MODULE 4

Probability Theory

You will learn about the ‘Probability Theory’ in this module.

Module Learning Objectives

At the end of this module, you will be able to:

- Understand the principles of counting including addition rule, product rule, permutation and combination.
- Understand the introduction and important definitions of Probability Theory.
- Describe conditional probability.
- Explain Bayes Theorem and its applications.
- Explain different methods of discrete probability distribution like discrete uniform distribution, Poisson distribution, Bernoulli distribution and Binomial distribution.
- Understand the concepts of Covariance and Correlation.
- Understand the different methods of Continuous Probability Distribution.
- Explain the concept of Central Limit Theorem.
- Understand Hypothesis Testing and the types of errors.

Module Topics

The following topics that will be covered in the module:

1. Principles of Counting
2. Introduction and Definitions of Probability Theory
3. Conditional Probability

- **Multiplication rule:**

  The multiplication rule states that 'If E1 is an experiment with n1 outcomes and E2 is an experiment with n2 possible outcomes, and E1 and E2 are independent events, then the experiment which consists of performing E1 first and then E2 consists of **n1*n2** possible outcomes.

  For example, tossing a coin twice, will produce the following number of outcomes:

  For E1, n1 = 2; for E2, n2 = 2; n1*n2 = 4

- **Permutation:**

  The number of ways to choose a sample of r elements from a set of n distinct objects where order does matter and replacements are not allowed. When n = r this reduces to n!, a simple factorial of n. Permutation is computed using the formula, nPr = n! divided by (n-r)!, i.e., **n!/(n-r)!**

- **Combination:**

  The number of ways to choose a sample of r elements from a set of n distinct objects where order doesn't matter and replacements are not allowed. Combination is computed using the formula, nCr = n! divided by r!*(n-r)!, i.e., **n!/r!*(n-r)!**

1.1 Permutations

Permutation: An arrangement or order of a set of objects, and the order if arrangement is important. There are four theorems of Permutation.

- **Theorem 1:** Arranging n Objects: This theorem states that n objects can be arranged in n! Ways.
- **Theorem 2:** Number of Permutations: The number of permutations of n distinct objects taken 'r' at a time, denoted by nPr, where repetitions are not allowed, is given by:

nPr = $n(n-1)(n-2)\ldots(n-r+1) = \dfrac{n!}{(n-r)!}$

Notation of permutations can be: P_r^n or ${}^nP_r$

Recall the factorial notations that you have learned in school mathematics. n!, which is read as n-factorial, is defined as:

n! = n x (n-1) x (n-2) x ….. 3 x 2 x 1.

The theorems in Permutation are based on factorial notation.

Example for theorem 1: In how many ways can 3 persons be seated?

Since there are 3 persons, the number of ways is 3!.

From the factorial notation, 3! = 3 x 2 x 1 = **6 ways**

Theorem 2: Note that nPn = n!, since 0! = 1.

Example for theorem 2: In how many ways can 5 boxes be arranged in 3 spaces in a shelf?

By theorem 2, 5P3 = 5!/(5-3)! = 5!/2! = **60**

Theorem 3: Permutations of different kinds of objects: The number of different permutations of n objects of which n1 are of one kind, n2 are of a second kind, ... nk are of a k-th kind is:

$$\frac{n!}{n_1! \times n_2! \times n_3 \times \ldots \times n_k!}$$

Theorem 4: Arranging objects in a circle: There are (n−1)! ways to arrange n distinct objects in a circle (where the clockwise and anti-clockwise arrangements are regarded as distinct.)

Example for theorem 3: In how many ways can the seven letters of the word "arrange" be arranged in a row?

There are two 'a's, two 'r's, one n, one g and one e.

Based on theorem 3, the number of ways of arrangement is: 7!/(2!)(2!)(1!)(1!)(1!) = 5040/(2)(2)(1)(1)(1) = 5040/4 = **1260 ways**.

Example for theorem 4:

In how many ways can 5 people be arranged in a circle?

Based on theorem, (5-1)! = 4! = **24 ways**.

1.2 Combinations

Combinations:

- Combinations are for unordered selections.
- A combination of n objects taken r at a time is a selection which does not take the arrangement of the objects into account, i.e the order of arrangement is not important.
- The number of ways (or combinations) in which r objects can be selected from a set of n objects, where repetition is not allowed, is denoted by:

$$C_r^n = \frac{n!}{r!(n-r)!} \quad \text{Or} \quad C_r^n = \frac{P_r^n}{r!}$$

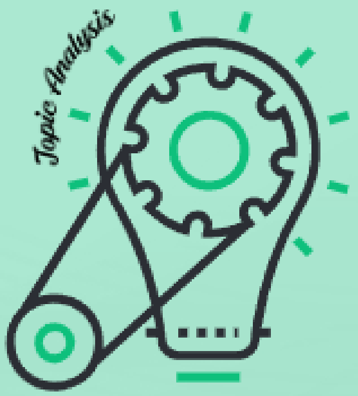
The ways of combination can be computed even if the objects are unordered. Some points to note are:

- nC0 = 1
- nCn = 1
- nCr = nCn-r

Example: Find the number of ways in which 3 components can be selected from a batch of 20 different components.

20C3 = 20!/(3!)(20-3)! = 20!/(3!)(17!) = **1140 ways**.

What did You Grasp?



1. In how many ways can a group of 4 boys be selected from 10 if the eldest boy is excluded?

   A) 84
   B) 126
   C) 210
   D) 44

2. Introduction to Probability Theory

Chance, Likely, Possibly...etc,

What do these words indicate?

These words indicate 'Lack of certainty'.

- There is no perfect yardstick to measure the lack of certainty or simply 'uncertainty'.
- But, it can be measured mathematically on the basis of some assumptions. This numerical measure is referred to as "PROBABILITY".

Will it rain today? It may.

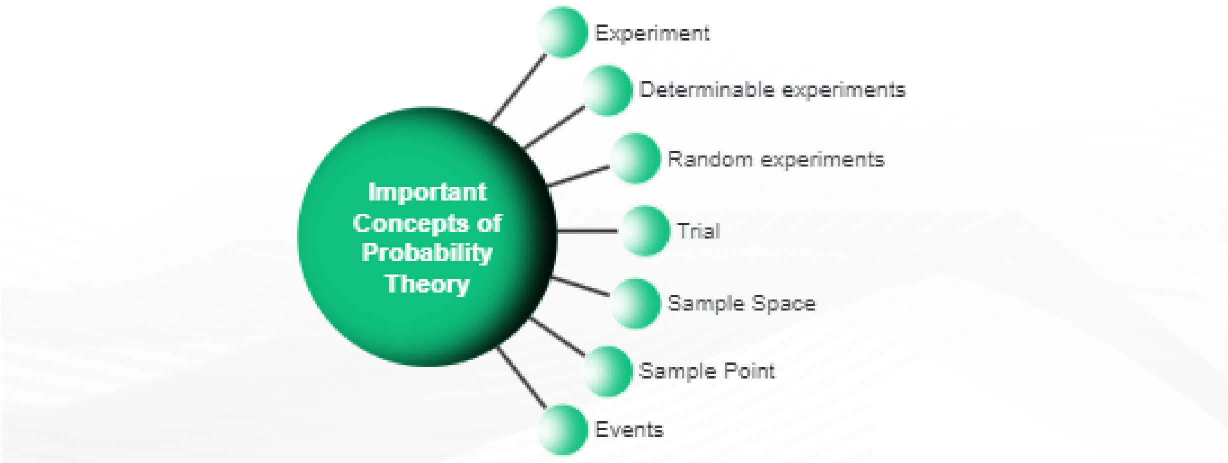
Will India win the match? There's a chance.

We come across the words may be, possibly, likely etc. on a daily basis. Things may happen or may not. There is always an uncertainty associated with these terms. Is there a way to measure uncertainty? There is no perfect way to measure uncertainty, but mathematically, based on assumptions, uncertainty can be measured. Here comes the probability theory. You can use the terms chance, odds, uncertainty, prevalence, risk, expectancy etc., to interpret probability.

In 1929, Bertrand Russell said during one of his lectures, ‘ Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means.” Probability theory is used to evaluate the reliability of conclusions and inferences based on data, thus it is fundamental to statistics.

2.1 Key Concepts and Definitions of Probability Theory

The important concepts of probability theory:



Important concepts of probability are listed below:

- **Experiment:** There are two categories of experiments
    - **Determinable experiments** - Experiments whose outcome may be predictable. For example, if we heat water, we know for sure that it will boil and change into vapour at some point in time.
    - **Random experiments** - Experiments whose outcomes are not predictable, i.e., a non-deterministic model. If in each trial of an experiment conducted under identical conditions, the outcome is not only always the same but may be any of the possible outcomes, such experiment is termed as a Random experiment. Examples of such events are - tossing a coin, rolling a die, selecting a card from a well-shuffled deck of 52 cards, etc.
- **Trial:** A trial is an action which results in one or several outcomes. Example: Tossing a coin, rolling a die.
- **Sample Space:** The set of all outcomes of a random experiment is called sample space. Denoted by ‘S”.

