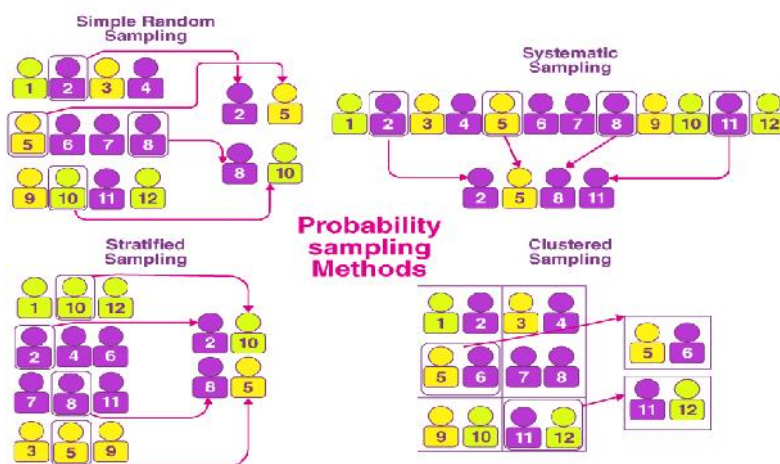
## 11.3  What is Non-Probability Sampling?

The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.

**Non-Probability Sampling Types:** Non-probability Sampling methods are further classified into different types, such as convenience sampling, consecutive sampling, quota sampling, judgmental sampling, snowball sampling. Here, let us discuss all these types of non-probability sampling in detail.

**Convenience Sampling:** In a convenience sampling method, the samples are selected from the population directly because they are conveniently available for the researcher. The samples are easy to select, and the researcher did not choose the sample that outlines the entire population.

**Example:**In researching customer support services in a particular region, we ask your few customers to complete a survey on the products after the purchase. This is a convenient way to collect data. Still, as we only surveyed customers taking the same product. At the same time, the sample is not representative of all the customers in that area.

**Consecutive Sampling:** Consecutive sampling is similar to convenience sampling with a slight variation. The researcher picks a single person or a group of people for sampling. Then the researcher researches for a period of time to analyze the result and move to another group if needed.

**Quota Sampling:** In the quota sampling method, the researcher forms a sample that involves the individuals to represent the population based on specific traits or qualities. The researcher chooses the sample subsets that bring the useful collection of data that generalizes the entire population.

**Purposive or Judgmental Sampling**: In purposive sampling, the samples are selected only based on the researcher's knowledge. As their knowledge is instrumental in creating the samples, there are the chances of obtaining highly accurate answers with a minimum marginal error. It is also known as judgmental sampling or authoritative sampling.

**Snowball Sampling:** Snowball sampling is also known as a chain-referral sampling technique. In this method, the samples have traits that are difficult to find. So, each identified member of a population is asked to find the other sampling units. Those sampling units also belong to the same targeted population.

Probability sampling vs. Non-probability Sampling Methods

The below table shows a few differences between probability sampling methods and non-probability sampling methods.

*Probability And Statistics*

Task: What is difference betweenProbability vs. Non probability sampling?

| Probability Sampling Methods | Non-probability Sampling Methods |
|---|---|
| Probability Sampling is a sampling technique in which samples taken from a larger population are chosen based on probability theory. | Non-probability sampling method is a technique in which the researcher chooses samples based on subjective judgment, preferably random selection. |
| These are also known as Random sampling methods. | These are also called non-random sampling methods. |
| These are used for research which is conclusive. | These are used for research which is exploratory. |
| These involve a long time to get the data. | These are easy ways to collect the data quickly. |
| There is an underlying hypothesis in probability sampling before the study starts. Also, the objective of this method is to validate the defined hypothesis. | The hypothesis is derived later by conducting the research study in the case of non-probability sampling. |

## 11.4  Uses of Probability Sampling

There are multiple uses of probability sampling:

Reduce Sample Bias: Using the probability sampling method, the bias in the sample derived from a population is negligible to non-existent. The selection of the sample mainly depicts the understanding and the inference of the researcher. Probability sampling leads to higher quality data collection as the sample appropriately represents the population.

Diverse Population: When the population is vast and diverse, it is essential to have adequate representation so that the data is not skewed towards one demographic. For example, if Square would like to understand the people that could make their point-of-sale devices, a survey conducted from a sample of people across the US from different industries and socio-economic backgrounds helps.

Create an Accurate Sample: Probability sampling helps the researchers plan and create an accurate sample. This helps to obtain well-defined data.

## 11.5  Uses of Non-Probability Sampling

Non-probability sampling is used for the following:

Create a hypothesis: Researchers use the non-probability sampling method to create an assumption when limited to no prior information is available. This method helps with the immediate return of data and builds a base for further research.

Exploratory research: Researchers use this sampling technique widely when conducting qualitative research, pilot studies, or exploratory research.

Budget and time constraints: The non-probability method when there are budget and time constraints, and some preliminary data must be collected. Since the survey design is not rigid, it is easier to pick respondents at random and have they take the survey or questionnaire.

How do you decide on the type of sampling to use?

For any research, it is essential to choose a sampling method accurately to meet the goals of your study. The effectiveness of your sampling relies on various factors. Here are some steps expert researchers follow to decide the best sampling method.

Jot down the research goals. Generally, it must be a combination of cost, precision, or accuracy.

Identify the effective sampling techniques that might potentially achieve the research goals.

Test each of these methods and examine whether they help in achieving your goal.

Select the method that works best for the research.

## 11.6 <u>What is a sampling error?</u>

A sampling error occurs when the sample used in the study is not representative of the whole population. Sampling errors often occur, and thus, researchers always calculate a margin of error during final results as a statistical practice. The margin of error is the amount of error allowed for a miscalculation to represent the difference between the sample and the actual population

What are the most common sampling errors in market research?

Here are the top four market research errors while sampling:

Population specification error: A population specification error occurs when researchers don't know precisely who to survey. For example, imagine a research study about kid's apparel. Who is the right person to survey? It can be both parents, only the mother, and the child. The parents make purchase decisions, but the kids may influence their choice.

Sample frame error: Sampling frame errors arise when researchers target the sub-population wrongly while selecting the sample. For example, picking a sampling frame from the telephone white pages book may have erroneous inclusions because people shift their cities. Erroneous exclusions occur when people prefer to un-list their numbers. Wealthy households may have more than one connection, thus leading to multiple inclusions.

**Selection error**: A selection error occurs when respondents self-select themselves to participate in the study. Only the interested ones respond. You can control selection errors by going the extra step to request responses from the entire sample. Pre-survey planning, follow-ups, and a neat and clean survey design will boost respondents' participation rate. Also, try methods like CATI surveys and in-person interviews to maximize responses.

**Sampling errors**: Sampling errors occur due to a disparity in the representativeness of the respondents. It majorly happens when the researcher does not plan his sample carefully. These sampling errors can be controlled and eliminated by creating a careful sample design, having a large enough sample to reflect the entire population, or using an online sample or survey audiences to collect responses.

**Controlling your sampling error**: Statistical theories help researchers measure the probability of sampling errors in sample size and population. The size of the sample considered from the population primarily determines the size of the sampling error. Larger sample sizes tend to encounter a lower rate of errors. Researchers use a metric known as the margin of error to understand and evaluate the margin of error. Usually, a confidence level of 95% is considered to be the desired confidence level.

**Pro Tip**: If you need help calculating your own margin of error, you can use our Margin of Error Calculator.

**Probability And Statistics**

What are the steps to reduce sampling errors?

Sampling errors are easy to identify. Here are a few simple steps to reduce sampling error:

Increase sample size: A larger sample size results in a more accurate result because the study gets closer to the actual population size.

Divide the population into groups: Test groups according to their size in the population instead of a random sample. For example, if people of a specific demographic make up 20% of the population, make sure that your study is made up of this variable.

Know your population: Study your population and understand its demographic mix. Know what demographics use your product and service and ensure you only target the sample that matters.

Sampling Errors Explained

Sampling errors are deviations in the sampled values from the values of the true population emanating from the fact that a sample is not an actual representative of a population of data.

Since there is a fault in the data collection, the results obtained from sampling become invalid. Furthermore, when a sample is selected randomly, or the selection is based on bias, it fails to denote the whole population, and sampling errors will certainly occur.

They can be prevented if the analysts select subsets or samples of data to represent the whole population effectively. Sampling errors are affected by factors such as the size and design of the sample, population variability, and sampling fraction.

Increasing the size of samples can eliminate sampling errors. However, to reduce them by half, the sample size needs to be increased by four times. If the selected samples are small and do not adequately represent the whole data, the analysts can select a greater number of samples for satisfactory representation.

The population variability causes variations in the estimates derived from different samples, leading to larger errors. The effect of population variability can be reduced by increasing the size of the samples so that these can more effectively represent the population.

Moreover, sampling errors must be considered when publishing survey results so that the accuracy of the estimates and the related interpretations can be established.
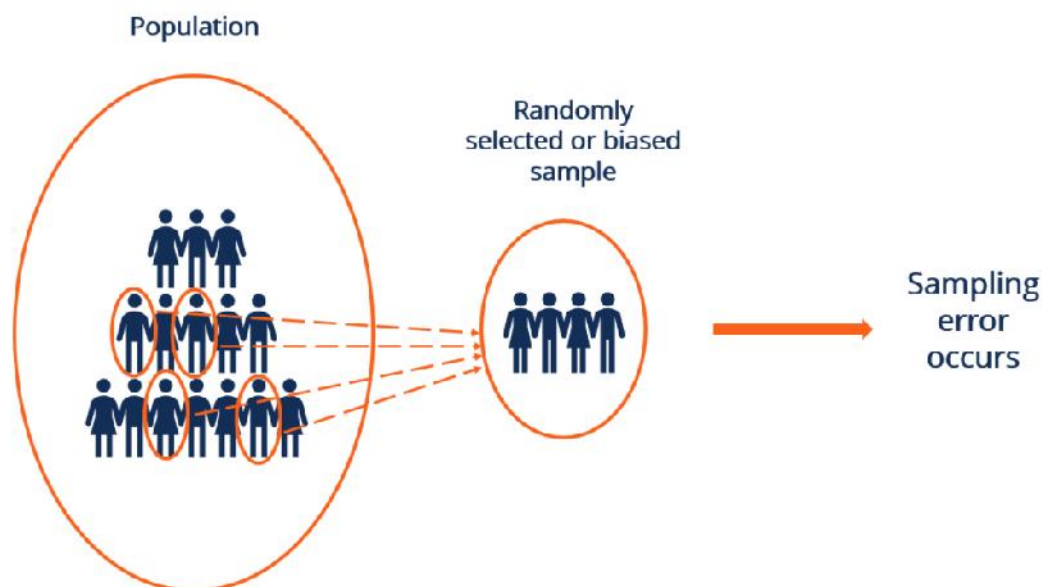


Example of Sampling Errors

Suppose the producers of Company XYZ want to determine the viewership of a local program that airs twice a week. The producers will need to determine the samples that can represent various types of viewers. They may need to consider factors like age, level of education, and gender.

**Lovely Professional University**

For example, people between the ages of 14 and 18 usually have fewer commitments, and most of them can spare time to watch the program twice weekly. On the contrary, people between the age of 18 and 35 usually have tighter schedules and will not have time to watch TV.

Hence, it is important to draw a sample proportionately. Otherwise, the results will not represent the real population.

Since the exact population parameter is not known, sampling errors for samples are generally unknown. However, analysts can use analytical methods to measure the amount of variation caused by sampling errors.

## 11.7  Categories of Sampling Errors

**Population Specification Error –** Happens when the analysts do not understand who to survey. For example, for a survey of breakfast cereals, the population can be the mother, children, or the entire family.

**Selection Error –** Occurs when the respondents' survey participation is self-selected, implying only those who are interested respond. Selection errors can be reduced by encouraging participation.

**Sample Frame Error –** Occurs when a sample is selected from the wrong population data.

**Non-Response Error –** Occurs when a useful response is not obtained from the surveys. It may happen due to the inability to contact potential respondents or their refusal to respond.

## 11.8  Sampling with Replacement and Sampling without Replacement

Sampling with replacement:

Consider a population of potato sacks, each of which has either 12, 13, 14, 15, 16, 17, or 18 potatoes, and all the values are equally likely. Suppose that, in this population, there is exactly one sack with each number. So the whole population has seven sacks. If I sample two with replacement, then I first pick one (say 14). I had a 1/7 probability of choosing that one. Then I replace it. Then I pick another. Every one of them still has 1/7 probability of being chosen. And there are exactly 49 different possibilities here (assuming we distinguish between the first and second.) They are: (12,12), (12,13), (12, 14), (12,15), (12,16), (12,17), (12,18), (13,12), (13,13), (13,14), etc.

Sampling without replacement:

Consider the same population of potato sacks, each of which has either 12, 13, 14, 15, 16, 17, or 18 potatoes, and all the values are equally likely. Suppose that, in this population, there is exactly one sack with each number. So the whole population has seven sacks. If I sample two without replacement, then I first pick one (say 14). I had a 1/7 probability of choosing that one. Then I pick another. At this point, there are only six possibilities: 12, 13, 15, 16, 17, and 18. So there are only 42 different possibilities here (again assuming that we distinguish between the first and the second.) They are: (12,13), (12,14), (12,15), (12,16), (12,17), (12,18), (13,12), (13,14), (13,15), etc.

What's the Difference?

When we sample with replacement, the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second. Mathematically, this means that the covariance between the two is zero.

In sampling without replacement, the two sample values aren't independent. Practically, this means that what we got on the for the first one affects what we can get for the second one. Mathematically, this means that the covariance between the two isn't zero.

Sampling theory is one of the techniques of statistical analysis. When there is research conducted on a group of people, then it is barely responsible to manage the data of each individual. And there comes the relevance of sampling theory. In this article, we have brought our readers a detailed discussion on what is sampling theory and everything about it that one should know.

*Probability And Statistics*

## 11.9  Definition of Sampling Theory

 The sampling theory definition of the statistic is the creation of a sample set. This is recognized as one of the major processes. It retains the accuracy in bringing out the correct statistical information. The population tree is a huge set and it turns out to be exhausting for the actual study and estimation process. Both money and time get exhausted in the process. The creation of the sample set saves time and effort and is a vital theory in the process of statistical data analysis.