  For rolling a dice, the sample space is, S = {1, 2, 3, 4, 5, 6}

  For tossing a coin, S = {H, T}

- **Sample Point:** Sample point is defined as each individual outcome of a random experiment.

  Flipping a coin: {H}, {T}

  Rolling a dice: {1}, {2}, {3}, {4}, {5}, {6}

- **Events:** An event is defined as any subset of the sample space. For example, the events in rolling a die can include: getting an even number, getting a prime number, etc.

2.2 Independent Events

Independent Events:

- If two events A and B are said to be "Independent" P(AB) = P(A)*P(B)
- Note 1: If A and B are independent, then,
  - A' and B' are independent.
  - A and B' are independent.
  - A' and B are independent.
- Here, A' denotes A complement, B' denotes B complement.
- *Note 2: Mutually Exclusive events are certainly not independent.*

The interpretation of two events being independent is that occurrence of one event does not affect the probability of the other event, i.e., P(A|B) = P(A). P(B|A) means "the probability of the event B given that A has already occurred" (you will learn about this in conditional probability).

Intersections of Independent Events

The probability of the intersection of a series of independent events A1, A2, A3,..An is simply given by

$P(A1 \cap A2 \ldots\ldots An) = P(A1)\, P(A2) \ldots\ldots P(An)$.

Example:

Problem: A fair die is tossed twice. Find the probability of getting a 4 or 5 on the first toss and a 1, 2, or 3 in the second toss.

Solution:

$P(A1) = P(4 \cup 5) = 2/6 = \frac{1}{3}$

$P(A2) = P(1 \cup 2 \cup 3) = 3/6 = \frac{1}{2}$

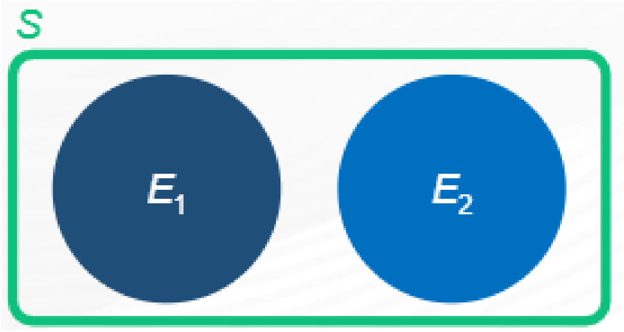
Since these are independent events,

$P(A1 \cap A2) = P(A1)\, P(A2) = (\frac{1}{3})(\frac{1}{2}) = \frac{1}{6}$.

2.3 Mutually Exclusive Events

Mutually exclusive (disjoint):

- Events are the events that cannot happen simultaneously.
- Two events A and B are said to be mutually exclusive if they do not contain any element in common. So the probability of the two events will be zero.
- P(A and B) = 0, i.e., $P(A \cap B) = 0$.
- When rolling a die, if A = getting an even number and B = getting an odd number, then A and B are said to be mutually exclusive events, as we cannot get both at the same time.



Two or more events are said to be mutually exclusive if the occurrence of any one of them means the others will not occur (That is, we cannot have 2 or more such events occurring at the same time). In the diagram, we see that there is no overlap in two mutually exclusive events. Hence, the probability of two mutually exclusive events is zero.

Now, suppose "A or B" denotes the event that "either A or B both occur", then

(a) If A and B are not mutually exclusive events: P(A or B) = P(A) + P(B) − P(A and B)

We can also write: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Complements of Events

The event A' (pronounce as A 'dash'), the complement of event A, is the event consisting of everything in the sample space S that is not contained within the event A.

2.4 Collectively Exhaustive Events

- In a sample space S, if events A and B are said to be Collectively Exhaustive events, then

  $A \cup B$ =S (sample space)

  $P(A \cup B) = 1$

  Eg: When rolling a dice, if

- A = Getting an odd number, and

- B = getting an even number
- A and B are Collectively Exhaustive Events.

In probability, exhaustive is a condition of two or more events which serves a great role in finding the probability as it changes if the events are exhaustive or not. A set of events are said to be exhaustive if there is a certain chance of occurrence of at least one of them when they are all considered together.

For example: Consider the experiment of a fair die being thrown. Then there are six outcomes and all of them are equally likely to occur. Also ,the events of getting different numbers taken together are exhaustive as together at least one of them is certain to happen. For getting a 2 or 5, sure will get one of the numbers during the experiment. So events are exhaustive.

If we compare this with mutually exhaustive events, where no more than one event can occur at a given time, the set of all possible die rolls is both collectively exhaustive and mutually exclusive. The events 1 and 6 are mutually exclusive but not collectively exhaustive. The events “even” (2,4 or 6) and “not-6” (1,2,3,4, or 5) are collectively exhaustive but not mutually exclusive.

What did You Grasp?

1. State True or False.
   A sample space refers to each individual outcome of a random experiment.
   A) True
   B) False

2. If the probability that person A will be alive in 20 years is 0.7 and the probability that person B will be alive in 20 years is 0.5, what is the probability that they will both be alive in 20 years?
   A) 0.5
   B) 0.45
   C) 0.35
   D) 0

3. It is known that the probability of obtaining zero defectives in a sample of 40 items is 0.34 whilst the probability of obtaining 1 defective item in the sample is 0.46. What is the probability of obtaining not more than 1 defective item in a sample?
   A) 1
   B) 0.2
   C) 0.8
   D) 0.35

2.5 Conditional Probability

- Conditional Probability of an event A, given that event B, has already happened is given by: P(A|B).
- If A and B are two events in a sample space S, given that P(B) > 0, then the conditional probability of A, assuming B has already happened, is given by:

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

- If an event is taken as a sample space for another event which is a subset of the former event.

The formula for conditional probability is derived based on the following logic. When we know that event B has occurred, every outcome that is outside B should be discarded. Thus, our sample space is reduced to the set B. Now the only way that A can happen is when the outcome belongs to the set A ∩ B. We divide P(A ∩ B) by P(B), so that the conditional probability of the new sample space becomes 1, i.e., P(B|B)=P(B∩B) divided by P(B) = 1.

Example:

A pair of four-sided dice is rolled and the sum is determined. What is the probability that a sum of 3 is rolled before a sum of 5 is rolled in a sequence of rolls of the dice?

Solution:

The sample space of this random experiment is

S = {(1, 1) (1, 2) (1, 3) (1, 4)

(2, 1) (2, 2) (2, 3) (2, 4)

(3, 1) (3, 2) (3, 3) (3, 4)

(4, 1) (4, 2) (4, 3) (4, 4)}.

Let A be the event of getting a sum of 3 and B be the event of getting a sum of 5. The probability that a sum of 3 is rolled before a sum of 5 is rolled can be thought of as the conditional probability of a sum of 3, given that a sum of 3 or 5 has occurred. That is, P(A/A U B). Hence,

P(A/A U B) = P(A ∩ (A U B))/P(A U B) = P(A)/P(A) + P(B) = 2/2 + 4 = ⅓ .

What did You Grasp?

1. What is the probability that the total of two dice will be greater than 8, given that the first die is a 6?

   A) ⅓
   B) 1/9
   C) ⅙
   D) ⅔

2.6 Law of Total Probability

If A_1, A_2, A_3, …. An is a partition of a sample space, (they are mutually exclusive and totally exhaustive events) then the probability of any event B can be obtained from the probabilities $P(A_i)$ and $P(B|A_i)$ using the formula (based on conditional probability):

i.e.
$$P(B) = P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)$$

$$P(B) = \sum_i P\ (B|Ai)P(Ai)$$

Let's see an example of Total Probability.

I have three bags that each contain 100 marbles.

Bag 1 has 75 red and 25 blue marbles.

Bag 2 has 60 red and 40 blue marbles.

Bag 3 has 45 red and 55 blue marbles.

I choose one of the bags at random and then pick a marble from the chosen bag, also at random. What is the probability that the chosen marble is red?

Solution:

Let R be the event that the chosen marble is red. Let Bi be the event that I choose Bag i. We already know that

$P(R|B_1)=0.75$,

$P(R|B_2)=0.60$,

$P(R|B_3)=0.45$

We choose our partition as B_1, B_2, B_3. Note that this is a valid partition because, firstly, the Bi's are disjoint (only one of them can happen), and secondly, because their union is the entire sample space as one the bags will be chosen for sure, i.e., $P(B_1 \cup B_2 \cap B_3) = 1$.

Using the law of total probability, we can write

$P(R) = P(R|B_1)P(B_1)+P(R|B_2)P(B_2)+P(R|B_3)P(B_3)$

$=(0.75)⅓ + (0.60) ⅓ + (0.45)⅓$

$=0.60$

2.7 Bayes' Theorem

- If A_1, A_2, A_3, …. An is a partition of a sample space, (they are mutually exclusive and totally exhaustive events) then the posterior probabilities of the event Ai, conditional on any event B can be obtained from the probabilities, P(Ai) > 0 and P(B) > 0.
- Suppose that we know P(A|B), but the probability P(B|A) is unknown. Using the definition of conditional probability, we have:
- $P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$
- Dividing by P(A), we get the Bayes' rule, that is :

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

There are many situations where the ultimate outcome of an experiment depends on what happens in various intermediate stages. This issue is resolved by the Bayes' Theorem.