Once you know the difference between these two terms, you are eligible to understand the information laid ahead.  In this part, you will learn how to identify the population and sample. The population can be referred to as the whole group of which you want to have a conclusion after making a statistical analysis. Samples are the groups within a population from which the data are to be collected.

The population can be categorized in terms of geographical locations, income, age groups, and many other categories. The population can be a narrow or a broad group as per the requirement. It would appear clearer with sampling theory examples.  For instance, if you are willing to conclude making statistical analysis about a topic on the adults, then the population can be a huge broad group. And on the other hand, if you are researching a particular company, then the population is a narrow one. The whole set of elements or entities is referred to as Population.

The sample frame is nothing but the set of sample elements that are under observation. The creation of samples is all that defines sampling theory. This is precisely the set of people that will be actively participating in the process of statistical analysis. The sampling model is when the set of the population has infinite elements.

### Process of Sampling

In this part of the article, we will discuss a few details regarding the process of sampling. So the steps are mentioned in the steps below:

The first step is a wise choice of the population set.

The second step is focusing on the sample set and the size of it.

Then, one needs to choose an identifiable property based on which the samples will be created out of the population set.

Then, the samples can be chosen using any of the types of sampling theory – Simple random, systematic, or stratified. Each of them is thoroughly discussed in the article ahead.

Checking the inaccuracy, if there is any.

Hence, the set is achieved in the result.

## Why Is Sampling Important for Researchers?

Everyone who has ever worked on a research project knows that resources are limited; time, money and people never come in an unlimited supply. For that reason, most research projects aim to gather data from a sample of people, rather than from the entire population (the census being one of the few exceptions).

 This is because sampling allows researchers to:

## Save Time

Contacting everyone in a population takes time. And, invariably, some people will not respond to the first effort at contacting them, meaning researchers have to invest more time for follow-up. Random sampling is much faster than surveying everyone in a population, and obtaining a non-random sample is almost always faster than random sampling. Thus, sampling saves researchers lots of time.

## Save Money

The number of people a researcher contacts is directly related to the cost of a study. Sampling saves money by allowing researchers to gather the same answers from a sample that they would receive from the population.

Non-random sampling is significantly cheaper than random sampling, because it lowers the cost associated with finding people and collecting data from them. Because all research is conducted on a budget, saving money is important.

## Collect Richer Data

Sometimes, the goal of research is to collect a little bit of data from a lot of people (e.g., an opinion poll). At other times, the goal is to collect a lot of information from just a few people (e.g., a user study or ethnographic interview). Either way, sampling allows researchers to ask participants more questions and to gather richer data than does contacting everyone in a population.

## The Importance of Knowing Where to Sample

Efficient sampling has a number of benefits for researchers. But just as important as knowing how to sample is knowing where to sample. Some research participants are better suited for the purposes of a project than others. Finding participants that are fit for the purpose of a project is crucial, because it allows researchers to gather high-quality data.

For example, consider an online research project. A team of researchers who decides to conduct a study online has several different sources of participants to choose from. Some sources provide a random sample, and many more provide a non-random sample. When selecting a non-random sample, researchers have several options to consider. Some studies are especially well-suited to an online panel that offers access to millions of different participants worldwide. Other studies, meanwhile, are better suited to a crowd sourced site that generally has fewer participants overall but more flexibility for fostering participant engagement.

To make these options more tangible, let's look at examples of when researchers might use different kinds of online samples.

Task: What is sampling theory?

## Different Use Cases of Online Sampling

### Academic Research

Academic researchers gather all kinds of samples online. Some projects require random samples based on probability sampling methods. Most other projects rely on non-random samples. In these non-random samples, researchers may sample a general audience from crowdsourcing websites or selectively target members of specific groups using online panels. The variety of research projects conducted within academia lends itself to many different types of online samples.

### Market Research

Market researchers often want to understand the thoughts, feelings and purchasing decisions of customers or potential customers. For that reason, most online market research is conducted in online panels that provide access to tens of millions of people and allow for complex demographic targeting. For some projects, crowdsourcing sites, such as Amazon Mechanical Turk, allow researchers to get more participant engagement than is typically available in online panels, because they allow researchers to select participants based on experience and to award bonuses.

### Public Polling

Public polling is most accurate when it is conducted on a random sample of the population. Hence, lots of public polling is conducted with nationally representative samples. There are, however, an increasing number of opinion polls conducted with non-random samples. When researchers poll people using non-random methods, it is common to adjust for known sources of bias after the data are gathered.

### User Testing

User testing requires people to engage with a website or product. For this reason, user testing is best done on platforms that allow researchers to get participants to engage deeply with their study. Crowdsourcing platforms are ideal for user testing studies, because researchers can often control participant compensation and reward people who are willing to make the effort in a longer study.

Online research is big business. There are hundreds of companies that provide researchers with access to online participants, but only a few facilitate research across different types of online panels or direct you to the right panel for your project.
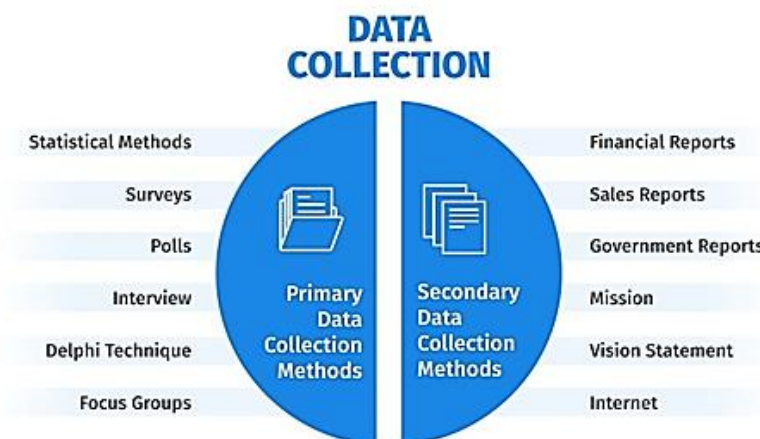
## 11.10 Data Collection Methods

Data is a collection of facts, figures, objects, symbols, and events gathered from different sources. **Organizations collect data to make better decisions**. Without data, it would be difficult for organizations to make appropriate decisions, and so data is collected at various points in time from different audiences.

For instance, before launching a new product, an organization needs to collect data on product demand, customer preferences, competitors, etc. In case data is not collected beforehand, the organization's newly launched product may lead to failure for many reasons, such as less demand and inability to meet customer needs.

Although data is a valuable asset for every organization, it does not serve any purpose until analyzed or processed to get the desired results.

You can categorize data collection methods into primary methods of data collection and secondary methods of data collection.



**Primary Data Collection Methods**

Primary data is collected from the first-hand experience and is not used in the past. The data gathered by primary data collection methods are specific to the research's motive and highly accurate.

Primary data collection methods can be divided into two categories: quantitative methods and qualitative methods.

**Quantitative Methods:**

Quantitative techniques for market research and demand forecasting usually make use of statistical tools. In these techniques, demand is forecast based on historical data. These methods of primary data collection are generally used to make long-term forecasts. Statistical methods are highly reliable as the element of subjectivity is minimum in these methods.

**Time Series Analysis**

The term time series refers to a sequential order of values of a variable, known as a trend, at equal time intervals. Using patterns, an organization can predict the demand for its products and services for the projected time.

**Smoothing Techniques**

In cases where the time series lacks significant trends, smoothing techniques can be used. They eliminate a random variation from the historical demand. It helps in identifying patterns and demand levels to estimate future demand. The most common methods used in smoothing demand forecasting techniques are the simple moving average method and the weighted moving average method.

**Barometric Method**

Also known as the leading indicators approach, researchers use this method to speculate future trends based on current developments. When the past events are considered to predict future events, they act as leading indicators.

**Qualitative Methods:**

Qualitative methods are especially useful in situations when historical data is not available. Or there is no need of numbers or mathematical calculations. Qualitative research is closely associated with words, sounds, feeling, emotions, colors, and other elements that are non-quantifiable. These techniques are based on experience, judgment, intuition, conjecture, emotion, etc.

Quantitative methods do not provide the motive behind participants' responses, often don't reach underrepresented populations, and span long periods to collect the data. Hence, it is best to combine quantitative methods with qualitative methods.

## Surveys

Surveys are used to collect data from the target audience and gather insights into their preferences, opinions, choices, and feedback related to their products and services. Most survey software often a wide range of question types to select.

You can also use a ready-made survey template to save on time and effort. Online surveys can be customized as per the business's brand by changing the theme, logo, etc. They can be distributed through several distribution channels such as email, website, offline app, QR code, social media, etc. Depending on the type and source of your audience, you can select the channel.

Once the data is collected, survey software can generate various reports and run analytics algorithms to discover hidden insights. A survey dashboard can give you the statistics related to response rate, completion rate, filters based on demographics, export and sharing options, etc. You can maximize the effort spent on online data collection by integrating survey builder with third-party apps.

**Polls**: Polls comprise of one single or multiple choice question. When it is required to have a quick pulse of the audience's sentiments, you can go for polls. Because they are short in length, it is easier to get responses from the people.

Similar to surveys, online polls, too, can be embedded into various platforms. Once the respondents answer the question, they can also be shown how they stand compared to others' responses.

**Interviews**: In this method, the interviewer asks questions either face-to-face or through telephone to the respondents. In face-to-face interviews, the interviewer asks a series of questions to the interviewee in person and notes down responses. In case it is not feasible to meet the person, the interviewer can go for a telephonic interview. This form of data collection is suitable when there are only a few respondents. It is too time-consuming and tedious to repeat the same process if there are many participants.

**Delphi Technique**: In this method, market experts are provided with the estimates and assumptions of forecasts made by other experts in the industry. Experts may reconsider and revise their estimates and assumptions based on the information provided by other experts. The consensus of all experts on demand forecasts constitutes the final demand forecast.

*Probability And Statistics*

---

**Focus Groups**: In a focus group, a small group of people, around 8-10 members, discuss the common areas of the problem. Each individual provides his insights on the issue concerned. A moderator regulates the discussion among the group members. At the end of the discussion, the group reaches a consensus.

**Questionnaire**: A questionnaire is a printed set of questions, either open-ended or closed-ended. The respondents are required to answer based on their knowledge and experience with the issue concerned. The questionnaire is a part of the survey, whereas the questionnaire's end-goal may or may not be a survey.

**Secondary Data Collection Methods**: Secondary data is the data that has been used in the past. The researcher can obtain data from the sources, both internal and external, to the organization.

**Internal sources of secondary data:**

Organization's health and safety records

Mission and vision statements

Financial Statements

Magazines

Sales Report

CRM Software

Executive summaries

**External sources of secondary data:**

Government reports

Press releases

Business journals

Libraries

Internet

The secondary data collection methods, too, can involve both quantitative and qualitative techniques. Secondary data is easily available and hence, less time-consuming and expensive as compared to the primary data. However, with the secondary data collection methods, the authenticity of the data gathered cannot be verified.

## Summary

- In Statistics, the **sampling method** or **sampling technique** is the process of studying the population by gathering information and analyzing that data. It is the basis of the data where the sample space is enormous.

- There are several different sampling techniques available, and they can be subdivided into two groups

- The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space.

- This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of the population

- In the systematic sampling method, the items are selected from the target population by selecting the random selection point and selecting the other methods after a fixed sample interval.

- It is calculated by dividing the total population size by the desired population size
- In a stratified sampling method, the total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group, the statisticians randomly select the sample.

## Keywords

- The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection.
- In a convenience sampling method, the samples are selected from the population directly because they are conveniently available for the researcher. The samples are easy to select, and the researcher did not choose the sample that outlines the entire population.
- Consecutive sampling is similar to convenience sampling with a slight variation. The researcher picks a single person or a group of people for sampling.
- In the quota sampling method, the researcher forms a sample that involves the individuals to represent the population based on specific traits or qualities.
- Snowball sampling is also known as a chain-referral sampling technique.
- In this method, the samples have traits that are difficult to find. So, each identified member of a population is asked to find the other sampling units. Those sampling units also belong to the same targeted population

## Self Assessment

1. _____uses randomization to select sample members.

A. Non-probability sampling
B. Probability sampling
C. Both of these
D. None of these

2. _____uses non-random techniques.

A. Non-probability sampling
B. Probability sampling
C. Both of these
D. None of these.

3. In this case each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected.

A. Systematic sampling
B. Stratified sampling
C. Simple Random sampling
D. None of these.

4. Individuals are selected at regular intervals from the sampling frame. The intervals are chosen to ensure an adequate sample size. If you need a sample size $n$ from a population of size $x$, you should select every $x/n^{th}$ individual for the sample

   A. Non Systematic sampling
   B. Stratified sampling
   C. Simple Random sampling
   D. Systematic Sampling

5. In this method, the population is first divided into subgroups who all share a similar characteristic.

   A. Non Systematic sampling
   B. Quota Sampling
   C. Simple Random sampling
   D. Stratified Sampling

6. _____is perhaps the easiest method of sampling, because participants are selected based on availability and willingness to take part.

   A. Convenience sampling
   B. Quota Sampling
   C. Simple Random sampling
   D. Stratified Sampling

7. This method is commonly used in social sciences when investigating hard-to-reach groups.

   A. Snowball Sampling
   B. Quota Sampling
   C. Simple Random sampling
   D. Stratified Sampling

8. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size

   A. Snowball Sampling
   B. Quota Sampling
   C. Simple Random sampling
   D. Stratified Sampling

9. _____is the number of completed responses your survey receives.

   A. Sample Size
   B. Population
   C. Random population

D.  All of these

10. _____means once we draw an item, then we do not replace it back to the sample space

before drawing a second item.

A.  Probability without replacement
B.  Probability with replacement
C.  None of these
D.  Both of these

11. When we _____, the two sample values are independent.

A.  Probability without replacement
B.  Probability with replacement
C.  None of these
D.  Both of these

12. In_____, the two sample values aren't independent.

A.  Probability without replacement
B.  Probability with replacement
C.  None of these
D.  Both of these

13. A _____is a probability distribution of a statistic obtained from a larger number of

samples drawn from a specific population.