2.7.1 Bayes' Rule

The following are two Bayes' rules:

- **Rule 1:** For any two events A and B, where $P(A) \neq 0$,

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

- **Rule 2:** If B_1, B_2, B_3 …. Bn, form a partition of the sample space S, and A is any event with $P(A) \neq 0$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Bayes' Theorem (also known as Bayes' rule) is a deceptively simple formula used to calculate conditional probability. For two events, A and B, Bayes' theorem allows you to figure out P(A|B) (the probability that event A happened, given that test B was positive) from p(B|A) (the probability that test B happened, given that event A happened).

Example:

Imagine you want to find out a patient's probability of having liver disease if they are an alcoholic.

- **A** could mean the event "Patient has liver disease." Past data tells that 10% of patients entering your clinic have liver disease. P(A) = 0.10.
- B could mean the litmus test that "Patient is an alcoholic." 5% of the clinic's patients are alcoholics. P(B) = 0.05.

You might also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your B|A: the probability that a patient is alcoholic, given that they have liver disease, is 7%.

By Bayes' theorem:

P(A|B) = (0.07 * 0.1)/0.05 = 0.14

In other words, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%). This is a large increase from the 10% suggested by past data. But it's still unlikely that any particular patient has liver disease.

2.8 Random Variables

- A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon.
- Though it is named as a variable, it is a function from Sample space S to R (set of real numbers).

$$X: S \to R$$

- A random variable is a real-valued variable whose value is determined by an underlying random experiment.
- There are two types of random variables, discrete and continuous.

Objects in a sample space may not be numbers. Thus, we use the notion of a random variable to quantify the qualitative elements of the sample space.

To analyze random experiments, we usually focus on some numerical aspects of the experiment. For example, in a soccer game, we may be interested in the number of goals, shots, shots on goal, corners kicks, fouls, etc. If we consider an entire soccer match as a random experiment, then each of these numerical results gives some information about the outcome of the random experiment. These are examples of random variables. In a nutshell, a random variable is a real-valued variable whose value is determined by an underlying random experiment.

For example, if we toss a coin 5 times, the random experiment and the sample space can be written as: S = {TTTTT, TTTTH, TTTHH, ….. HHHHH}.

The sample space S here has $2^5 = 32$ elements. Suppose that in this experiment, we are interested in the number of heads. We can define a random variable X whose value is the number of observed heads. The value of X will be one of 0, 1, 2, 3, 4 or 5 depending on the outcome of the random experiment.

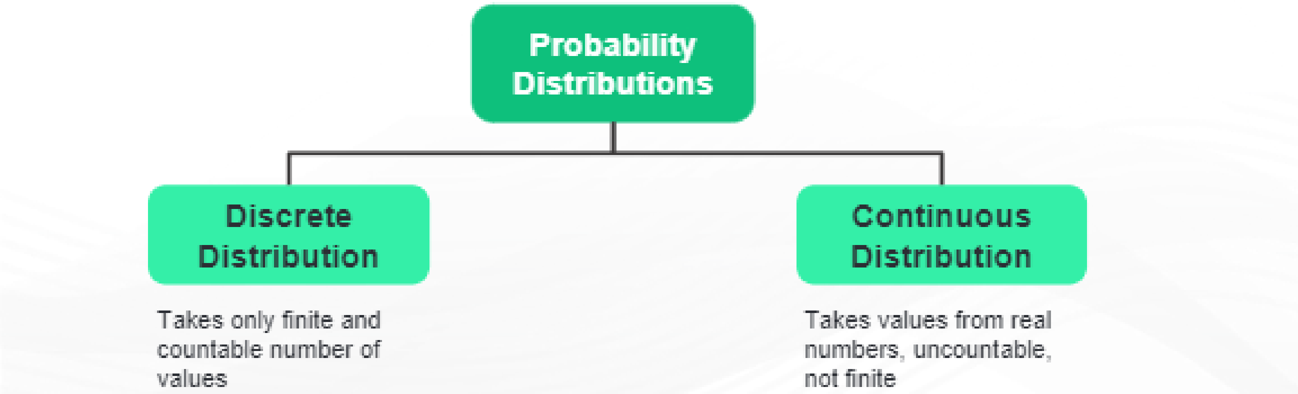
Since a random variable is a function, we can talk about its range. The range of a random variable X, shown by Range(X) or R_X, is the set of possible values for X. In the above example, Range(X) = R_X = {0, 1, 2, 3, 4, 5}.

A random variable is characterized by either its probability density function or its cumulative distribution function. The other characteristics of a random variable are its mean, variance and moment generating function.

There are two types of random variables: discrete random variables and continuous random variables, and based on this there are two types of probability distribution functions.

3. Probability Distributions

There are two major categories of probability distributions.



A discrete random variable is one whose set of assumed values is countable (arises from counting). A discrete probability distribution is a table (or a formula) listing all possible values that a discrete variable can take on, together with the associated probabilities.

A continuous random variable is one whose set of assumed values is uncountable (arises from a measurement.).

3.1 Discrete Probability Distribution

- A discrete distribution describes the probability of occurrence of each value of a discrete random variable.
- A discrete random variable is a random variable that has countable values.
- When we say that the probability distribution of an experiment is discrete then the sum of probabilities of all possible values of the random variable must be equal to 1. That is, if X is a discrete random variable, then,

$$\sum P(Xi)=1$$

(Here, 'i = 1, 2, 3, ……. n' is the set of all values that the variable X can take.)

A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,........ Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete.

Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten, etc.

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

Some of the discrete probability distributions are Poisson distribution, Bernoulli distribution, Binomial distribution, etc. We'll learn about some of these distributions in detail in the upcoming sections.

3.1.1 Important Formulae

The following are two important formulae of Discrete probability distribution:

Mean $$\begin{aligned}\mu_x &= x_1p_1 + x_2p_2 + \cdots + x_kp_k \\ &= \sum x_ip_i\end{aligned}$$

Variance $$\sigma_x^2 = \sum (x_i - \mu_x)^2 p_i$$

The standard deviation σX is the square root of the variance.

Suppose a variable X can take the values 1, 2, 3, or 4. The probabilities associated with each outcome are described by the following table:

For the outcomes 1, 2, 3, 4, the probabilities are 0.1, 0.3, 0.4 and 0.2, respectively.

The probability that X is equal to 2 or 3 is the sum of the two probabilities: P(X = 2 or X = 3) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7. Similarly, the probability that X is greater than 1 is equal to 1 - P(X = 1) = 1 - 0.1 = 0.9, by the complement rule.

3.1.2 Discrete Uniform Distribution

- A random variable has a uniform distribution when each value of the random variable is equally likely, and values are uniformly distributed throughout some interval.
- A discrete uniform distribution is a symmetric probability distribution whereby a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n.

- The cumulative distribution function (CDF) of the discrete uniform distribution can be expressed, for any k ∊ [a,b], as:

$$F(k;a,b)=\frac{\lfloor k\rfloor-a+1}{b-a+1}$$

An example of a discrete uniform distribution on the first N integers is the statistical experiment of rolling one die, where the random variable X represents the outcome of the die.

For the standard six-sided die, we have the probability P(X) = ⅙ for each outcome. Furthermore, the expected value is E(X) = 6+1 / 2=3.5, so over the long run, the average of the outcomes should be midway between 3 and 4. We also find that the variance is Var(X) = 6^2−1 / 12 = 35 / 12 ≈ 2.9167, and the standard deviation of the outcomes is σX = sqrt(35 / 12) ≈ 1.7078.

3.1.3 Poisson Distribution

- Developed by the French mathematician Simeon Denis Poisson in 1837.
- Probability distribution of a Poisson random variable X representing the number of successes occurring in a given time interval or a specified region of space is given by the formula:

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

(x = 0, 1, 2, 3… n

e = 2.71828 (but use your calculator's 'e' button)

μ = mean number of successes in the given time interval or region of space)

Poisson Distribution was developed by the French mathematician Simeon Denis Poisson in 1837.

A Poisson random variable must satisfy the following conditions:

- The number of successes in two disjoint time intervals is independent.
- The probability of a success during a small time interval is proportional to the entire length of the time interval.

The formula for calculating the probability distribution of a Poisson random variable is given in the slide.

Applications of Poisson distribution include:

- The number of deaths by horse kicking in the Prussian army (the first application)
- Birth defects and genetic mutations
- Rare diseases (like Leukemia, but not AIDS because it is infectious and so not independent) - especially in legal cases

- Car accidents
- Traffic flow and an ideal gap distance
- Number of typing errors on a page
- Hair found in McDonald's hamburgers
- Spread of an endangered animal in Africa
- Failure of a machine in one month

3.1.3.1 Example of Poisson Distribution

Problem:

If electricity power failures occur according to a Poisson distribution with an average of 3 failures every 20 weeks, calculate the probability that there will not be more than one failure during a particular week.