A.  Sampling Distribution
B.  Normal Distribution
C.  Poison Distribution
D.  None of these

14. Of the following sampling methods, which is a probability method?

A.  Judgment
B.  Quota
C.  Simple random
D.  Convenience

15. Of the following sampling methods, which is a Non probability method?

A.  Systematic Sampling
B.  Quota sampling
C.  Simple random
D.  None of these

16. Sample is regarded as a subset of?

A. Data

B. Set

C. Distribution

D. Population

## Answers for Self Assessment

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. | B | 2. | A | 3. | C | 4. | D | 5. | D |
| 6. | A | 7. | A | 8. | A | 9. | A | 10. | A |
| 11. | B | 12. | A | 13. | A | 14. | C | 15. | B |
| 16. | D | | | | | | | | |

## Review Questions

Q1.justify this with Suitable example "Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. "

Q2. What is the most common type of sampling?

Q3What are the 4 types of non-probability sampling?

Q4 what is the difference between purposive and convenience sampling?

Q5 what is the difference between snowball sampling and convenience sampling?

Q6 what is sampling design example?

Q7 what is difference between probability and non-probability sampling?

Q8 what are the characteristics of probability sampling?

## Further Readings

An Introduction to Probability and Statistics Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

First Course in Probability,Book by Sheldon M. Ross

Schaums Theory and Problems of Statistics Book by Murray R. Spiegel

Introduction to Probability, Statistics, and Random ...Book by Hossein Pishro-Nik

## Web Links

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability

# Unit 12: Hypothesis Testing

| CONTENTS |
| --- |
| Objectives |
| Introduction |
| 12.1     Definition of Hypothesis |
| 12.2     Importance of Hypothesis |
| 12.3     Understanding Types of Hypothesis |
| 12.4     Formulating a Hypothesis |
| 12.5     Hypothesis Testing |
| 12.6     Hypothesis vs. prediction |
| Summary |
| Keywords |
| Self Assessment |
| Answers for Self Assessment |
| Review Questions |
| Further Readings |

## Objectives

- Understand basics of Hypothesis.
- Learn concepts of Null Hypothesis.
- Define basic charctertics ofHypothesis.
- Understand concept of Alternate Hypothesis.
- Learn basic of   level of significance.

## Introduction

Hypothesis is usually considered as an important mechanism in Research. Hypothesis is a tentative assumption made in order to test its logical or empirical consequences. If we go by the origin of the word, it is derived from the Greek word- 'hyposthenia' meaning 'to put under' or to 'to suppose'. Etymologically hypothesis is made up of two words, "hypo" and "thesis" which means less than or less certain than a thesis. It is a presumptive statement of a proposition or a reasonable guess, based upon the available evidence, which the researcher seeks to prove through his study. A hypothesis will give a plausible explanation that will be tested. A hypothesis may seem contrary to the real situation. It may prove to be correct or incorrect. Hypothesis need to be clear and precise and capable of being tested. It is to be limited in scope and consistent with known or established facts and should be amenable to testing within the stipulated time. It needs to explain what it claims to explain and should have empirical reference.

## 12.1  Definition of Hypothesis

"A hypothesis can be defined as a tentative explanation of the research problem, a possible outcome of the research, or an educated guess about the research outcome". Goode and Hatt have defined it as "a proposition which can be put to test to determine its validity". "Hypotheses are single tentative guesses, good hunches – assumed for use in devising theory or planning experiments intended to be given a direct experimental test when possible". According to Lundberg, "A hypothesis is a tentative generalization, the validity of which remains to be tested.

In its most elementary stage, the hypothesis may be any hunch, guess, imaginative idea, which becomes the basis for action or investigation".Hence, a hypothesis is a hunch, assumption, suspicion, assertion or an idea about a phenomenon, relationship or situation, the reality or truth of which you do not know. A researcher calls these assumptions/ hunches hypotheses and they become the basis of an enquiry. In most studies the hypothesis will be based upon your own or someone else's observation. Hypothesis brings clarity, specificity and focus to a research problem, but is not essential for a study. You can conduct a valid investigation without constructing formal hypothesis. The formulation of hypothesis provides a study with focus. It tells you what specific aspects of a research problem to investigate. A hypothesis tells you what data to collect and what not to collect, thereby providing focus to the study. As it provides a focus, the construction of a hypothesis enhances objectivity in a study. A hypothesis may enable you to add to the formulation of a theory. It enables you to specifically conclude what is true or what is false. Ludberg observes, quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable.

Notes: A hypothesis is a tentative generalization, the validity of which to be tested.

**Nature of Hypothesis:**

The hypothesis is a clear statement of what is intended to be investigated. It should be specified before research is conducted and openly stated in reporting the results.

**This allows to:**

- Identify the research objectives.
- Identify the key abstract concepts involved in the research.
- Identify its relationship to both the problem statement and the literature review.
- A problem cannot be scientifically solved unless it is reduced to hypothesis form.
- It is a powerful tool of advancement of knowledge, consistent with existing knowledgeand conducive to further enquiry.
- It can be tested – verifiable or falsifiable.
- Hypotheses are not moral or ethical questions.
- It is neither too specific nor to general.
- It is a prediction of consequences.
- It is considered valuable even if proven false

## 12.2  Importance of Hypothesis

Hypothesis though an important part of research may not be required in all types of research. The research which are based on fact finding (historical or descriptive research) do not need hypothesis. Hillway also says that "When fact-finding alone is the aim of the study, a hypothesis is not required". Whenever possible, a hypothesis is recommended for all major studies to explain observed facts, conditions or behaviour and to serve as a guide in the research process.

- Hypothesis facilitates the extension of knowledge in an area. They provide tentativeexplanations of facts and phenomena, and can be tested and validated. It sensitizes the investigator to certain aspects of the situations which are relevant from the standpointof the problem in hand.
- Hypothesis provide the researcher with rational statements, consisting of elements expressed in a logical order of relationships which seeks to describe or to explain conditions or events that have yet not been confirmed by facts. The hypothesis enables the researcher to relate logically known facts to intelligent guesses about unknown conditions. It is a guide to the thinking process and the process of discovery.

- Hypothesis provides direction to the research. It defines what is relevant and what is irrelevant. The hypothesis tells the researcher what he needs to do and find out in his study. Thus it prevents the review of irrelevant literature and provides a basis for selecting the sample and the research procedure to be used in the study.

- Hypothesis implies the statistical techniques needed in the analysis of data, and the relationship between the variables to be tested. It also helps to delimit his study in scope so that it does not become broad or unwieldy.

*Task:* Write an example for Null hypothesis?

- Hypothesis provides the basis for reporting the conclusion of the study. It serves as a framework for drawing conclusions. In other word, we can say that it provides the outline for setting conclusions in a meaningful way.

- So, Hypothesis has a very important place in research although it occupies a very small placein the body of a thesis.

### Sources of Hypothesis

A good hypothesis can only be derived from experience in research. Though hypothesis should precede the collection of data, but some degree of data collection, literature review or a pilot study will help in the development and gradual refinement of the hypothesis. A researcher should have quality of an alert mind to derive a hypothesis and quality of critical mind of rejecting faulty hypothesis.

The following sources can help the researcher in coming up with a good hypothesis:

- Review of literature.
- Discussion with the experts in the given field to understand the problem, its origin and
- Objectives in seeking a solution.
- Intuition of the researcher also sometimes helps in forming a good hypothesis.
- Previous empirical studies done on the given area.

## 12.3  Understanding Types of Hypothesis

Research Problems are too general by themselves to enable us to carryout meaningful analysis. They need to be specified in a more focused way. Hypotheses are specific statements that relate to the problem, the answers to which are likely to be yes or no, depending upon what is uncovered from the research.

Examples of Hypothesis can be:

- Suicide is related to general level of religiosity/secularization of society.
- Alienation and political participation are negatively related.

Such statements specify links between different phenomena, in order to explain different patterns of behavior that appear to occur. However, such patterns of association do not necessarily demonstrate that a causal relationship exists. We cannot for an instance say, 'socio-economic deprivation causes suicide.' If that was the case, then all those in Britain defined by various yardsticks as living in a state of relative poverty would inevitably commit suicide. This is very unlikely to happen.

## Variable:

So, to understand the types of hypothesis, we need to understand the concept of variables first. The variables are empirical properties that take two or more values or in other words a variable is any entity that can take on different values. In simple terms, anything that can vary or that is not constant can be considered a variable. For instance, age can be considered a variable because age can take different values for different people or for the same person at different times. Similarly, country can be considered a variable because a person's country can be assigned a value. A variable is a concept or abstract idea that can be described in measurable terms. In research, this term refers to the measurable characteristics, qualities, traits, or attributes of a particular individual, object or situation being studied. Variables differ in many respects, most notably in the role they are given in our research and in the type of measures that can be applied to them. The statement of problem usually provides only general direction for the research study. It does not include all the specific information. There is some basic terminology that is extremely important in how we communicate specific information about research problems and research in general. So, weight, height, income are all examples of variables. In Research, there is a need to make a distinction between various kinds of variables. There are many classifications given for variables. We will try to understand only the Dependent Variable and Independent Variable.

## Independent Variables:

The variables which are manipulated or controlled or changed. These are also known asmanipulated variables. Researchers often mistake independent variable and assume that it isindependent of any manipulation. It is called independent because variable is isolated from anyother factor. In research, we try to determine whether there is a cause and effect relationship.

In fact, when you are looking for some kind of relationship between variables you are trying tosee if the independent variable causes some kind of change in the other variables, or dependentvariables.

## Dependent Variables:

Dependent variables are the outcome variables and are the variables for which we calculate statistics. The variable which changes on account of independent variable is known as dependent variable. It is something that depends on other factors. For example, a test score could be a dependent variable because it could change depending on several factors such as how much you studied, how much sleep you got the night before you took the test, or even how hungry you were when you took it. Usually when you are looking for a relationship between two things you are trying to find out what makes the dependent variable change the way it does.

As we have discussed that a variable is an image, perception or concept that can be measured, hence capable of taking on different values. The variables that you wish to explain are regarded as dependent variables or criterion variables. The other variable expected to explain the change in the dependent variable is referred to as an independent variable or predictor variable. The dependent variable is the expected outcome of the independent variable and independent variable produce dependant variables.Variables can have three types of relationships among them.

- A positive relationship is one where an increase in one would lead to increase in theother.
- A negative relationship is one where an increase in one variable lead to decrease in theother.
- A zero relationship is one which shows no significant relationship between the twovariables.
- Once we have understood variables, we can discuss the various types of hypothesis.

Task: What is difference between Dependent and independent variables ?

**Lovely Professional University**

Research Hypothesis:The Research Hypothesis could be understood in terms of Simple Research hypothesis and Complex Research Hypothesis. A simple research hypothesis predicts the relationship between a single independent variable and a single dependent variable. A Complex hypothesis predicts the relationship between two or more independent variables and two or more dependent Variables. A research hypothesis must be stated in a testable form for its proper evaluation and it should indicate a relationship between variables in clear, concise and understandable Language. Research Hypothesis are classified as being directional or non-directional.

**Directional Hypotheses**: These are usually derived from theory. They may imply that the researcher is intellectually committed to a particular outcome. They specify the expected direction of the relationship between variables i.e. the researcher predicts not only the existence of a relationship but also its nature.

**Non-directional Hypotheses**: Used when there is little or no theory, or when findings of previous studies are contradictory. They may imply impartiality. Do not stipulate the direction of the relationship.

Associative and causal Hypotheses:Associative Hypotheses: Propose relationships between variables - when one variable changes, the other changes. Do not indicate cause and effect.

**Causal Hypothesis**: Propose a cause and effect interaction between two or more variables. The independent variable is manipulated to cause effect on the dependent variable. The dependent variable is measured to examine the effect created by the independent variable.

**Statistical Hypothesis**:To test whether the data support or refute the research hypothesis, it needs to be translated into a statistical hypothesis. It is given in statistical terms. In the context of inferential statistics, it is statement about one or more parameters that are measures of the population under study. Inferential statistics is used for drawing conclusions about population values. To use inferential statistics, we need to translate the research hypothesis into a testable form, which is called the null hypothesis. A testable hypothesis contains variables that are measurable or able to be manipulated. They need to predict a relationship that can be 'supported' or 'not supported' based on data collection and analysis.

**Null Hypothesis**: These are used when the researcher believes there is no relationship between two variables or when there is inadequate theoretical or empirical information to state a research hypothesis. The null hypothesis represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. Has serious outcome if incorrect decision is made. Designated by: Ho or Hn.

Null hypotheses can be:

- Simple or complex
- Associative or causal

**The Alternative Hypothesis**: The alternative hypothesis is a statement of what a hypothesis test is set up to establish. Designated by: H1 or Ha. It is opposite of Null Hypothesis. It is only reached if Ha is rejected. Frequently "alternative" is actual desired conclusion of the researcher.

We give special consideration to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if when the null is rejected. The final conclusion, once the test has been carried out, is always given in terms of the null hypothesis. We either 'reject Ho in favor of Ha' or 'do not reject Ho'; we never conclude 'reject Ha', or even 'accept Ha'. If we conclude 'do not reject Ho', this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against Ho in favor of Ha; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

For example:

Ha= the males visited cinema more than females.

Ho= the males and females do not differ in respect of the frequency of seeing cinema. So, Alternative hypothesis is usually the one which one wishes to prove and the Null hypothesis is the one which one wishes to disapprove.

## 12.4  Formulating a Hypothesis

There are no precise rules for formulating hypothesis and deducing consequences but there are some difficulties that arise in formulating the hypothesis. However, there are certain necessary conditions that are conducive to their formulation. They are:

-Richness of background knowledge: In the absence of knowledge concerning a subject matter, one can make no well-founded judgment of relevant hypothesis. Background knowledge is essential for perceiving relationships among the variables and to determine what findings other researchers have reported on the problem under study. New knowledge, new discoveries and new inventions should always form continuity with the already existing corpus of knowledge and therefore it becomes all the more essential to be well versed with the already existing knowledge.

Hypothesis can be formulated correctly by persons who have rich experience and academic background, but they can never be formulated by those who have poor background knowledge.

Logical and Scientific approach: Formulation of proper hypothesis depends on one's experience and logical insight. Hypothesis does not have a clear cut and definite theoretical background. Partly, it is a matter of lifting upon an idea on some problem and it is not always possible to have complete information of, and acquaintance with the scientific methods for formulating hypothesis. This lack of scientific knowledge presents difficulty in formulation of hypothesis. A researcher may begin a study by selecting one of the theories in his own area of interest and deduce a hypothesis from this theory through logic which is possible only when the researcher has a proper understanding of the scientific method and has a versatile intellect.

At times, conversations and consultations with colleagues and experts from different fields are also helpful in formulating important and useful hypothesis.

**Characteristics of a Good Hypothesis**:

Hulley says a good hypothesis must be based on a good research question. It should be simple, specific and stated in advance. So, a hypothesis could be called as a good hypothesis if it possesses the following characteristics:

- Hypothesis should be simple so that it is easily understood by everyone.
- Hypothesis should be clear, specific and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.
- Hypothesis should be capable of being tested.
- Hypothesis should state relationship between variables.
- Hypothesis should be consistent with most known facts. i.e. it must be consistent with a substantial body of established facts.
- The hypothesis must explain the facts that gave rise to the need for explanation. It must actually explain what it claims to explain. 10.

## 12.5  Hypothesis Testing

When the purpose of the research is to test a research hypothesis, it is termed as hypothesis testing research. It can be of experimental design or the non-experimental design. Research in which the independent variable is manipulated is termed 'experimental hypothesis-testing research' and a research in which an independent variable is not manipulated is called 'non experimental hypothesis testing research'.

Example: As we have discussed the Null hypothesis (Ho) and Alternative Hypothesis (Ha) earlier so while testing hypothesis we generally proceed on the basis of Null hypothesis (Ho), keeping the Alternative hypothesis in view. We do so because on the assumption that Null hypothesis is true, one can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the Alternative hypothesis. Hence the use of null Hypothesis is quite frequent. While testing the Hypothesis the following things to be kept in mind:

Level of significance: This is a very important concept in the context of hypothesis testing. It is always some percentage (usually 5%) which should be chosen with great care, thought and reason. In case we take the significance level at 5%, then this implies that Ho will be rejected when the sampling result (i.e observed evidence) has a less than 0.05 probability of occurring if Ho is true. In other words, the 5% level of significance means that researcher is willing to take as much as a 5% risk of rejecting the Null hypothesis when it happens to be true. Thus, the significance level is the maximum value of the probability of rejecting Ho when it is true and is usually determined in advance before testing the hypothesis.

The criteria for rejecting the null hypothesis may differ. Sometimes the Null hypothesis is rejected only when the quantity of the outcome is so large that the probability of its having occurred by mere chance is 1 times out of 100. We consider the probability of its having occurred by chance to be too little and we reject the chance theory of the Null hypothesis and take the occurrence to be due to genuine tendency. On the other occasions, we may reject the Null hypothesis even when the quantity of the reported outcome is likely to occur by chance 5 times out of 100. Statistically the former is known as the rejection of Null hypothesis at 0.1 level and the latter is known as the rejection at 0.5 level. It may be pointed out that if the researcher is able to reject the Null hypothesis, he cannot directly uphold the declarative hypothesis. If an outcome is not held to be due to chance, it does not mean that it is due to the very cause and effect relationship asserted in the particular declarative statement. It may be due to something else which the researcher may have failed to control.

Declaration rule or test of hypothesis: Given a Null hypothesis (Ho) and Alternative hypothesis (Ha), we make a rule which is known as decision rule according to which we accept Ho (i.e reject Ha)or reject Ho(i.e accept Ha). For instance, if Ho is, that a certain lot is good (there are very few defective items in it) against Ha that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis. We might test 10 times in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept Ho otherwise we will reject Ho (or accept Ha). This sort of basis is known as decision rule.

Two-tailed and one-tailed test: In the context of hypothesis testing, these two terms are quite important and must be clearly understood. A two-tailed rejects the Null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesized value of the mean of the population. Such a test is appropriate when the Null hypothesis is some specified value and the Alternative hypothesis is a value not equal to the specified value of Null hypothesis. In a two-tailed test, there are two rejection regions, one on each tail of the curve which can be illustrated as under:

If the significance level is 5% and the two-tailed test is to be applied, the probability of the rejection area will be 0.005 (equally divided on both tails of the curve is 0.0025) and that of the acceptance region will be 0.95.

But there are situations when only one-tailed test is considered appropriate. A onetailed test would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesized value. We should always remember that accepting Ho, on the basis of sample information does not constitute the proof that Ho, is true. We only mean that there is no statistical evidence to reject it.

Errors in Testing of Hypothesis: There are basically two types of errors we make in the context of testing of Hypothesis. These are called as Type-I error and the Type-II error. In type-I error, we may reject Null hypothesis when Null hypothesis is true. Type-II error is when we accept Null hypothesis when the Null Hypothesis is not true. In other words, Type-I error means rejection of hypothesis which should have been accepted and Type-II error means accepting the hypothesis which should have been rejected. Type-I error is denoted by alpha known as alpha error, also called the level of significance of test and Type-II error is denoted by beta known as beta error.

|  | Accept Null hypothesis | Reject Null hypothesis |
|---|---|---|
| Null hypothesis (true) | Correct decision | Type-I error (alpha error) |
| Null hypothesis (false) | Type-II error (beta error) | Correct decision |

The probability of Type-I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If Type-I error is fixed at 5%, it means that there are about 5 chance in100 that we will reject Null hypothesis when Null hypothesis is true. We can control Type-I error just by fixing at a lower level. For instance, if we fix it at 1%, we will say that the maximum probability of committing Type-I error would only be 0.01.

But with the fixed sample size, when we try to reduce Type-I error, the probability of committing Type-II error increases. Both types of errors cannot be reduced simultaneously. There is tradeoff between two types of errors which means that the probability of making one type error can only be reduced if we are willing to increase the probability of making the other type of error. One must set a very high level for Type-I error in one's testing technique of a given hypothesis. Hence, in the testing of hypothesis, one must make all possible efforts to strike an adequate balance between Type-I and Type-II errors.

### What is Statistical Significance?

In Statistics, "significance" means "not by chance" or "probably true". We can say that if a statistician declares that some result is "highly significant", then he indicates by stating that it might be very probably true. It does not mean that the result is highly significant, but it suggests that it is highly probable.

### Level of Significance Definition

The level of significance is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true. The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error. The level of significance is the measurement of the statistical significance. It defines whether the null hypothesis is assumed to be accepted or rejected. It is expected to identify if the result is statistically significant for the null hypothesis to be false or rejected.

### Level of Significance Symbol

The level of significance is denoted by the Greek symbol α (alpha). Therefore, the level of significance is defined as follows:

Significance Level = p (type I error) = α

The values or the observations are less likely when they are farther than the mean. The results are written as "significant at x%".

Example: The value significant at 5% refers to p-value is less than 0.05 or $p < 0.05$. Similarly, significant at the 1% means that the p-value is less than 0.01.

The level of significance is taken at 0.05 or 5%. When the p-value is low, it means that the recognised values are significantly different from the population value that was hypothesised in the beginning. The p-value is said to be more significant if it is as low as possible. Also, the result would be highly significant if the p-value is very less. But, most generally, p-values smaller than 0.05 are known as significant, since getting a p-value less than 0.05 is quite a less practice.

## How to Find the Level of Significance?

To measure the level of statistical significance of the result, the investigator first needs to calculate the p-value. It defines the probability of identifying an effect which provides that the null hypothesis is true. When the p-value is less than the level of significance ($\alpha$), the null hypothesis is rejected. If the p-value so observed is not less than the significance level $\alpha$, then theoretically null hypothesis is accepted. But practically, we often increase the size of the sample size and check if we reach the significance level. The general interpretation of the p-value based upon the level of significance of 10%:

If $p > 0.1$, then there will be no assumption for the null hypothesis

If $p > 0.05$ and $p \leq 0.1$, it means that there will be a low assumption for the null hypothesis.

If $p > 0.01$ and $p \leq 0.05$, then there must be a strong assumption about the null hypothesis.

If $p \leq 0.01$, then a very strong assumption about the null hypothesis is indicated.

Level of Significance Example

If we obtain a p-value equal to 0.03, then it indicates that there are just 3% chances of getting a difference larger than that in our research, given that the null hypothesis exists. Now, we need to determine if this result is statistically significant enough.

We know that if the chances are 5% or less than that, then the null hypothesis is true, and we will tend to reject the null hypothesis and accept the alternative hypothesis. Here, in this case, the chances are 0.03, i.e. 3% (less than 5%), which eventually means that we will eliminate our null hypothesis and will accept an alternative hypothesis.

## Null Hypothesis Examples

A null hypothesis, denoted by H0, proposes that two factors or groups are unrelated and that there is no difference between certain characteristics of a population or process. You must test the likelihood of the null hypothesis, in tandem with an alternative hypothesis, in order to disprove or discredit it. Some examples of a null hypothesis include:

There is no significant change in a person's health during the times when they drink green tea only or root beer only.

There is no significant change in an individual's work habits whether they get eight hours or nine hours of sleep.

There is no significant change in the growth of a plant if one uses distilled water only or vitamin-rich water only to water it.

## Alternative Hypothesis Examples

An alternative hypothesis, denoted by H1 or HA , is a claim that is contradictory to the null hypothesis. Researchers will pair the alternative hypothesis with the null hypothesis in order to prove that there is no relation. If the null hypothesis is disproven, then the alternative hypothesis will be accepted. If the null hypothesis is not rejected, then the alternative hypothesis will not be accepted. Some examples of alternative hypotheses are:

*Probability And Statistics*

A person's health improves during the times when they drink green tea only, as opposed to root beer only.Work habits improve during the times when one gets 8 hours of sleep only, as opposed to 9 hours of sleep only.

The growth of the plant improved during the times when it received vitamin-rich water only, as opposed to distilled water only.

### Statistical Hypothesis Examples

A statistical hypothesis is an examination of a portion of a population or statistical model. In this type of analysis, you use statistical information from an area. For example, if you wanted to conduct a study on the life expectancy of people from Savannah, you would want to examine every single resident of Savannah. This is not practical. Therefore, you would conduct your research using a statistical hypothesis or a sample of Savannah's population.

50% of Savannah's population lives beyond the age of 70.

80% of the U.S. population gets a divorce because of irreconcilable differences.

45% of the poor in the U.S. are illiterate.

## 12.6  Hypothesis vs. prediction

| Term | Hypothesis | Prediction |
|---|---|---|
| Definition | Explanation of a phenomenon | Event that will occur if phenomenon is true |
| What it does | Explains why something happens | Forecasts future event |
| How its written | Statement with variables | If, then statement |
| Example | Cholesterol of higher than 400 leads to heart disease | If someone has cholesterol of higher than 400, then they have heart disease |

### Prediction Examples

Need a few more examples of predictions? Explore these unique predictions to clarify the difference between hypothesis and prediction.

If the individual consumes greasy foods, then the person will have more skin oils and breakouts.

If the individual gets eight hours of sleep, then the individual will be more productive.

If the employer institutes a relaxation session in the workday, then the employees will be happier.

If the individual gets fewer than 8 hours of sleep, then the individual will be less productive.

If the employees are happier, then the workplace will be more productive

## 5-STEPS PROCEDURE FOR TESTING HYPOTHESIS

| STEPS | ACTIONS | DESCRIPTIONS |
|---|---|---|
| STEP 1 | State Null and Alternative hypothesis | Null Hypothesis : Ho = 0<br>Alternative Hypothesis : H1 = 0<br>Note :<br>1. **Two-tailed test** if alternative hypothesis does not state direction [ greater or less].<br>2. **One-tailed test** if alternative state direction. |
| STEP 2 | Select Level of Significance | 1. .01 level [1% level] - for consumer research<br>2. .05 level [5% level] – for quality assurance<br>3. .10 level [10% level] – for political pooling |
| STEP 3 | Identify the test Statistics | **z and t as test statistic , [*Note use t test when n is less than 30. If n is 30 and more use z test]*,** and others **F test [to test more than 2 means ]** and **Chi-square** for non-parameter statistic. |
| STEP 4 | Formulate Decision Rule | **Find the critical value** of z from Normal Distribution table , or value t from t distribution table , or Chi-Square, or F distribution table , where appropriate. |
| STEP 5 | Arrive at decision | **Only ONE DECISION** is possible in Hypothesis Testing Do **not reject Null Hypothesis**, or reject Null Hypothesis and Accept Alternative Hypothesis |

## Summary

- A hypothesis is a precise, testable statement of what the researcher(s) predict will be the outcome of the study.

- This usually involves proposing a possible relationship between two variables: the independent variable (what the researcher changes) and the dependent variable (what the research measures).

- In research, there is a convention that the hypothesis is written in two forms, the null hypothesis, and the alternative hypothesis (called the experimental hypothesis when the method of investigation is an experiment).

## Keywords

- The null hypothesis states that there is no relationship between the two variables being studied (one variable does not affect the other).

- It states results are due to chance and are not significant in terms of supporting the idea being investigated

- A one-tailed directional hypothesis predicts the nature of the effect of the independent variable on the dependent variable.

- E.g., adults will correctly recall more words than children.

- **Sample size and selection.** Your data needs to be representative of the target study population. You should use statistical methods to estimate the feasibility of your sample size.

- **Determine the criteria** for a successful pilot study based on the objectives of your study. How will your pilot study address these criteria?

## Self Assessment

1. A statement made about a population for testing purpose is called?

A. Statistic.

B. Hypothesis.

C. Level of Significance.

D. Statistic

2. The hypothesis that there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.

A. Null Hypothesis

B. Statistical Hypothesis

C. Simple Hypothesis

D. Composite Hypothesis

3. The hypothesis that there is some significant difference between specified populations, any observed difference being due to sampling or experimental error.

A. Null Hypothesis

B. Statistical Hypothesis

C. Alternate Hypothesis

D. Composite Hypothesis

4. If the null hypothesis is false then which of the following is accepted?

A. Null Hypothesis

B. Positive Hypothesis

C. Negative Hypothesis

D. Alternative Hypothesis.

5. The rejection probability of Null Hypothesis when it is true is called as?

A. Level of Confidence

B. Level of Significance

C. Level of Margin

D. Level of Rejection

6. If the Critical region is evenly distributed then the test is referred as?

A. Two tailed

B. One tailed

C. Three tailed

D. Zero tailed

7. A _____non-directional hypothesis predicts that the independent variable will have an effect on the dependent variable, but the direction of the effect is not specified.

A. Two tailed

B. One tailed

C. Three tailed

D. Zero tailed

8. A _____ hypothesis predicts that the independent variable will have an effect on the dependent variable, but the direction of the effect is specified.