Solution:

The average number of failures per week is: μ = 3/20 = 0.15

$$P(x0) + P(x1) = \frac{e^{-0.15}\,0.15^0}{0!} + \frac{e^{-0.15}\,0.15^1}{1!}$$

= 0.98981

In this example, we've added the probabilities for 0 failures and 1 failure, since it is given that "Not more than one failure".

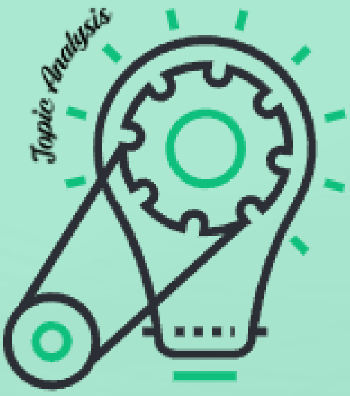
3.1.3.2 Mean and Variance of Poisson Distribution

If μ is the average number of successes occurring in a given time interval or region in the Poisson distribution, then the mean and the variance of the Poisson distribution are both equal to μ.

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

In a Poisson distribution, only one parameter, μ is needed to determine the probability of an event.

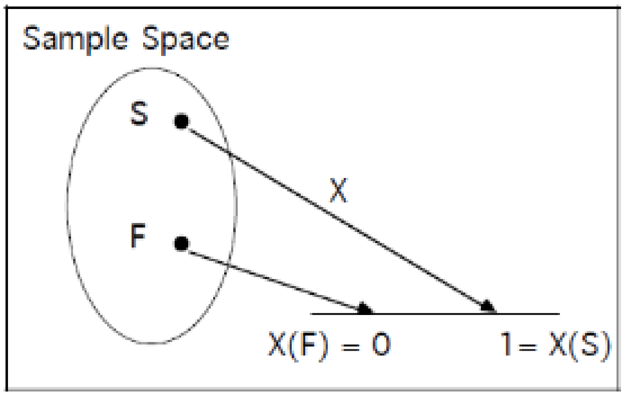
What did You Grasp?



1. A life insurance salesman sells on an average 3 life insurance policies per week. Use Poisson's law to calculate the probability that in a given week he will sell 2 or more policies, but less than 5 policies.

   A) 0.61611
   B) 0.32926
   C) 0.62358
   D) 0.52369

3.1.4 Bernoulli Distribution



Bernoulli Distribution:

The probability density function of a random variable is

$f(0) = P(X = 0) = 1 - p$

$f(1) = P(X = 1) = p,$

(where p denotes the probability of success. Hence,

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

The random variable X is called the Bernoulli random variable if its probability density function is of the form

$f(x) = p^x (1 - p)^{1-x}, x = 0, 1,$ where p is the probability of success.

We denote the Bernoulli random variable by writing $X \sim BER(p)$.

Example:

What is the probability of getting a score of not less than 5 in a throw of a six-sided die?

Solution:

Although there are six possible scores {1, 2, 3, 4, 5, 6}, we are grouping them into two sets, namely {1, 2, 3, 4} and {5, 6}. Any score in

{1, 2, 3, 4} is a failure and any score in {5, 6} is a success. Thus, this is a Bernoulli trial with:

P(X = 0) = P(failure) = 4/6

P(X = 1) = P(success) = 2/6.

Hence, the probability of getting a score of not less than 5 in a throw of a six-sided die is 2/6.

3.1.4.1 Bernoulli's Theorem

Bernoulli Theorem:

If X is a Bernoulli random variable with parameter p, then the mean, variance and moment generating functions are respectively given by:

$$\mu_X = p$$
$$\sigma_X^2 = p\,(1-p)$$
$$M_X(t) = (1-p) + p\,e^t.$$

Proof for mean, standard deviation and moment generating function:

$$\mu_X = \sum_{x=0}^{1} x\,f(x) = \sum_{x=0}^{1} x\,p^x\,(1-p)^{1-x} = p.$$

$$\sigma_X^2 = \sum_{x=0}^{1}(x-\mu_X)^2\,f(x) = \sum_{x=0}^{1}(x-p)^2\,p^x\,(1-p)^{1-x} = p^2\,(1-p) + p\,(1-p)^2 = p\,(1-p)\,[p+(1-p)] = p\,(1-p)$$

$$M(t) = E\left(e^{tX}\right) = \sum_{x=0}^{1} e^{tx}\,p^x\,(1-p)^{1-x} = (1-p) + e^t\,p.$$

Understand Bernoulli's theorem and its proof. Note that for the Bernoulli distribution all its moments about zero are same and equal to p.

3.1.5 Binomial Distribution

A random variable X is said to follow Binomial distribution if its probability mass function is given by:

$$P(X) = C_x^n p^x q^{n-x}$$

where,

n = the number of trials; x = 0, 1, 2, ... n; p = the probability of success in a single trial; q = the probability of failure in a single trial (i.e. q = 1 − p);

C_x^n is a combination.

Binomial Distribution applies to when chances are two alternatives.ie, p = success and q = failure and p + q = 1.

Example:

The probability of winning a match for team A is 0.60. Find the probability of winning 3 matches out of 5.

Solution:

Probability of winning, p = 0.60

Probability of losing, q = 0.40

Probability of winning 3 matches out of 5, $P(x=3)5 = 5C3(0.6)^3(0.4)^2$

$P(x=3) = 5C3 \times 0.216 \times 0.16 = 10 * 0.03456$

Hence, the probability is **0.3456**.

3.1.5.1 Properties of a Binomial Experiment

A Binomial experiment has the following properties:

- The experiment consists of n repeated trials.
- Each trial results in an outcome that may be classified as a success or a failure (hence the name, binomial).
- The probability of a success, denoted by p, remains constant from trial to trial and repeated trials are independent.
- The number of successes X in n trials of a binomial experiment is called a binomial random variable.

Go through the pointers in the slide and understand the properties of a Binomial distribution. The Binomial distribution can arise whenever we select a random sample of n units with replacement. Each unit in the population is classified into one of two categories according to whether it does or does not possess a certain property.

3.1.5.2 Constants of Binomial Distribution

If p is the probability of success and q is the probability of failure in a binomial trial, then the expected number of successes in n trials (i.e. the mean value of the binomial distribution) is:

$E(X) = \mu = np$

The variance of the binomial distribution is:

$V(X) = \sigma 2 = npq$

Standard deviation = $\sigma = \sqrt{npq}$

In a binomial distribution, only 2 parameters, namely n, and p, are needed to determine the probability.

What did You Grasp?

1. A die is tossed 5 times. What is the probability of no fives turning up?

   A) 0.5787
   B) 0.2356
   C) 0.3472
   D) 0.2569

3.2 Covariance

Covariance:

Provides a measure of the strength of the correlation between two or more sets of random variates.

Is the relationship between two variables in a given data set.

Example: Let E(x) be the expected value of a given variable x, and E(y) be the expected value of variable y, then the covariance between x and y is given by:

$$\text{cov}(x, y) = E[xy] - E[x]\,E[y]$$

The covariance generalizes the concept of variance to multiple random variables. Instead of measuring the fluctuation of a single random variable, the covariance measures the fluctuation of two variables with each other.

3.2.1 Properties of Covariance

Given a constant 'a' and random variables X, Y, and Z, the following properties are:

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X) \geq 0$$
$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$
$$\mathrm{Cov}(aX,Y) = a\mathrm{Cov}(X,Y)$$
$$\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z).$$

Go through the slide and understand the properties of covariance.

3.3 Correlation

Correlation:

- Depends upon two variables, change in one variable effects a change in second variable.
- Its value lies in the range of -1 and +1. Whereas in covariance two variables vary together, which can be negative or positive.
- Is the measure of the strength of the relation between two variables. How strongly two variables are connected is defined as the correlation.
- Is related to covariance by the given formula: cor(x,y)= cov(x,y)/σx σy

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it's a measure of how things are related. The study of how variables are correlated is called correlation analysis.

Some examples of data that have a high correlation:

Your caloric intake and your weight.

Your eye color and your relatives' eye colors.

The amount of time your study and your GPA.

Some examples of data that have a low correlation (or none at all):

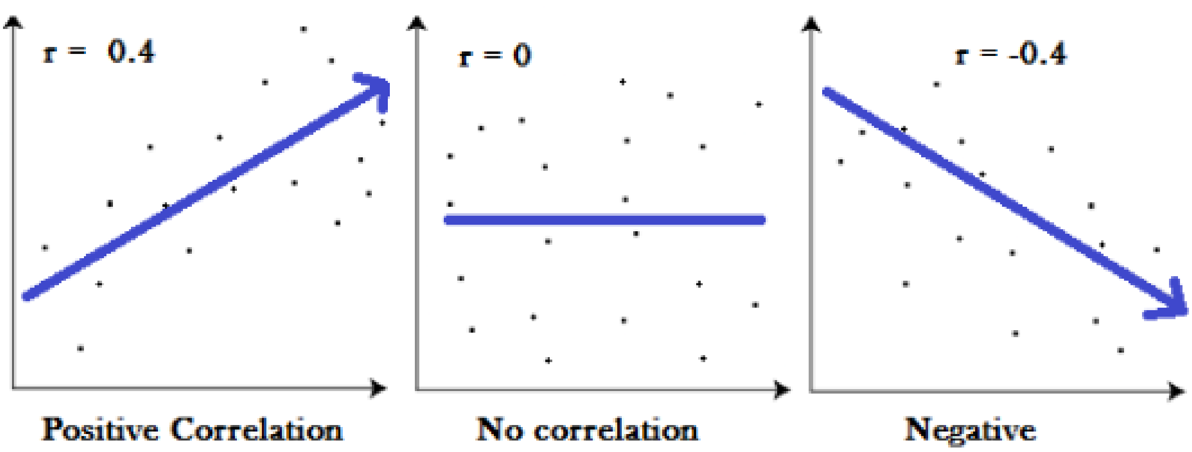
A dog's name and the type of dog biscuit they prefer.

The cost of a car wash and how long it takes to buy a soda inside the station.

3.3.1 Correlation Coefficient

A correlation coefficient is a way to put a value to the relationship.

The following graph shows the correlation of -1, 0 and 1.



Correlation coefficients have a value of between -1 and 1. A "0" means there is no relationship between the variables at all, while -1 or 1 means that there is a perfect negative or positive correlation (negative or positive correlation here refers to the type of graph the relationship will produce).

4. Continuous Random Variables

Continuous Random Variable:

- A variable, which takes an infinite number of possible values.
- Are usually measurements.
- Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.
- Is not defined at specific values. Instead, it is defined over an interval of values, and is represented by the area under a curve (in advanced mathematics, this is known as an integral).
- The probability of observing any single value is equal to 0, since the number of values, which may be assumed by the random variable is infinite.

Suppose a random variable X may take all values over an interval of real numbers. Then the probability that X is in the set of outcomes A, P(A), is defined to be the area above A and under a curve. The curve, which represents a function p(x), must satisfy the following:

1: The curve has no negative values ($p(x) > 0$ for all x)

2: The total area under the curve is equal to 1.

A curve meeting these requirements is known as a Density curve.

4.1 Continuous Probability Distribution

The probability distribution of a continuous random variable.

The equation that describes the continuous probability distribution is called probability density function.

The criteria for a probability density function are as follows:

- The random variable Y is a function of X; that is, $y = f(x)$.
- The value of y is greater than or equal to zero for all values of x.
- The total area under the curve of the function is equal to one.

A continuous probability distribution differs from a discrete distribution in a few ways.

The probability distribution of a continuous random variable is called continuous probability distribution. The probability density function is the equation that describes the continuous probability distribution.

The conditions that the probability density function should follow, are given above. A continuous probability distribution differs from a discrete probability distribution in a few ways.

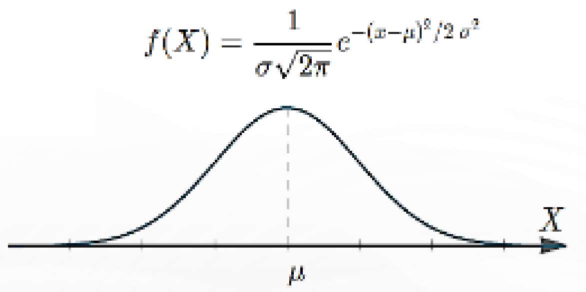
- The probability that a continuous random variable will assume a particular value is zero.
- Therefore, a continuous probability distribution cannot be expressed in tabular form.
- An equation or formula is used to describe a continuous probability distribution, which is the probability density function.

Some of the examples for continuous probability distribution are: normal distribution, exponential distribution, t-distribution, chi-square distribution, and so on. We'll see in detail, some of these distributions in the forthcoming sections.

4.1.1 Normal Distribution

- The Normal Probability Distribution is very common in the field of statistics.
- The random variable X is said to be normally distributed with mean μ and standard deviation σ if its probability distribution is given by

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\,\sigma^2}$$



A normal curve

(where −∞ < μ < ∞ and 0 < σ² < 1 are arbitrary parameters. If X has a normal distribution with parameters μ and σ², then we write X ~ N(μ, σ²).)

Among continuous probability distributions, the normal distribution is very well known since it arises in many applications. Normal distribution was discovered by a French mathematician Abraham DeMoivre.

Whenever you measure things like people's height, weight, salary, opinions or votes, the graph of the results is very often a normal curve. The main importance of normal distribution lies on the central limit theorem which says that the sample mean has a normal distribution if the sample size is large.

4.1.1.1 Properties of Normal Distribution

The properties of normal distribution are as follows:

- The normal curve is symmetrical about the mean μ.
- The mean is at the middle and divides the area into halves.

- The total area under the curve is equal to 1.
- It is completely determined by its mean and standard deviation σ (or variance σ2)

In a normal distribution, only 2 parameters are needed, namely μ and σ2.

Go through the slide and understand the properties of normal distribution.

4.1.1.2 Area Under the Normal Curve

- The probability that a variable is within range in a normal distribution is calculated by finding the area under the normal curve.
- The area depends on the values of μ (mean) and σ (standard deviation).
- The z-score table is used to find the area under the normal curve.
- Z-score is the standardized value of observation x from a distribution that has mean μ and standard deviation σ.
- In a z-score table, the left most column means the number of standard deviations above the the mean to 1 decimal place, the top row gives the second decimal place, and the intersection of a row and column gives the probability.

Area under the normal curve helps us in finding the probability of a variable being in the range of a normal distribution. The z-score table is used to find the area under the normal curve. There are a lot of sources to access the z-score table. An example for such source is: http://users.stat.ufl.edu/~athienit/Tables/Ztable.pdf.

As given above, in a z-score table, the left most column means the number of standard deviations above the the mean to 1 decimal place, the top row gives the second decimal place, and the intersection of a row and column gives the probability.

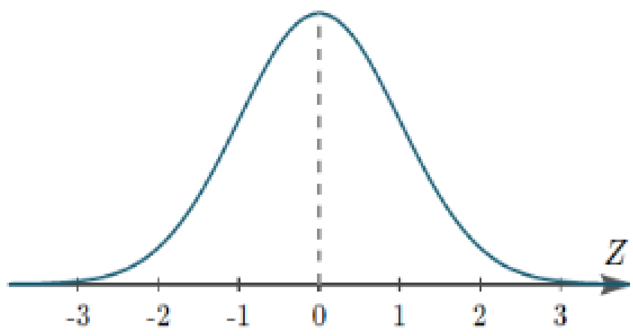
For example, if we want to know the probability that a variable is no more than 0.51 standard deviations above the mean, we need to see the 6th row down in page 2 of the pdf, that starts with 0.0

(corresponding to 0.5) and the 2nd column (corresponding to 0.01). The intersection of the 6th row and 2nd column is 0.6950. This tells us that there is a 69.50% percent chance that a variable is less than 0.51 sigmas above the mean.

4.1.1.3 Standard Normal Distribution

- A random variable X is called a standard normal variate if its mean is zero and its standard deviation is unity, i.e. 1.
- We can transform all the observations of any normal random variable X with mean μ and variance σ to a new set of observations of another normal random variable Z with mean 0 and variance 1 using the following transformation:

$$Z = \frac{X - \mu}{\sigma}$$



Standard Normal Curve $\mu = 0, \sigma = 1$

A normal distribution with mean μ and standard deviation σ can be converted into standard normal distribution performing change of scale and origin. The new distribution of the normal random variable Z with mean 0 and variance 1 (or standard deviation 1) is called a standard normal distribution. Standardizing the distribution like this makes it much easier to calculate probabilities.

4.1.2 Continuous Uniform Distribution

The probability density function and cumulative distribution function for a continuous uniform distribution on the interval [a,b] are:

$$P(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for } x > b \end{cases}$$

$$D(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \le x \le b \\ 1 & \text{for } x > b. \end{cases}$$

These can be written in terms of the Heaviside step function $H(x)$ as

$$P(x) = \frac{H(x-a) - H(x-b)}{b-a}$$

$$D(x) = \frac{(x-a)H(x-a) - (x-b)H(x-b)}{b-a},$$

the latter of which simplifies to the expected $D(x) = (x-a)/(b-a)$ for $a < x < b$.

The continuous distribution is implemented as `UniformDistribution[a, b]`.

A uniform distribution is sometimes also known as a rectangular distribution, is a distribution that has constant probability.

4.1.2.1 Characteristic Function for a Uniform Distribution

The following formula explains the characteristic function:

For a *continuous* uniform distribution, the characteristic function is

$$\phi(t) = \frac{2}{(b-a)\,t} \sin\left[\tfrac{1}{2}(b-a)\,t\right] e^{i(a+b)t/2}.$$

If $a = 0$ and $b = 1$, the characteristic function simplifies to

$$\begin{aligned} \phi(t) &= \frac{2\sin\left(\tfrac{1}{2}t\right)e^{it/2}}{t} \\ &= \frac{i - i\cos t + \sin t}{t}. \end{aligned}$$

The formula for finding the characteristic function for a uniform distribution is given above.