A. Two tailed

B. One tailed

C. Three tailed

D. Zero tailed

9. You can determine the feasibility of your research design with a _____before you start.

A. Pilot study
B. Convenience Sampling
C. Random sampling
D. None of these

10 .What is First step of Procedure for Testing Hypothesis

A. State null and alternate hypothesis

B. State level of significance

C. Identify test statistics

D. Formulate decision rule

11. The _____is usually a hypothesis of equality between population parameters.

A. Null hypothesis

B. Alternate Hypothesis

C. Both of these

D. None of these

12. The _____is effectively the opposite of a null hypothesis.

A. Null hypothesis

B. Alternate Hypothesis

C. Both of these

D. None of these

13. $\mu_{after} = \mu_{before}$ (the mean sales is the same before and after spending more on advertising) is

A. Null hypothesis

B. Alternate Hypothesis

C. Both of these

D. None of these

14. μafter > μbefore (the mean sales increased after spending more on advertising) is

A. Null hypothesis

B. Alternate Hypothesis

C. Both of these

D. None of these

15. Null and alternative hypotheses are statements about:

A. Population parameters.

B. Sample parameters.

C. Sample statistics.

D. None of these

## Answers for Self Assessment

| 1. | B | 2. | A | 3. | C | 4. | D | 5. | B |
|----|---|----|---|----|---|----|---|----|---|
| 6. | A | 7. | A | 8. | B | 9. | A | 10. | A |
| 11. | A | 12. | B | 13. | A | 14. | B | 15. | A |

## Review Questions

1) How do you explain a hypothesis in any example?
2) Is a hypothesis a prediction?
3) What are the 3 required parts of a hypothesis?
4) Write example for alternate hypothesis?
5) How can you explain Null hypothesis?
6) What is the difference of null and alternative hypothesis?
7) What is meant by level of significance?
8) What is the difference between significance level and confidence level?

## Further Readings

An Introduction to Probability and StatisticsBook by A. K. Md. Ehsanes Salah and V. K. Rohatgi

First Course in Probability, ABook by Sheldon M. RossSchaums Theory and Problems of Statistics

Book by Murray R. SpiegelIntroduction to Probability, Statistics, and Random Book

by Hossein Pishro-Nik

**Web links**

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability

# Unit 13: Tests of Significance

## Objectives

- Understand basics of significance testing.
- Learn concepts of statistical testing.
- Define basic charctertics of T test and Z test.
- Understand concept of testing Hypothesis.

## Introduction

In Statistics, **tests of significance** are the method of reaching a conclusion to reject or support the claims based on sample data. The statistics are a special branch of Mathematics which deals with the collection and calculation over numerical data. This subject is well known for research based on statistical surveys. During a statistical process, a very common as well as an important term we come across is "significance".**Statistical significance** is very important in research not only in Mathematics but in several different fields such as medicine, psychology and biology. There are many methods through which the significance can be tested. These are known as **significance tests**. Let us learn about significance testing in detail.

## 13.1  Definition of Significance Testing

In statistics, it is important to know if the result of an experiment is significant enough or not. In order to measure the significance, there are some predefined tests which could be applied. These tests are called the tests of significance or simply the significance tests This statistical testing is subjected to some degree of error. For some experiments, the researcher is required to define the

probability of sampling error in advance. In any test which does not consider the entire population, the sampling error does exist. The testing of significance is very important in statistical research.

The significance level is the level at which it can be accepted if a given event is statistically significant. This is also termed as p-value. It is observed that the bigger samples are less prone to chance, thus the sample size plays a vital role in measuring the statistical significance. One should use only representative and random samples for significance testing. In short, the significance is the probability that a relationship exists. Significance tests tell us about the probability that if a relationship we found is due to random chance or not and to which level. This indicates about the error that would be made by us if the found relationship is assumed to exist.

### Tests of Significance in Statistics

Technically speaking, the statistical significance refers to the probability of a result of some statistical test or research occurring by chance. The main purpose of performing statistical research is basically to find the truth. In this process, the researcher has to make sure about the quality of sample, accuracy, and good measures which need a number of steps to be done. The researcher has to determine whether the findings of experiments have occurred due to a good study or just by fluke.

The significance is a number which represents probability indicating the result of some study has occurred purely by chance. The statistical significance may be weak or strong. It does not necessarily indicate practical significance. Sometimes, when a researcher does not carefully make use of language in the report of their experiment, the significance may be misinterpreted.

The psychologists and statisticians look for a 5% probability or less which means 5% results occur due to chance. This also indicates that there is a 95% chance of results occurring NOT by chance. Whenever it is found that the result of our experiment is statistically significant, it refers that we should be 95% sure the results are not due to chance.

## 13.2  Process of Significance Testing

In the process of testing for statistical significance, there are the following steps:

- Stating a Hypothesis for Research
- Stating a Null Hypothesis
- Selecting a Probability of Error Level
- Selecting and Computing a Statistical Significance Test
- Interpreting the results

### Types of Errors

There are basically two types of errors:

Type I

Type II

Type I Error

The type I error occurs when the researcher finds out that the relationship assumed through research hypothesis does exist; but in reality, there is evidence that it does not exist. In this type of error, the researcher is supposed to reject the research hypothesis and accept the null hypothesis, but its opposite happens. The probability that researchers commit Type I error is denoted by alpha (α).

Type II Error

The type II error is just opposite the type I error. It occurs when it is assumed that a relationship does not exist, but in reality it does. In this type of error, the researcher is supposed to accept the research hypothesis and reject the null hypothesis, but he does not and the opposite happens. The probability that a type II error is committed is represented by beta (β).

Types of Statistical Tests

One-tailed and two-tailed are two types of statistical tests that are used alternatively for the computation of the statistical significance of some parameter in a given set of data. These are also termed as one-sided and two-sided tests.

In research, the one-tailed test can be used when the deviations of the estimated parameter in one direction from an assumed benchmark value are considered theoretically possible.

On the other hand, the two-tailed test should be utilized when the deviations in both directions of benchmark value are considered as theoretically possible.The word "tail" is used in the names on these tests since the extreme points of the distributions in which observations tend to reject the null hypothesis are quite small and "tail off" to zero similar to the bell curve or normal distribution. The choice of one-tailed or two-tailed significance test depends upon the research hypothesis.

Example: The one-tailed test can be utilized for the test of the null hypothesis such as, boys will not score significantly higher marks than girls in 10 Standard. In this example, the null hypothesis does indirectly assume the direction of the difference.

The two-tailed test could be utilized in the testing of the null hypotheses: There is no significant difference in scores of boys and girls in 10 Standard.

## 13.3  What is p-Value Testing?

In the context of the statistical significance of a data, the p-value is an important terminology for hypothesis testing. The p-value is said to be a function of observed sample results which is being used for testing of statistical hypothesis. A threshold value is to be selected before the test is performed. This value is known as the significance level that is traditionally 1% or 5%. It is denoted by α.

In the case when the p-value is smaller than or equal to significance level (α), the data is said to be inconsistent for our assumption of the null hypothesis to be true. Therefore, the null hypothesis should be rejected and an alternative hypothesis is supposed to be accepted or assumed as true.

Note that the smaller the p-value is, the bigger the significance should be as it indicates that the research hypothesis does not adequately explain the observation. If the p-value is calculated accurately, then such test controls type I error rate not to be greater than the significance level (α). The use of p-values in statistical hypothesis testing is very commonly seen in a wide variety of areas such as psychology, sociology, science, economics, social science, biology, criminal justice etc.

### An introduction to t-tests

A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

You want to know whether the mean petal length of iris flowers differs according to their species. You find two different species of irises growing in a garden and measure 25 petals of each species. You can test the difference between these two groups using a t-test.

- The null hypothesis ($H_0$) is that the true difference between these group means is zero.
- The alternate hypothesis ($H_a$) is that the true difference is different from zero.

### When to use a t-test

A t-test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test or a post-hoc test.

*Probability And Statistics*

The t-test is a parametric test of difference, meaning that it makes the same assumptions about your data as other parametric tests. The t-test assumes your data:

- Are independent
- Are (approximately) normally distributed.

  Have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)

If your data do not fit these assumptions, you can try a nonparametric alternative to the t-test, such as the Wilcoxon Signed-Rank test for data with unequal variances.

**What type of t-test should I use?**

When choosing a t-test, you will need to consider two things: whether the groups being compared come from a single population or two different populations, and whether you want to test the difference in a specific direction.

**One-sample, two-sample, or paired t-test?**

If the groups come from a single population (e.g. measuring before and after an experimental treatment), perform a paired t-test.

If the groups come from two different populations (e.g. two different species, or people from two separate cities), perform a two-sample t-test (a.k.a. independent t-test).

If there is one group being compared against a standard value (e.g. comparing the acidity of a liquid to a neutral pH of 7), perform a one-sample t-test.

**One-tailed or two-tailed t-test?**

If you only care whether the two populations are different from one another, perform a two-tailed t-test.

If you want to know whether one population mean is greater than or less than the other, perform a one-tailed t-test.

Example: In your test of whether petal length differs by species:

Your observations come from two separate populations (separate species), so you perform a two-sample t-test.

You don't care about the direction of the difference, only whether there is a difference, so you choose to use a two-tailed t-test.

## 13.4  Z-test

Z-test is a statistical method to determine whether the distribution of the test statistics can be approximated by a normal distribution. It is the method to determine whether two sample means are approximately the same or different when their variance is known and the sample size is large (should be >= 30).

**When to Use Z-test:**

The sample size should be greater than 30. Otherwise, we should use the t-test. Samples should be drawn at random from the population.  standard deviation of the population should be known. Samples that are drawn from the population should be independent of each other. The data should be normally distributed, however for large sample size, it is assumed to have a normal distribution

**Hypothesis Testing**

A hypothesis is an educated guess/claim about a particular property of an object. Hypothesis testing is a way to validate the claim of an experiment.

- **Null Hypothesis:** The null hypothesis is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to some claimed value. We either reject or fail to reject the null hypothesis. Null Hypothesis is denoted by $H_0$.
- **Alternate Hypothesis:** The alternative hypothesis is the statement that the parameter has a value that is different from the claimed value. It is denoted by $H_A$.

**Level of significance:** It means the degree of significance in which we accept or reject the null hypothesis. Since in most of the experiments 100% accuracy is not possible for accepting or rejecting a hypothesis, so we, therefore, select a level of significance. It is denoted by alpha ($\propto$).

**Steps to perform Z-test:**
- First, identify the null and alternate hypotheses.
- Determine the level of significance ($\propto$).
- Find the critical value of z in the z-test using
- Calculate the z-test statistics. Now compare with the hypothesis and decide whether to reject or not to reject the null hypothesis

## 13.5  Type of Z-test

- **Left-tailed Test:** In this test, our region of rejection is located to the extreme left of the distribution. Here our null hypothesis is that the claimed value is less than or equal to the mean population value



α
Left-tailed Test
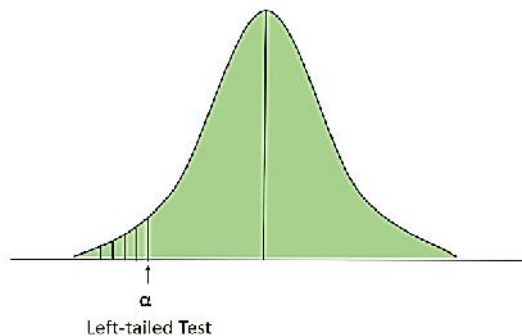
- **Right-tailed Test:** In this test, our region of rejection is located to the extreme right of the distribution. Here our null hypothesis is that the claimed value is less than or equal to the mean population value.



α
Right-tailed test
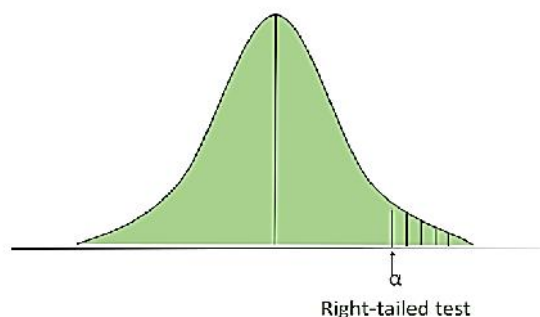
Two-tailed test: In this test, our region of rejection is located to both extremes of the distribution. Here our null hypothesis is that the claimed value is equal to the mean population value

α/2    Double-tailed Test    α/2
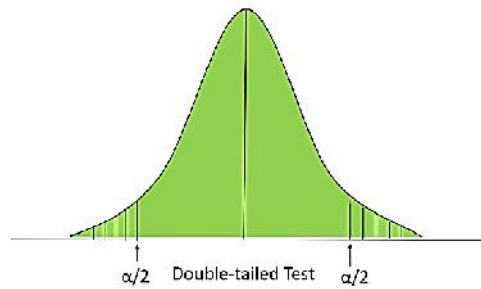
## 13.6  Key Differences Between T-test and Z-test

The difference between t-test and z-test can be drawn clearly on the following grounds:

1. The t-test can be understood as a statistical test which is used to compare and analyses whether the means of the two population is different from one another or not when the standard deviation is not known. As against, Z-test is a parametric test, which is applied when the standard deviation is known, to determine, if the means of the two datasets differ from each other.
2. The t-test is based on Student's t-distribution. On the contrary, z-test relies on the assumption that the distribution of sample means is normal. Both student's t-distribution and normal distribution appear alike, as both are symmetrical and bell-shaped. However, they differ in the sense that in a t-distribution, there is less space in the centre and more in the tails.
3. One of the important conditions for adopting t-test is that population variance is unknown. Conversely, population variance should be known or assumed to be known in case of a z-test.
4. Z-test is used to when the sample size is large, i.e. $n > 30$, and t-test is appropriate when the size of the sample is small, in the sense that $n < 30$.

## 13.7  What is the Definition of F-Test Statistic Formula?

It is a known fact that statistics is a branch of Mathematics that deals with the collection, classification and representation of data. The tests that use F - distribution is represented by a single word in Statistics called F test. F test is usually used as a generalized statement for comparing two variances. F test statistic formula is used in various other tests such as regression analysis, the chow test and Scheffe test. F tests can be conducted by using several technological aids. However, manual calculation of is a little complex and time consuming. This article gives an in detail description of the F test formula and its usage.

**Definition of F-Test Formula**

F test is a test statistic that has an F distribution under the null hypothesis. It is used in comparing the statistical model with respect to the available data set. The name for the test is given in the honor of Sir. Ronald A Fisher by George W Snedecor. To perform an F test using technology, the following aspects are to be taken care of.

Example:

State the null hypothesis along with the alternative hypothesis.