4.1.2.2 Moment Generating Function of a Uniform Distribution

The following formula explains the moment generating function of a uniform distribution:

If we assume t ≠ 0, then:

$$\begin{aligned} M(t) &= E\left(e^{tX}\right) \\ &= \int_a^b e^{tx}\,\frac{1}{b-a}\,dx \\ &= \frac{1}{b-a}\left[\frac{e^{tx}}{t}\right]_a^b \\ &= \frac{e^{tb}-e^{ta}}{t\,(b-a)}. \end{aligned}$$

If t = 0, and M(0) = 1,

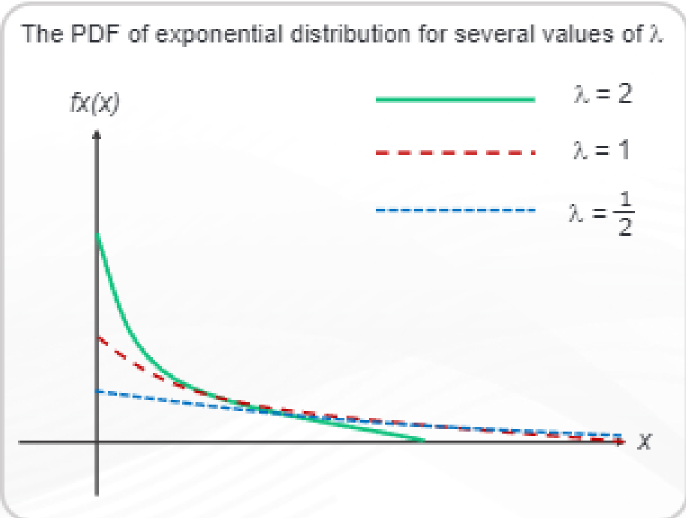
$$M(t) = \begin{cases} 1 & \text{if } t = 0 \\ \frac{e^{tb}-e^{ta}}{t\,(b-a)}, & \text{if } t \neq 0 \end{cases}$$

The formula for calculating the moment generating function for a uniform distribution is given above.

4.1.3 Exponential Distribution

Exponential Distribution:

- Is one of the widely used continuous distributions. It is often used to model the time elapsed between events.
- Is first defined mathematically, and its mean and expected value are derived.
- The intuition for the distribution will then be developed.



An interesting property of the exponential distribution is that it can be viewed as a continuous analogue of the geometric distribution. You toss a coin (repeat a Bernoulli experiment) until you observe the first heads (success). Now, suppose that the coin tosses are Δ seconds apart and in each toss, the probability of success is $p = \Delta\lambda$. Also, suppose that Δ is very small, so the coin tosses are very close together in time and the probability of success in each trial is very low. Let X be the time you observe the first success.

To get some intuition for this interpretation of the exponential distribution, suppose you are waiting for an event to happen. For example, you are at a store and are waiting for the next customer. In each millisecond, the probability that a new customer enters the store is very small. You can imagine that, in each millisecond, a coin (with a very small P(H)) is tossed, and if it lands heads a new customer enters. If you toss a coin every millisecond, the time until a new customer arrives approximately follows an exponential distribution.

4.1.3.1 Formula for Exponential Distribution

A continuous random variable X is said to have an exponential distribution, with parameter $\lambda > 0$, shown as $X \sim Exponential(\lambda)$, its probability density function (PDF) is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Go through the slide and understand the formula for calculating exponential distribution.

4.1.3.2 CDF of an Exponential Distribution

It is convenient to use the unit step function defined as

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

so we can write the PDF of an $Exponential(\lambda)$ random variable as

$$f_X(x) = \lambda e^{-\lambda x} u(x).$$

Let us find its CDF, mean and variance. For $x > 0$, we have

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

So we can express the CDF as

$$F_X(x) = (1 - e^{-\lambda x}) u(x).$$

The steps involved in calculating the CDF of exponential distribution is given in the slide.

4.1.3.3 Variance of Exponential Distribution

Let $X \sim Exponential(\lambda)$. We can find its expected value as follows, using integration by parts:

$$\begin{aligned} EX &= \int_0^\infty x \lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \int_0^\infty y e^{-y} dy \qquad \text{choosing } y = \lambda x \\ &= \frac{1}{\lambda} \Big[- e^{-y} \\ &- y e^{-y} \Big]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Now let's find Var(X). We have

$$\begin{aligned} EX^2 &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy \\ &= \frac{1}{\lambda^2} \Big[-2e^{-y} - 2ye^{-y} - y^2 e^{-y} \Big]_0^\infty \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

The steps involved in deriving the formula for calculating the variance of an exponential distribution is given in the slide.

The formula for calculating the variance of an exponential distribution:

$$\mathrm{Var}(X) = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

$$\boxed{\text{If } X \sim Exponential(\lambda), \text{ then } EX = \frac{1}{\lambda} \text{ and } \mathrm{Var}(X) = \frac{1}{\lambda^2}.}$$

The formula for calculating the variance of an exponential distribution is given in the slide.

4.1.4 p-Value

- When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the null hypothesis.
- The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it.
- All hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling you about the population)

A p-value is used in hypothesis testing to help you support or reject the null hypothesis. The p-value is the evidence against a null hypothesis. The smaller the p-value, the strong the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p-value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That's pretty tiny. On the other hand, a large p-value of 0.9 (90%) means your results have a 90% probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important ("significant") your results.

4.1.4.1 p-Value Interpretation

- The p-value is a number between 0 and 1 and interpreted in the following way:
- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- The p-values very close to the cutoff (0.05) are considered to be marginal (could go either way).
- Always report the p-value so your readers can draw their own conclusions.

The p-value is used to test the significance of the hypothesis.

For example, suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it's more than that. You conduct a hypothesis test because you believe the null hypothesis, Ho, that the mean delivery time is 30 minutes max, is incorrect. Your alternative hypothesis (Ha) is that the mean time is greater than 30 minutes.

You randomly sample some delivery times and run the data through the hypothesis test, and your p-value turns out to be 0.001, which is much less than 0.05. In real terms, there is a probability of 0.001 that you will mistakenly reject the pizza place's claim that their delivery time is less than or equal to 30 minutes.

Since typically we are willing to reject the null hypothesis when this probability is less than 0.05, you conclude that the pizza place is wrong; their delivery times are in fact more than 30 minutes on average. (You could also be wrong by having sampled an unusually high number of late pizza deliveries just by chance.)

4.1.5 Confidence Interval

Confidence interval:

- Is an interval estimate combined with a probability statement.

  (This means that if we used the same sampling method to select different samples and computed an interval estimate for each sample, we would expect the true population parameter to fall within the interval estimates 95% of the time.)
- If repeated samples were taken and the 95% confidence interval was computed for each sample, 95% of the intervals would contain the population mean.
- A 95%confidence interval has a 0.95 probability of containing the population mean. 95% of the population distribution is contained in the confidence interval.

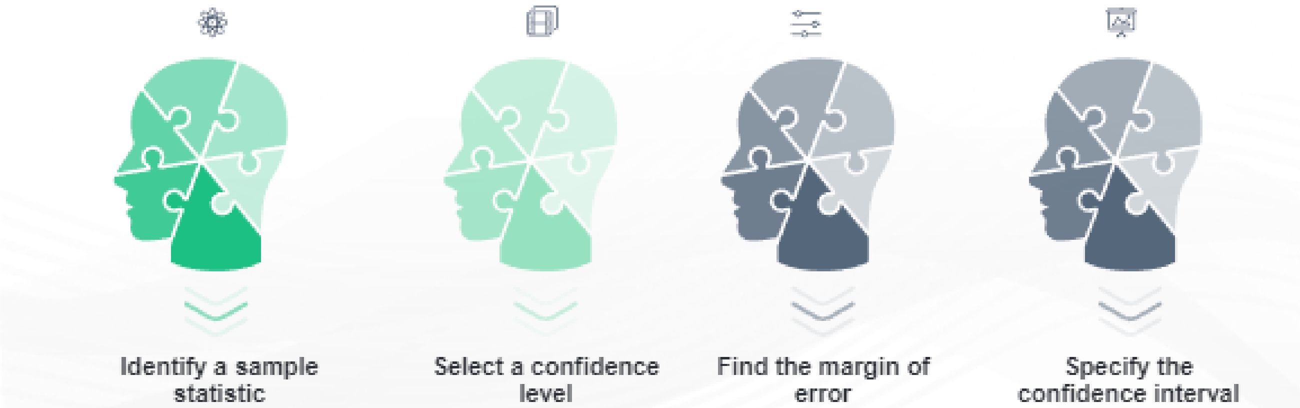
Statisticians use a confidence interval to describe the amount of uncertainty associated with a sample estimate of a population parameter. The confidence level describes the uncertainty associated with a sampling method.

Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter; a 95% confidence level means that 95% of the intervals would include the parameter; and so on.

The purpose of taking a random sample from a lot of population and computing a statistic, such as the mean from the data, is to approximate the mean of the population. A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest.

4.1.5.1 Constructing a Confidence Interval

There are four steps involved in constructing a confidence interval:



To express a confidence interval, you need three pieces of information.

- Confidence level
- Statistic
- Margin of error

There are four steps to constructing a confidence interval.