Compute the value of 'F' with the help of the standard formula.

Determine the value of the F statistic. The ratio of variance of the group of means to the mean of the within group variances.

As the last step, support or reject the Null hypothesis.

F-Test Equation to Compare Two Variances:

In statistics, the F-test formula is used to compare two variances, say σ1 and σ2, by dividing them. As the variances are always positive, the result will also be positive always. Hence, the F test equation used to compare two variances is given as:

F_value = variance1/variance2

i.e. F_value = $\sigma^2 1 / \sigma^2 2$

F test formula helps us to compare the variances of two different sets of values. To use F distribution under null hypothesis, it is important to determine the mean of the two given observations at first and then calculate the variance. σ2

$$= \Sigma \, (x - \bar{x})^2 \, / n - 1$$

In the above formula,

$\sigma^2$ is the variance

x is the values given in a set of data

x bar  is the mean of the given data set

n is the total number of values in the data set

While running an F test, it is very important to note that the population variances are equal. In more simple words, it is always assumed that the variances are equal to unity or 1. Therefore, the variances are always equal in case of null hypothesis.

## 13.8  F Test Statistic Formula Assumptions

F test equation involves several assumptions. In order to use the F - test formula, the population should be distributed normally. The samples considered for the test should be independent events. In addition to these, it is also important to consider the following points.

Calculation of right tailed tests is easier. To force the test into a right tailed test, the larger variance is pushed in the numerator.

In case of two tailed tests, alpha is divided by two prior to determination of critical value.

Variances are the squares of the standard deviations.

If the obtained degree of freedom is not listed in the F table, it is always better to use a larger critical value to decrease the probability of type 1 errors.

F-Value Definition: Example Problems

Example 1:

Perform an F test for the following samples.

Sample 1 with variance equal to 109.63  and sample size equal to 41.

Sample 2 with variance equal to 65.99 and sample size equal to 21.

Solution:

Step 1:

The hypothesis statements are written as:

H_0: No difference in variances

**Lovely Professional University**

H_a: Difference in variances

Step 2:

Calculate the value of F critical. In this case, highest variance is taken as the numerator and the lowest variance in the denominator.

F_value = σ21σ22

F_value = 109.6365.99

F_value = 1.66

Step 3:

The next step is the calculation of degrees of freedom.

The degrees of freedom is calculated as Sample size - 1

The degree of freedom for sample 1 is 41 -1 = 40.

The degree of freedom for sample 2 is 21 - 1 = 20.

Step 4:

There is no alpha level described in the question and hence a standard alpha level of 0.05 is chosen. During the test, the alpha level should be reduced to half the initial value and hence it becomes 0.025.

Step 5:

Using the F table, the critical F value is determined with alpha at 0.025. The critical value for (40 , 20) at alpha equal to 0.025 is 2.287.

Step 6:

It is now the time for comparing the calculated value with the standard value in the table. Generally, the null hypothesis is rejected if the calculated value is greater than the table value. In this F value definition example, the calculated value is 1.66 and the table value is 2.287.

It is clear from the values that 1.66 < 2.287. Hence, null hypothesis cannot be rejected.

Fun Facts about F-Value Definition:

In case of statistical calculations where null hypothesis can be rejected, the F value can be less than 1 however, not exactly equal to zero.

The F critical value cannot be exactly equal to zero. If the F value is exactly zero, it indicates that the mean of every sample is exactly the same and the variance is zero.

## 13.9  T-test formula

T-tests can be performed either manually by using a formula or through some software.

$$t = \frac{\bar{x} + \mu}{\frac{\sigma^2}{\sqrt{n}}}$$

Where x is the mean of the sample, and µ is the assumed mean, σ is the standard deviation, and n is the number of observations.

T-test for the difference in mean:

$$t = \frac{\overline{X1} - \overline{X2}}{\sqrt{\frac{(n1-1)\sigma 1^2 + (n2-1)\sigma 2^2}{n1+n2-2}} \times \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

Where **x1** and **x2** are the mean of two samples and **σ1** and **σ2** is the standard deviation of two samples, and **n1** and **n2** are the numbers of observation of two samples.

**One sample t-test (one-tailed t-test)**

- One sample t-test is a statistical test where the critical area of a distribution is one-sided so that the alternative hypothesis is accepted if the population parameter is either greater than or less than a certain value, but not both.
- In the case where the t-score of the sample being tested falls into the critical area of a one-sided test, the alternative hypothesis is to be accepted instead of the null hypothesis.
- A one-tailed test is used to determine if the population is either lower than or higher than some hypothesized value.
- A one-tailed test is appropriate if the estimated value might depart from the sample value in either of the directions, left or right, but not both.
- For this test, the null hypothesis states that there is no difference between the true mean and the assumed value whereas the alternative hypothesis states that either the assumed value is greater than or less than the true mean but not both.
- For instance, if our $H_0$: $\mu_0 = \mu$ and $H_a$: $\mu < \mu_0$, such a test would be a one-sided test or more precisely, a left-tailed test.
- Under such conditions, there is one rejection area only on the left tail of the distribution.
- If we consider $\mu = 100$ and if our sample mean deviates significantly from 100 towards the lower direction, $H_0$ or null hypothesis is rejected. Otherwise, $H_0$ is accepted at a given level of significance.
- Similarly, if in another case, $H_0$: $\mu = \mu_0$ and $H_a$: $\mu > \mu_0$, this is also a one-tailed test (right tail) and the rejection region is present on the right tail of the curve.
- In this case, when $\mu = 100$ and the sample mean deviates significantly from 100 in the upward direction, $H_0$ is rejected otherwise, it is to be accepted.

**Two sample t-test (two-tailed t-test)**

- Two sample t-test is a test a method in which the critical area of a distribution is two-sided and the test is performed to determine whether the population parameter of the sample is greater than or less than a specific range of values.
- A two-tailed test rejects the null hypothesis in cases where the sample mean is significantly higher or lower than the assumed value of the mean of the population.
- This type of test is appropriate when the null hypothesis is some assumed value, and the alternative hypothesis is set as the value not equal to the specified value of the null hypothesis.
- The two-tailed test is appropriate when we have $H_0$: $\mu = \mu_0$ and $H_a$: $\mu \neq \mu_0$ which may mean $\mu > \mu_0$ or $\mu < \mu_0$.
- Therefore, in a two-tailed test, there are two rejection regions, one in either direction, left or right, towards each tail of the curve.
- Suppose, we take $\mu = 100$ and if our sample mean deviates significantly from 100 in either direction, the null hypothesis can be rejected. But if the sample mean does not deviate considerably from $\mu$, the null hypothesis is accepted

**Lovely Professional University**

**Independent t-test**

- An Independent t-test is a test used for judging the means of two independent groups to determine the statistical evidence to prove that the population means are significantly different.
- Subjects in each sample are also assumed to come from different populations, that is, subjects in "Sample A" are assumed to come from "Population A" and subjects in "Sample B" are assumed to come from "Population B."
- The populations are assumed to differ only in the level of the independent variable.
- Thus, any difference found between the sample means should also exist between population means, and any difference between the population means must be due to the difference in the levels of the independent variable.
- Based on this information, a curve can be plotted to determine the effect of an independent variable on the dependent variable and vice versa

**T-test example**

If a sample of 10 copper wires is found to have a mean breaking strength of 527 kgs, is it feasible to regard the sample as a part of a large population with a mean breaking strength of 578 kgs and a standard deviation of 12.72 kgs? Test at 5% level of significance.

Taking the null hypothesis that the mean breaking strength of the population is equal to 578 kgs, we can write:

H0 : μ = 578 kgs

Ha : μ ≠ 578 kgs

x = 527 kgs ,σ = 12.72 , n = 10.

Based on the assumption that the population to be normal, the formula for the test statistic t can be written as:

T-test formula 1

$$t = \frac{\bar{x} + \mu}{\dfrac{\sigma^2}{\sqrt{n}}}$$

t= (527+578) / (12.722/√10)

As Ha is two-sided in the given question, a two-tailed test is to used for the determination of the rejection regions at a 5% level of significance which comes to as under, using normal curve area table:

R : | t | > 1.96

The observed value of t is -1.488 which is in the acceptance region since R: | t | > 1.96, and thus, H0 is accepted.

## 13.10 Fisher Z Transformation

The Fisher Z transformation is approximately a variance transformation for the sample correlation coefficient for two variables X and Y following a bivariate normal distribution. Fisher Z transformation was developed by Fisher. This indicates that the variance of z is close to the constant for all values assigned to the population correlation coefficient ?.

Fisher Z transformation is a method that transforms the Pearson's correlation coefficient r to the normally distributed variable z. The Z in the Fisher Z transformation stands for the normal z-score. It is named after Fisher who developed this transformation.

The uses of Fisher Z transformation are listed below:

The Fisher Z transformation is used to estimate the confidence interval for both correlation coefficients and the differences between two correlations. It is most usually used to test the significance of the difference between the correlation coefficients of two independent random samples.

It is mainly applicable to the Pearson's correlation coefficient. But it is also relevant to Spearman's rank correlation coefficient in some cases.

The formula for Fisher Z transformation is defined as:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

Here, ln is the natural logarithm and r is the coefficient of correlation.

There are two main characteristics of the distribution of the z statistics. They are as follows:

1. Standard error is known.

2. It is normal.

The applications of Fisher Z transformation are provided below:

1. To test whether a provided value is equal to the population correlation or not.

2. To test the equality of two population correlations.

3. To combine correlation estimates obtained from different samples.

## Summary

- A t-test is a type of inferential statistic used **to determine if there is a significant difference between the means of two groups**, which may be related in certain features.

- Z Test is **the statistical hypothesis** which is used in order to determine that whether the two samples means calculated are different in case the standard deviation is available and sample is large

- If you are studying one group, use a paired t-test to compare the group mean over time or after an intervention, or use a one-sample t**-test to compare the group mean to a standard value**.

- If you are studying two groups, use a two-sample t-test. If you want to know only whether a difference exists, use a two-tailed test.

- The common assumptions made when doing a t-test include those regarding **the scale of measurement, random sampling, normality of data distribution, adequacy of sample size**, and equality of variance in standard deviation.

# Keywords

- T-test refers to a type of parametric test that is applied to identify, how the means of two sets of data differ from one another when variance is not given
- T-test refers to a type of parametric test that is applied to identify, how the means of two sets of data differ from one another when variance is not given
- The t-test is based on Student's t-distribution. On the contrary, z-test relies on the assumption that the distribution of sample means is normal.
- Both student's t-distribution and normal distribution appear alike, as both are symmetrical and bell-shaped. However, they differ in the sense that in a t-distribution, there is less space in the centre and more in the tails.

# Self Assessment

1. A _____ is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.

A. T-test
B. Regression
C. Correlation
D. Factor analysis

2. Data value different from normal behavior of data in data set are

A. Analyzer
B. Outlier
C. Mean value
D. None of these

3. If sample size is less than 30 then which test is recommended

A. T test
B. Z test
C. None of these
D. Both of these

4. If sample size is more than 30 then which test is recommended

A. T test
B. Z test
C. None of these
D. Both of these

5. _____ is used in order to determine a how averages of different data sets differs from each other in case standard deviation or the variance is not known.

A. T test

B. Z test

C. None of these

D. Both of these

6. _____ is the statistical hypothesis which is used in order to determine that whether the two samples means calculated are different in case the standard deviation is available and sample is large.

A. T test

B. Z test

C. None of these

D. Both of these

7. _____, is a number representing how many standard deviations above or below the mean population the score derived from a z-test is.

A. Z score

B. R score

C. T score

D. X score

8. _____ can also be used to check if the data conforms to a regression model, which is acquired through least square analysis.

A. F test

B. M test

C. Z test

D. T test

9. Which test is suitable for comparing the means of two populations?

A. T test

B. F test

C. Regression

D. Correlation

10. A _____is the value of the test statistic which defines the upper and lower bounds of a confidence interval.

A. Critical value

B. Regression value

C. Correlation value

D. None of these

11. The graph for the _____is similar to the standard normal curve.

A. Student's t-distribution

B. Student's f-distribution

C. Student's z-distribution

*Probability And Statistics*

D. Student's g-distribution

12._____is a statistical test where the critical area of a distribution is one-sided so that the alternative hypothesis is accepted if the population parameter is either greater than or less than a certain value.

A. One sample t-test

B. Regression test

C. Correlation test

D. None of these

13. Variables are said to be _____if the changes in one variable results in a corresponding change in the other variable.

A. Correlated

B. Interpreted

C. None of these

D. Both of these

14. When the change in the two variables is such that with an increase in the value of one, the value of the other increases in a fixed proportion.

A. Perfect correlation

B. No correlation

C. Limited degree of correlation

D. All of these

15. If the changes in the value of one variable are not in association with the changes in the value of other variable there will be no correlation

A. Perfect correlation

B. No correlation

C. Limited degree of correlation

D. All of these

## Answers for Self Assessment

| 1. | A | 2. | B | 3. | A | 4. | B | 5. | A |
|----|---|----|---|----|---|----|---|----|---|
| 6. | B | 7. | A | 8. | A | 9. | A | 10. | A |
| 11. | A | 12. | A | 13. | A | 14. | A | 15. | B |

## Review Questions

- What is T test used for, explain it with example?
- What is use and applications of Z test?
- Explain the difference between T test and Z test?
- What is an example of an independent t test?

**Lovely Professional University**

- What is the difference between independent sample and one sample t test?
- Is F-test and ANOVA the same?
- What is p value in ANOVA?
- What is Fisher R to Z transformation?
- How do you convert correlation to z score?

## Further Readings

**An Introduction to Probability and Statistics**

Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

**First Course in Probability, A**

Book by Sheldon M. Ross

**Schaums Theory and Problems of Statistics**

Book by Murray R. Spiegel

**Introduction to Probability, Statistics, and Random ...**

Book by Hossein Pishro-Nik

**Web links**

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability

# Unit 14: Statistical Tools and Techniques

| CONETNTS |
| --- |
| Objectives |
| Introduction |
| 14.1　What Is Bayes' Theorem? |
| 14.2　How to Use Bayes Theorem for Business and Finance |
| 14.3　Bayes Theorem of Conditional Probability |
| 14.4　Naming the Terms in the Theorem |
| 14.5　What is SPSS? |
| 14.6　Features of SPSS |
| 14.7　R Programming Language – Introduction |
| 14.8　Statistical Features of R: |
| 14.9　Programming Features of R: |
| 14.10　Microsoft Excel |
| Keywords |
| Summary |
| Self Assessment |
| Answers for Self Assessment |
| Review Questions |
| Further Readings |

## Objectives

- Understand basics of Bayes' theorem.
- Learn concepts of Bayes' theorem
- Define basic terms of spss.
- Understand concept of R and excel in data analysis.
- Solve basic questions related to probability

## Introduction

Bayes' theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence. In finance, Bayes' theorem can be used to rate the risk of lending money to potential borrowers.