1. Identify a sample statistic. Choose the statistic (e.g, sample mean, sample proportion) that you will use to estimate a population parameter.
2. Select a confidence level. As we noted in the previous section, the confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
3. Find the margin of error. If you are working on a homework problem or a test question, the margin of error may be given. Often, however, you will need to compute the margin of error, based on one of the following equations.
   Margin of error = Critical value * Standard deviation of statistic
   Margin of error = Critical value * Standard error of statistic
4. Specify the confidence interval. The uncertainty is denoted by the confidence level. And the range of the confidence interval is defined by the following equation.

   Confidence interval = sample statistic + Margin of error

4.1.6 t-Value

T-tests:

- Are statistical hypothesis tests that you use to analyze one or two sample means.
- Depending on the t-test that you use, you can compare a sample mean to a hypothesized value, the means of two independent samples, or the difference between paired samples.

- Are a type of test statistic. Hypothesis tests use the test statistic that is calculated from your sample to compare your sample to the null hypothesis.
- If the test statistic is extreme enough, this indicates that your data are so incompatible with the null hypothesis that you can reject the null.

The term "t-test" refers to the fact that these hypothesis tests use t-values to evaluate your sample data. T-values are a type of test statistic.

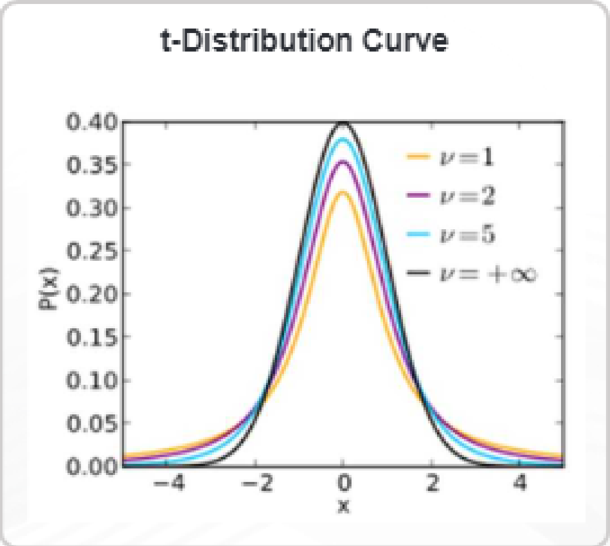
When you analyze your data with any t-test, the procedure reduces your entire sample to a single value, the t-value. These calculations factor in your sample size and the variation in your data. Then, the t-test compares your sample means(s) to the null hypothesis condition in the following manner:

- If the sample data equals the null hypothesis precisely, the t-test produces a t-value of 0.
- As the sample data become progressively dissimilar from the null hypothesis, the absolute value of the t-value increases.

To be able to interpret individual t-values, we have to place them in a larger context. T-distributions provide this broader context so we can determine the unusualness of an individual t-value.

4.1.7 t - Distribution

- The t-distribution (also called Student's t-distribution) is a family of distributions that look almost identical to the normal distribution curve, only a bit shorter and fatter.
- The t-distribution is used instead of the normal distribution when you have small samples.



The t distribution (Student's t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

There are actually many different t distributions. The particular form of the t distribution is determined by its degrees of freedom. The degrees of freedom refers to the number of independent observations in a set of data.

The larger the sample size, the more the t distribution looks like the normal distribution. In fact, for sample sizes larger than 20 (e.g. more degrees of freedom), the distribution is almost exactly like the normal distribution.

For calculations, we use the t-distribution table, which is available here: http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf.

When you look at the t-distribution table, you'll see "df." This means "degrees of freedom" and is just the sample size minus one.

- Step 1: Subtract one from your sample size. This will be your degrees of freedom.
- Step 2: Look up the df in the left hand side of the t-distribution table. Locate the column under your alpha level (the alpha level is usually given to you in the question).

4.1.7.1 Properties of t-Distribution

The properties of t-distribution are as follows:

- The t distribution ranges from $-\infty$ to ∞.
- The t distribution is bell shaped and symmetric like the standard normal curve.
- The shape of the t distribution changes with the change in degrees of freedom.
- The variance of the t distribution is always greater than one and is defined only for 3 or more degrees of freedom.
- The t distribution has a greater dispersion than the standard normal distribution.

The properties of t-distribution are listed above in the slide.

4.1.7.2 t-Distribution Formula

t-Distribution is given by the formula:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

where

$\bar{x}$ is the mean of the first sample;

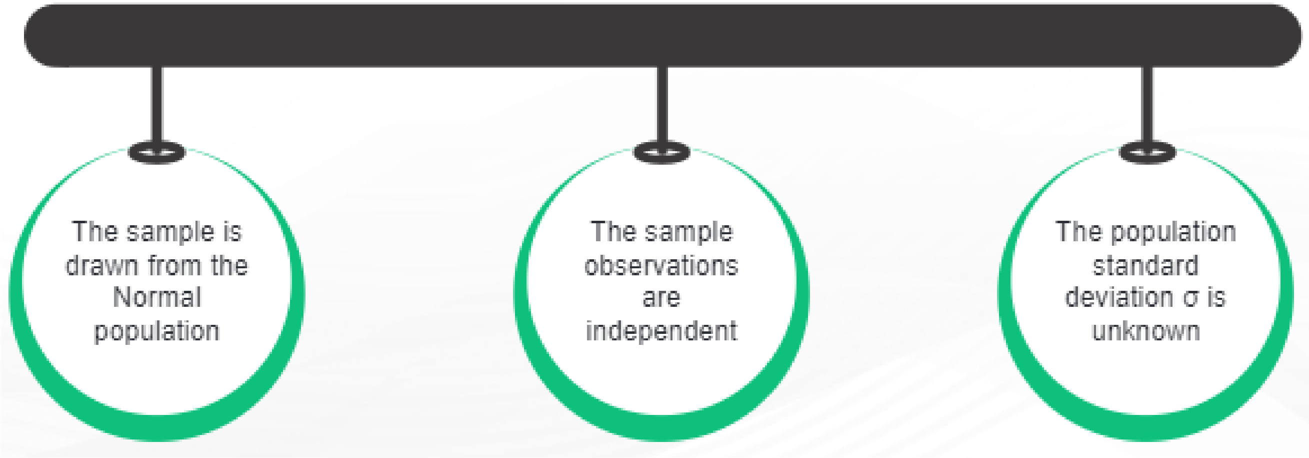
μ is the mean of the second sample;

$\frac{s}{\sqrt{N}}$ is the estimate of the standard error of the difference between the means.

The formula for calculating t-distribution is given in the slide.

4.1.7.3 Assumptions of t-Distribution

The following are the assumptions of student's t test:



Based on the assumptions given above, the test of hypothesis of population mean is given as follows:

When the population standard deviation σ is unknown then to test the population means we use the t statistic. The null hypothesis is there is no difference between the sample mean and the population mean and alternative hypothesis is there is difference between the sample mean and the population mean.

We find the t table value at α level of significance and n –1 degrees of freedom. If the calculated value is less than the critical value, then we accept the null hypothesis that the population mean and the sample mean are equal. Otherwise we reject the null hypothesis and conclude that the population mean and the sample mean are not equal.

4.1.7.4 Example for t-Distribution

Problem:

Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

There are two ways to solve this problem, using the T Distribution Calculator. Both approaches are presented in the next slide.

Solution A is the traditional approach. It requires you to compute the t statistic, based on data presented in the problem description. Then, you use the T Distribution Calculator to find the probability.

Solution B is easier. You simply enter the problem data into the T Distribution Calculator. The calculator computes a t statistic "behind the scenes", and displays the probability. Both approaches come up with exactly the same answer.

The first thing we need to do is compute the t statistic, based on the following equation:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

t = (290 - 300) / [50 / sqrt(15)]

t = -10 / 12.909945 = - 0.7745966

- Now, we can use the t distribution Calculator. Since we know the t statistic, we select "t score" from the random variable drop down. The degrees of freedom are equal to 15 - 1 = 14 and the t statistic is equal to - 0.7745966.
- The calculator displays the cumulative probability: 0.226. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

We've seen one way of calculating the probability. Now, let's look at the other way. This time, we will work directly with the raw data from the problem. We will not compute the t statistic; the t Distribution Calculator will do the math for us.

Since we will work with the raw data, we select "Sample mean" from the Random Variable drop down. Then, we enter the following data:

- The degrees of freedom are equal to 15 - 1 = 14.
- The population mean equals 300.
- The sample mean equals 290.
- The standard deviation of the sample is 50.

The calculator displays the cumulative probability: 0.226. Hence, there is a 22.6% chance that the average sampled light bulb will burn out within 290 days.

4.1.8 Chi-square Distribution

The distribution of the chi-square statistic is called the chi-square distribution. The **chi-square distribution** is defined by the following probability density function:

$Y = Y_0 * (X^2)^{(v/2 - 1)} * e^{-X2 / 2}$

- Y_0 is a constant that depends on the number of degrees of freedom,
- X_2 is the chi-square statistic,
- v = n - 1 is the number of degrees of freedom,
- e is a constant equal to the base of the natural logarithm system (approximately 2.71828).
- Y_0 is defined, so that the area under the chi-square curve is equal to one.