Bayes' theorem is also called Bayes' Rule or Bayes' Law and is the foundation of the field of Bayesian statistics.

Applications of the theorem are widespread and not limited to the financial realm. As an example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test.

Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities.

## 14.1  What Is Bayes' Theorem?

Prior probability, in Bayesian statistical inference, is the probability of an event before new data is collected. This is the best rational assessment of the probability of an outcome based on the current knowledge before an experiment is performed.

 Posterior probability is the revised probability of an event occurring after taking into consideration new information. Posterior probability is calculated by updating the prior  probability by using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

Bayes' theorem thus gives the probability of an event based on new information that is, or may be related, to that event.

The formula can also be used to see how the probability of an event occurring is affected by hypothetical new information, supposing the new information will turn out to be true.

For instance, say a single card is drawn from a complete deck of 52 cards. The probability that the card is a king is four divided by 52,

Which equals 1/13 or approximately 7.69%.

Example: Remember that there are four kings in the deck. Now, suppose it is revealed that the selected card is a face card. The probability the selected card is a king, given it is a face card, is four divided by 12, or approximately 33.3%, as there are 12 face cards in a deck.

## Formula For Bayes' Theorem

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

**where:**

$P(A) =$ The probability of A occurring

$P(B) =$ The probability of B occurring

$P(A|B) =$ The probability of A given B

$P(B|A) =$ The probability of B given A

$P\left(A \cap B\right)) =$ The probability of both A and B occurring

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

P(Fire|Smoke) means how often there is fire when we can see smoke
P(Smoke|Fire) means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know
"backwards" P(Smoke|Fire)

Example:

- **Dangerous fires are rare (1%)**

- **But smoke is fairly common (10%) due to barbecues,**
- **And 90% of dangerous fires make smoke**

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$P(Fire|Smoke) = \frac{P(Fire)\ P(Smoke|Fire)}{P(Smoke)}$$

$$= \frac{1\% \ x\ 90\%}{10\%}$$

$$= 9\%$$

So it is still worth checking out any smoke to be sure.

Example: Picnic Day

You are planning a picnic today, but the morning is cloudy

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

**What is the chance of rain during the day?**

We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.

The chance of Rain given Cloud is written P(Rain|Cloud)

So, let's put that in the formula:

$$P(Rain|Cloud) = \frac{P(Rain)\ P(Cloud|Rain)}{P(Cloud)}$$

- P(Rain) is Probability of Rain = 10%
- P(Cloud|Rain) is Probability of Cloud, given that Rain happens = 50%
- P(Cloud) is Probability of Cloud = 40%

$$P(Rain|Cloud) = \frac{0.1\ x\ 0.5}{0.4} = .125$$

Or a 12.5% chance of rain. Not too bad, let's have a picnic!

Task: Discuss real life application of Bayes' theorem?

## 14.2 How to Use Bayes Theorem for Business and Finance

In finance and business circles, corporate financial specialists have been applying Bayes' Theorem for centuries.

Consider these applications:

In evaluating interest rates. Companies rely on interest rates for multiple reasons - borrowing money, investing in the fixed income market, and trading in currencies overseas. Any unexpected shifts in interest rate values can hit a company hard in the pocketbook, and can negatively impact profits and revenues. With Bayes Theorem and estimated probabilities, companies can better evaluate systematic changes in interest rates, and steer their financial resources to take maximum advantage.

With net income. Businesses are keen to be on top of their net income streams, or the profit a business earns after subtracting expenses out of the equation. Net income is highly vulnerable to external events, like legal proceedings, weather, the cost of necessary equipment and materials, and geopolitical events, for starters. Plugging probability scenarios into the net income equation when these scenarios arise gives financial decision makers a stronger platform when managing resources and making critical decisions.

For extending credit. Under the Bayes Theorem conditional probability model, financial companies can make better decisions and better evaluate the risk of lending cash to unfamiliar or even existing borrowers. For example, an existing client may have had a good previous track record of repaying loans, but lately the client has been slow in playing This additional information, based on probability theory, can lead the company to treat the slow payment history as a red flag, and either hike interest rates on the loan or reject it altogether.

Example:

**The Enigma code**
In 1774, the brilliant French mathematician Pierre-Simon Laplace expanded upon Bayes' theorem, before the theorem all but disappeared from sight until the 20th Century, when British codebreaker Alan Turing used it during the Second World War to help crack the 'unbreakable' Enigma code, a development that helped the Allies win the war.

Turing developed a system based on Bayesian theory that enabled him to guess a stretch of letters in an Enigma message, calculate the probabilities, and add more clues as they arrived. With this method he could reduce the number of wheel settings to be tested, which subsequently led him to cracking the code.

With the advent of the computer age, the use of Bayesian theory has exploded, into such areas as artificial intelligence, robotics, law, imaging technologies and medical diagnostics. In 1996, Bill Gates said that Microsoft's competitive advantage was its use of Bayesian networks. Bayes techniques are also used in spam filters, voice recognition systems, recommendation systems and in Google search.

Despite Bayes' theorem being a clever mathematical formula, the good news is that you don't need to be a mathematician to be able to apply Bayesian thinking to investing or your everyday life

## 14.3    Bayes Theorem of Conditional Probability

Before we dive into Bayes theorem, let's review marginal, joint, and conditional probability.

Recall that marginal probability is the probability of an event, irrespective of other random variables. If the random variable is independent, then it is the probability of the event directly, otherwise, if the variable is dependent upon other variables, then the marginal probability is the probability of the event summed over all outcomes for the dependent variables, called the sum rule.

**Marginal Probability**: The probability of an event irrespective of the outcomes of other random variables, e.g. $P(A)$.

The joint probability is the probability of two (or more) simultaneous events, often described in terms of events A and B from two dependent random variables, e.g. X and Y. The joint probability is often summarized as just the outcomes, e.g. A and B.

**Joint Probability**: Probability of two (or more) simultaneous events, e.g. $P(A \text{ and } B)$ or $P(A, B)$.

The conditional probability is the probability of one event given the occurrence of another event, often described in terms of events A and B from two dependent random variables e.g. X and Y.

**Conditional Probability**: Probability of one (or more) event given the occurrence of another event, e.g. P(A given B) or P(A | B).

The joint probability can be calculated using the conditional probability; for example:

P (A, B) = P(A | B) * P(B)

This is called the product rule. Importantly, the joint probability is symmetrical, meaning that:

P (A, B) = P(B, A)

The conditional probability can be calculated using the joint probability; for example:

P (A | B) = P(A, B) / P(B)

The conditional probability is not symmetrical; for example:

P (A | B) != P(B | A)

We are now up to speed with marginal, joint and conditional probability. If you would like more background on these fundamentals, see the tutorial:

**A Gentle Introduction to Joint, Marginal, and Conditional Probability**

An Alternate Way To Calculate Conditional Probability

Now, there is another way to calculate the conditional probability.

Specifically, one conditional probability can be calculated using the other conditional probability; for example:

P(A|B) = P(B|A) * P(A) / P(B)

The reverse is also true; for example:

P(B|A) = P(A|B) * P(B) / P(A)

This alternate approach of calculating the conditional probability is useful either when the joint probability is challenging to calculate (which is most of the time), or when the reverse conditional probability is available or easy to calculate.

This alternate calculation of the conditional probability is referred to as Bayes Rule or Bayes Theorem, named for Reverend Thomas Bayes, who is credited with first describing it. It is grammatically correct to refer to it as Bayes' Theorem (with the apostrophe), but it is common to omit the apostrophe for simplicity.

Task: What is difference between conditional probability and Bayes Rule?

## 14.4 Naming the Terms in the Theorem

The terms in the Bayes Theorem equation are given names depending on the context where the equation is used.

It can be helpful to think about the calculation from these different perspectives and help to map your problem onto the equation.

Firstly, in general, the result P(A|B) is referred to as the **posterior probability** and P(A) is referred to as the **prior probability**.

- P(A|B): Posterior probability.
- P(A): Prior probability.

Sometimes P(B|A) is referred to as the **likelihood** and P(B) is referred to as the **evidence**.
- P(B|A): Likelihood.
- P(B): Evidence.

This allows Bayes Theorem to be restated as:

- Posterior = Likelihood * Prior / Evidence

We can make this clear with a smoke and fire case.

**What is the probability that there is fire given that there is smoke?**

Where P(Fire) is the Prior, P(Smoke|Fire) is the Likelihood, and P(Smoke) is the evidence:

- P(Fire|Smoke) = P(Smoke|Fire) * P(Fire) / P(Smoke)

You can imagine the same situation with rain and clouds.

## 14.5  What is SPSS?

SPSS stands for "Statistical Package for the Social Sciences". It is an IBM tool. This tool first launched in 1968. This is one software package. This package is mainly used for statistical analysis of the data.SPSS is mainly used in the following areas like healthcare, marketing, and educational research, market researchers, health researchers, survey companies, education researchers, government, marketing organizations, data miners, and many others.It provides data analysis for descriptive statistics, numeral outcome predictions, and identifying groups. This software also gives data transformation, graphing and direct marketing features to manage data smoothly.

Why SPSS?

They came under IBM SPSS Statistics, and most of the users refer to it as SPSS only.It is straight forward, and its English-like command language helps the user to go through the flow.SPSS introduces the following four programs that help researchers with their complex data analysis needs.

**Statistics Program**

SPSS's statistics program gives a large amount of basic statistical functionality; some include frequencies, cross-tabulation, bivariate statistics, etc.

**Modeler Program**

Researchers are able to build and validate predictive models with the help of advanced statistical procedures.

Text Analytics for Surveys Program

It gives robust feedback analysis. which in turn get a vision for the actual plan.

**Visualization Designer**

Researchers found this visual designer data to create a wide variety of visuals like density charts and radial box plots.

## 14.6  Features of SPSS

The data from any survey collected gets easily exported to SPSS for detailed and good analysis.

In SPSS, data gets stored in.SAV format. These data mostly come from surveys. This makes the process of manipulating, analyzing and pulling data very simple.

SPSS have easy access to data with different variable types. These variable data is easy to understand. SPSS helps researchers to set up model easily because most of the process is automated.

After getting data in the magic of SPSS starts. There is no end to what we can do with this data.

SPSS has a unique way to get data from critical data also. Trend analysis, assumptions, and predictive models are some of the characteristics of SPSS.

SPSS is easy for you to learn, use and apply.

It helps in to get data management system and editing tools handy.

SPSS offers you in-depth statistical capabilities for analyzing the exact outcome.

SPSS helps us to design, plotting, reporting and presentation features for more clarity.

**Example: Statistical Methods of SPSS**

Many statistical methods can be used in SPSS, which are as follows:

- Prediction for a variety of data for identifying groups and including methodologies such as cluster analysis, factor analysis, etc.
- Descriptive statistics, including the methodologies of SPSS, are frequencies, cross-tabulation, and descriptive ratio statistics, which are very useful.
- Also, Bivariate statistics, including methodologies like analysis of variance (ANOVA), means, correlation, and nonparametric tests, etc.
- Numeral outcome prediction such as linear regression.

It is a kind of self-descriptive tool which automatically considers that you want to open an existing file, and with that opens a dialog box to ask which file you would like to open. This approach of SPSS makes it very easy to navigate the interface and windows in SPSS if we open a file.

Besides the statistical analysis of data, the SPSS software also provides data management features; this allows the user to do a selection, create derived data, perform file reshaping, etc. Another feature is data documentation. This feature stores a metadata dictionary along with the data file.

**It has two types of views those are Variable View and Data View:**

**Variable View**

Name: This is a column field, which accepts the unique ID. This helps in sorting the data. For example, the different demographic parameters such as name, gender, age, educational qualification are the parameters for sorting data.

The only restriction is special characters which are not allowed in this type.

Label: The name itself suggests it gives the label. Which also gives the ability to add special characters.

Type: This is very useful when different kind of data's are getting inserted.

Width: We can measure the length of characters.

Decimal: While entering the percentage value, this type helps us to decide how much one needs to define the digits required after the decimal.

Value: This helps the user to enter the value.

Missing: This helps the user to skip unnecessary data which is not required during analysis.

Align: Alignment, as the name suggests, helps to align left or right. But in this case, for ex. Left align.

Measure: This helps to measure the data being entered in the tools like ordinal, cardinal, nominal.

The data has to enter in the sheet named "variable view". It allows us to customize the data type as required for analyzing it.

To analyze the data, one needs to populate the different column headings like Name, Label, Type, Width, Decimals, Values, Missing, Columns, Align, and Measures.

These headings are the different attributes which, help to characterize the data accordingly.

**Data View**

The data view is structured as rows and columns. By importing a file or adding data manually, we can work with SPSS.

SPSS Data View & Variable View

An SPSS data file *always* has two tabs in the left bottom corner:

- **Data View** is where we inspect our actual data and

- **Variable View** is where we see additional information about our data.
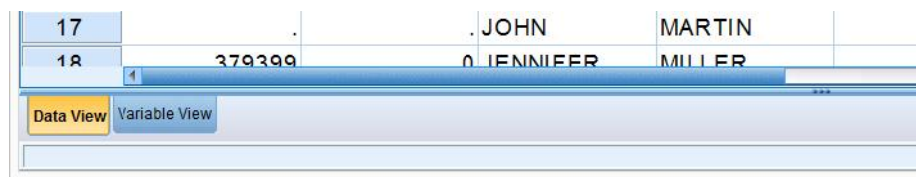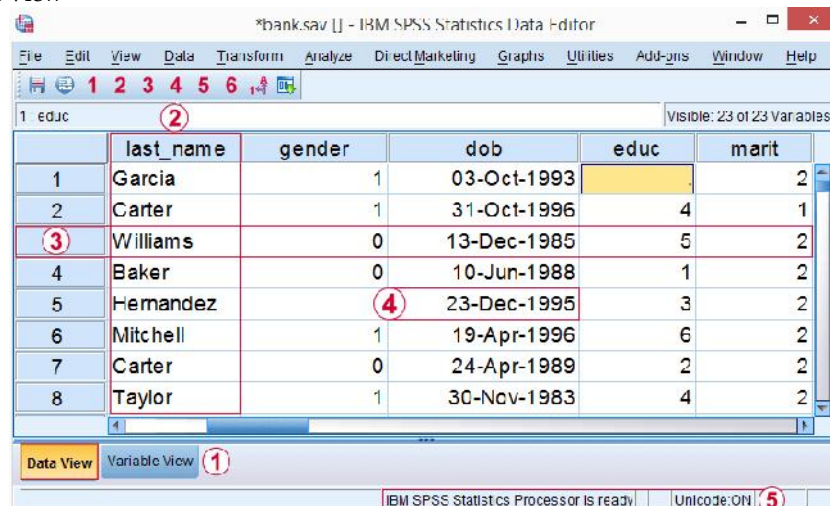


You can **switch** between Data View and Variable View by

- clicking the tabs in the left bottom corner;

- using the Ctrl + t shortkey;

- double-clicking a variable name in Data View;

- double-clicking an outline number in Variable View.

Let's first take a close look at the main parts of the Data View tab. We'll then proceed with variable

view.

SPSS Data View



1.  The data editor has **tabs** for switching between Data View and Variable View. For now, make sure you're in Data View.

2.  Columns of cells are called **variables**. Each variable has a unique name ("gender") which is shown in the column header.

**Lovely Professional University**

3.   Rows of cells are called **cases**. Oftentimes, each respondent in a study is represented as a single case.

4.   In SPSS, **values** refer to cell contents.

5.   The **status bar** may give useful information on the data, for instance whethera WEIGHT, FILTER, SPLIT FILE or Unicode mode is in effect.

**SPSS Variable View**



1. In the left bottom corner we find **tabs** for switching between Variable View and Data View. For now, select Variable View.

2. In Variable View, **variables** are shown as rows of cells. The first column shows the **variable name** for each variable.

3. The fifth column may or may not contain a **variable label**. This describes the exact meaning of each variable.

4. The sixth column shows **value labels**: descriptions of the meaning of one, many or all values that a variable may contain.

5. In short, Variable View does not show the data itself but, rather, information *about* the data. This is sometimes called "metadata" or "the codebook".

## 14.7 R Programming Language – Introduction

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.

It was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R programming language is an implementation of the S programming language. It also combines with lexical scoping semantics inspired by Scheme. Moreover, the project conceives in 1992, with an initial version released in 1995 and a stable beta version in 2000.

Why R Programming Language?

R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R.It's a platform-independent language. This means it can be applied to all operating system.It's an open-source free language. That means anyone can install it in any organization without purchasing a license.

R programming language is not only a statistic package but also allows us to integrate with other languages (C, C++). Thus, you can easily interact with many data sources and statistical packages.The R programming language has a vast community of users and it's growing day by day.R is currently one of the most requested programming languages in the Data Science job market that makes it the hottest trend nowadays.

## 14.8  Statistical Features of R:

Basic Statistics: The most common basic statistics terms are the mean, mode, and median. These are all known as "Measures of Central Tendency." So, using the R language we can measure central tendency very easily.

Static graphics: R is rich with facilities for creating and developing interesting static graphics. R contains functionality for many plot types including graphic maps, mosaic plots, biplots, and the list goes on.

Probability distributions: Probability distributions play a vital role in statistics and by using R we can easily handle various types of probability distribution such as Binomial Distribution, Normal Distribution, Chi-squared Distribution and many more.

Data analysis: It provides a large, coherent and integrated collection of tools for data analysis.

## 14.9  Programming Features of R:

R Packages: One of the major features of R is it has a wide availability of libraries. R has CRAN(Comprehensive R Archive Network), which is a repository holding more than 10, 0000 packages.

Distributed Computing: Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Two new packages ddR and multidplyr used for distributed programming in R were released in November 2015.

**Programming in R:**

Since R is much similar to other widely used languages syntactically, it is easier to code and learn in R. Programs can be written in R in any of the widely used IDE like R Studio, Rattle, Tinn-R, etc. After writing the program save the file with the extension.r.

**Advantages of R:**

R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.

As R programming language is an open source. Thus, you can run R anywhere and at any time.

R programming language is suitable for GNU/Linux and Windows operating system.

R programming is cross-platform which runs on any operating system.

In R, everyone is welcome to provide new packages, bug fixes, and code enhancements.

**Disadvantages of R:**

In the R programming language, the standard of some packages is less than perfect.

Although, R commands give little pressure to memory management. So, R programming language may consume all available memory.

In R basically, nobody to complain if something doesn't work.

R programming language is much slower than other programming languages such as Python and MATLAB.

**Applications of R:**

We use R for Data Science. It gives us a broad variety of libraries related to statistics. It also provides the environment for statistical computing and design.R is used by many quantitative analysts as its programming tool. Thus, it helps in data importing and cleaning.R is the most prevalent language. So many data analysts and research programmers use it. Hence, it is used as a fundamental tool for finance.Tech giants like Google, Facebook, bing, Twitter, Accenture, Wipro and many more using R nowadays.R and Python both play a major role in data science. It becomes confusing for any newbie to choose the better or the most suitable one among the two, R and Python. So, take a look at R vs Python for Data Science to choose which language is more suitable for data science.

**Statistics**

R and its libraries implement various statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, spatial and time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and its community is noted for contributing packages. Many of R's standard functions are written in R, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time.

Another of R's strengths is static graphics; it can produce publication-quality graphs that include mathematical symbols. Dynamic and interactive graphics are available through additional packages.

**Programming**

R is an interpreted language; users typically access it through a command-line interpreter. If a user types 2+2 at the R command prompt and presses enter, the computer replies with 4.

Like languages such as APL and MATLAB, R supports matrix arithmetic. R's data structures include vectors, matrices, arrays, data frames (similar to tables in a relational database) and lists. Arrays are stored in column-major order. R's extensible object system includes objects for (among others): regression models, time-series and geo-spatial coordinates. R has no scalar data type. Instead, a scalar is represented as a length-one vector.

Many features of R derive from Scheme. R uses S-expressions to represent both data and code. Functions are first-class objects and can be manipulated in the same way as data objects, facilitating meta-programming that allows multiple dispatch. Variables in R are lexically scoped and dynamically typed. Function arguments are passed by value, and are lazy — that is to say, they are only evaluated when they are used, not when the function is called.

R supports procedural programming with functions and, for some functions, object-oriented programming with generic functions. A generic function acts differently depending on the classes of the arguments passed to it. In other words, the generic function dispatches the method implementation specific to that object's class. For example, R has a generic print function that can print almost every class of object in R with print(object name)

Although used mainly by statisticians and other practitioners seeking an environment for statistical computation and software development, R can also operate as a general matrix calculation toolbox – with performance benchmarks comparable to GNU Octave or MATLAB.

## 14.10 Microsoft Excel

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA). Excel forms part of the Microsoft Office suite of software.

Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering, and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives (using pivot tables and the scenario manager). A PivotTable is a tool for data analysis. It does this by simplifying large data sets via PivotTable fields It has a programming aspect, Visual Basic for Applications, allowing the user to employ a wide variety of numerical methods, for example, for solving differential equations of mathematical physics, and then reporting the results back to the spreadsheet. It also has a variety of interactive features allowing user interfaces that can completely hide the spreadsheet from the user, so the spreadsheet presents itself as a so-called application, or decision support system (DSS), via a custom-designed user interface, for example, a stock analyzer, or in general, as a design tool that asks the user questions and provides answers and reports. In a more elaborate realization, an Excel application can automatically poll external databases and measuring instruments using an update schedule, analyze the results, make a Word report or PowerPoint slide show, and e-mail these presentations on a regular basis to a list of participants. Excel was not designed to be used as a database.

Microsoft allows for a number of optional command-line switches to control the manner in which Excel starts using other Windows applications such as Microsoft Access and Microsoft Word, as well as Excel can communicate with each other and use each other's capabilities. The most common are Dynamic Data Exchange: although strongly deprecated by Microsoft, this is a common method to send data between applications running on Windows, with official MS publications referring to it as "the protocol from hell". As the name suggests, it allows applications to supply data to others for calculation and display. It is very common in financial markets, being used to connect to important financial data services such as Bloomberg and Reuters.

OLE Object Linking and Embedding allows a Windows application to control another to enable it to format or calculate data. This may take on the form of "embedding" where an application uses another to handle a task that it is more suited to, for example a PowerPoint presentation may be embedded in an Excel spreadsheet or vice versa.

**Using external data**

Excel users can access external data sources via Microsoft Office features such as (for example) .odc connections built with the Office Data Connection file format. Excel files themselves may be updated using a Microsoft supplied ODBC driver.

Excel can accept data in real-time through several programming interfaces, which allow it to communicate with many data sources such as Bloomberg and Reuters (through add-ins such as Power Plus Pro).

DDE: "Dynamic Data Exchange" uses the message passing mechanism in Windows to allow data to flow between Excel and other applications. Although it is easy for users to create such links, programming such links reliably is so difficult that Microsoft, the creators of the system, officially refer to it as "the protocol from hell". In spite of its many issues DDE remains the most common way for data to reach traders in financial markets.

Network DDE Extended the protocol to allow spreadsheets on different computers to exchange data. Starting with Windows Vista, Microsoft no longer supports the facility.

Real Time Data: RTD although in many ways technically superior to DDE, has been slow to gain acceptance, since it requires non-trivial programming skills, and when first released was neither adequately documented nor supported by the major data vendors.

Alternatively, Microsoft Query provides ODBC-based browsing within Microsoft Excel

**Why Excel Is Still Essential to Data Analytics**

Excel spreadsheets have been around for more than 30 years and they're still valuable. The original concept isn't much different than what we use today, it just looks better and has a lot of new capabilities. But aren't Excel spreadsheets outdated? It's manual and there are better software programs.

Spreadsheets are still relevant and a great tool to learn about data. It's true it's not the only or most fitting solution for all data projects, but it remains as a reliable and affordable tool for analytics. It's a foundational structure for intelligent data because it deepens your understanding of the analytics process. Many industries and businesses continue to emphasize the importance of Excel skills because it remains as an intelligent way to extract actionable insights. Revenue patterns, operations, marketing trends, and more can be analyzed through Excel spreadsheets, but the real advantage is the process.

## Keywords

- Bayes' theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability.
- Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.
- In finance, Bayes' theorem can be used to rate the risk of lending money to potential borrowers.
- SPSS stands for "Statistical Package for the Social Sciences". It is an IBM tool. This tool first launched in 1968. This is one software package.
- R Packages: One of the major features of R is it has a wide availability of libraries.

## Summary

- Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS.
- It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA). Excel forms part of the Microsoft Office suite of software
- R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.
- As R programming language is an open source. Thus, you can run R anywhere and at any time.
- R programming language is suitable for GNU/Linux and Windows operating system.
- R programming is cross-platform which runs on any operating system.

## Self Assessment

1. _____is the measure of the likelihood that an event will occur.

A. Probability
B. Statistics
C. Sample space
D. Random Experiment

2. The _____Theorem is a mathematic model, based on statistics and probability that aims to calculate the probability of one scenario based on its relationship with another scenario.

A. Multiplication
B. Addition
C. Bayes

D. Random theorem

3. The initial probability is based on the present level of information.

A. Prior Probability
B. Posterior Probability
C. Previous Probability
D. All of these

4. A _____ measures the probability of an event given that (by assumption, presumption, assertion or evidence) another event has occurred.

A. Conditional probability
B. Posterior Probability
C. Previous Probability
D. All of these

5. _____ a conditional probability, is the probability of observing event A given that B is true.

A. $P(A \mid B)$
B. $P(B \mid A)$
C. $P(AA \mid B)$
D. $P(A \mid BB)$

6. _____ is the probability of observing event B given that A is true.

A. $P(A \mid B)$
B. $P(B \mid A)$
C. $P(AA \mid B)$
D. $P(A \mid BB$

7. _____ is the study of a person or agents' choices.

A. Decision theory
B. Regression theory
C. Correlation theory
D. None of these

8. _____ is the summation of all the numbers in a dataset divided by the total number of values.

A. Mean
B. Median
C. Mode
D. None of these

9. When the data-set has numbers that are too far away from each other, we use the _____ to find a middle point.

A. Mean

B. Median

C. Mode

D. None of these


10. _____-is the most frequently occurring value in a set of observations.

A. Mean

B. Median

C. Mode

D. None of these


11. _____-is a nice way to identify normal variation and abnormal variation in Task.

A. Process Control Chart

B. Median

C. Mean

D. Variance


12. One of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions is.

A. SPSS

B. DMiner

C. MMiner

D. None of these


13. _____is an integrated development environment (IDE) for R.

A. RStudio

B. RS studio

C. S studio

D. None of these


14. _____involves organizing and summarizing the data for better and easier understanding by describing the data.

A. Descriptive statistics

B. Inferential statistics

C. Regression analysis

D. Confidence level


15. _____is the method of estimating the population parameter based on the sample information. It applies dimensions from sample groups in an experiment to contrast the conduct group and make overviews on the large population sample.

A. Descriptive statistics

B. Inferential statistics

C. Regression analysis

D.  Confidence level

## Answers for Self Assessment

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. | A | 2. | C | 3. | A | 4. | A | 5. | A |
| 6. | B | 7. | A | 8. | A | 9. | B | 10. | C |
| 11. | A | 12. | A | 13. | A | 14. | A | 15. | B |

## Review Questions

1)  What is Bayes theorem in simple terms?
2)  What is Bayes Theorem example?
3)  What is the difference between conditional probability and Bayes Theorem?
4)  How is Bayes theorem used in real life?
5)  What is SPSS and its advantages?
6)  What are the major features of SPSS?
7)  What is RStudio used for?
8)  Explain any five data analytics functions in Excel?
9)  How Bayes rule cracked the Enigma code?

## Further Readings

- An Introduction to Probability and Statistics

  Book by A. K. Md. Ehsanes Salah and V. K. Rohatgi

- First Course in Probability, A

  Book by Sheldon M. Ross

- Schaums Theory and Problems of Statistics

  Book by Murray R. Spiegel

- Introduction to Probability, Statistics, and Random ...

  Book by HosseinPishro-Nik

## Web Links

https://www.tutorialspoint.com

www.webopedia.com

https://www.britannica.com/science/probability