4. Bayes Theorem
5. Discrete Probability Distribution
6. Covariance and Correlation
7. Continuous Probability Distribution
8. Central Limit Theorem
9. Hypothesis Testing

1. Principles of Counting

Counting principles:

- Important to handle large scale data and to get a good understanding of probability.
- The branch of mathematics that deals with various counting techniques is called 'combinatorics'.

Basic techniques of counting:

- Addition Rule
- Multiplication Rule
- Permutation
- Combination

Counting techniques are important to handle large masses of data. Understanding of the Principles of counting is helpful for gaining a better understanding of probability theory. The following are the basic techniques of counting:

- **Addition rule:**

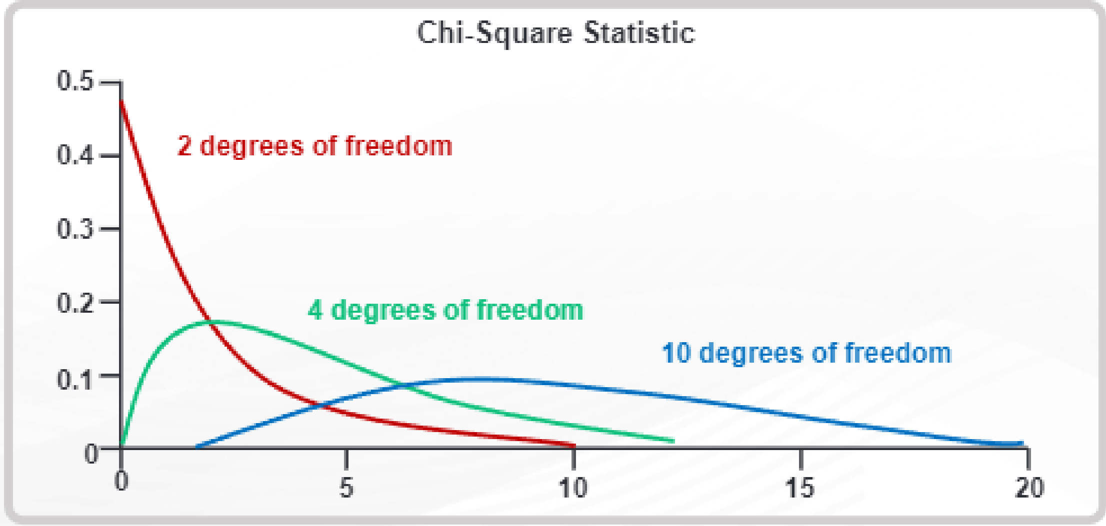
  The addition rule states that "If E1 with n1 outcomes and E2 with n2 outcomes are mutually exclusive events, and if E is the situation where either E1 or E2 will occur, then

  n = n1 + n2

The chi-square distribution results when v independent variables with standard normal distributions are squared and summed. In a testing context, the chi-square distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the chi-square distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

4.1.8.1 Chi-square Statistics

The graph illustrates chi-square distribution of different sample sizes.



In the graph above, the red curve shows the distribution of chi-square values computed from all possible samples of size 3, where degrees of freedom is n - 1 = 3 - 1 = 2.

Similarly, the green curve shows the distribution for samples of size 5 (degrees of freedom equal to 4);

The blue curve, for samples of size 11 (degrees of freedom equal to 10).

4.1.8.2 Properties of Chi-square Distribution

The chi-square distribution has the following properties:

- The mean of the distribution is equal to the number of degrees of freedom: $\mu = v$.
- The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2*v$.
- When the degrees of freedom is greater than or equal to 2, the maximum value for Y occurs when $X^2 = v - 2$.
- As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

The properties of chi-square distribution are listed in the slide.

4.1.8.3 Example for Chi-square Distribution

Problem:

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose, the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

Your facilitator will now guide you in solving the given problem.

4.1.8.3 Example for Chi-square Distribution (Contd.)

Solution:

- The standard deviation of the population is 4 minutes.
- The standard deviation of the sample is 6 minutes.
- The number of sample observations is 7.

To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

$X^2 = [(n - 1) * s^2] / \sigma^2$

$X^2 = [(7 - 1) * 6^2] / 4^2 = 13.5$

where X^2 is the chi-square statistic, n is the sample size, s is the standard deviation of the sample, and σ is the standard deviation of the population.

Let's revisit the problem presented in the previous slide. The manufacturing department ran a quality control test, using 7 randomly selected batteries. In their test, the standard deviation was 6 minutes, which equated to a chi-square statistic of 13.5.

Suppose they repeated the test with a new random sample of 7 batteries. What is the probability that the standard deviation in the new test would be greater than 6 minutes?

Solution

We know the following:

The sample size n is equal to 7.

The degrees of freedom are equal to n - 1 = 7 - 1 = 6.

The chi-square statistic is equal to 13.5 (from the above example).

Given the degrees of freedom, we can determine the cumulative probability that the chi-square statistic will fall between 0 and any positive value. To find the cumulative probability that a chi-square statistic falls between 0 and 13.5, we enter the degrees of freedom (6) and the chi-square statistic (13.5) into the Chi-Square Distribution Calculator. The calculator displays the cumulative probability: 0.96.

This tells us that the probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This means that the probability that the standard deviation would be greater than 6 minutes is 1 - 0.96 or 0.04.

5. Central Limit Theorem

- The central limit theorem (CLT) is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.
- Let $X_1, X_2, ..., X_n$ be a random sample of size n and sample mean from a distribution with mean μ and variance $\sigma^2 < \infty$, then the limiting distribution of

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- is standard normal, that is Zn converges in distribution to Z where Z denotes a standard normal random variable.

The central limit theorem (also known as Lindeberg-Levy Theorem) states that even though the population distribution may be far from being normal, yet for large sample size n, the distribution of the standardized sample mean is approximately standard normal with better approximations obtained with the larger sample size.

All samples will follow an approximately normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample size.

According to the central limit theorem, the mean of a sample of data will be closer to the mean of the overall population in question as the sample size increases, notwithstanding the actual distribution of the data, and whether it is normal or non-normal.

As a general rule, sample sizes equal to or greater than 30 are considered sufficient for the central limit theorem to hold, meaning the distribution of the sample means is fairly normally distributed.

5.1 Convergence in Distribution

- The type of convergence used in the Central Limit Theorem is called the convergence in distribution, which is defined as follows:
- Suppose X is a random variable with cumulative density function F(x) and the sequence $X_1, X_2, ...$ of random variables with cumulative density functions $F_1(x), F_2(x), ...$, respectively. The sequence X_n converges in distribution to X if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

(for all values x at which F(x) is continuous. The distribution of X is called the limiting distribution of X_n)

Whenever a sequence of random variables X1, X2, ... converges in distribution to the random variable X, it will be denoted by X_n X.

5.2 Properties of Central Limit Theorem

Central Limit Theorem:

- Specifies a theoretical distribution.
- Is formulated by the selection of all possible random samples of a fixed size 'n'.
- Sample mean is calculated for each sample and the distribution of sample means is considered.

The properties of central limit theorem are listed in the slide.

5.3 Examples for Central Limit Theorem

- If an investor is looking to analyze the overall return for a stock index made up of 1,000 stocks, he can take random samples of stocks from the index to get an estimate for the return of the total index.
- The samples must be random, and at least 30 stocks must be evaluated in each sample for the central limit theorem to hold. Random samples ensure a broad range of stock across industries and sectors is represented in the sample.
- Stocks previously selected must also be replaced for selection in other samples to avoid bias. The average returns from these samples approximate the return for the whole index and are approximately normally distributed. The approximation holds even if the actual returns for the whole index are not normally distributed.

Your facilitator will explain the given example on Central Limit Theorem.

5.4 Central Limit Theorem for Sums

- Suppose, X is a random variable with a distribution that may be known or unknown (it can be any distribution), and suppose:

  Ax = the mean of X

  Cx = the standard deviation of x
- The central limit for sums states that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample increases.

The normal distribution has a mean equal to the original mean multiplied by the sample size

The standard deviation is equal to the original standard deviation multiplied by the square root of the sample size The Central Limit Theorem CLT for Sums Using the Central Limit Theorem

5.5 Example for Central Limit Theorem

Problem:

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn from the population.

Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.

Solution:

Let X = one value from the original unknown population. The probability question asks you to find a probability for the sum (or total of) 80 values.

ÆX = the sum or total of 80 values. Since μx= 90, σx= 15, and n=80, ÆX ~N((80),(90),(80)(15)) – Mean of the sums = (n)(μx) = (80)(90) = 7,200 – Standard deviation of the sums = n σx = 80 15 – Sum of 80 values = Æx = 7,500

6. Hypothesis Testing

- There are two main methods used in inferential statistics:

  1. Estimation

  2. Hypothesis testing

- A Hypothesis is a statement or an assumption about relationships between variables.

  or

- A Hypothesis is a tentative explanation for certain behaviors, phenomenon or events that have occurred or will occur.

We already learned in the previous module about the basics and the types of statistics: descriptive and inferential statistics. There are two major methods used in inferential statistics:

- Estimation
- Hypothesis testing

A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation.

Examples:

1. A new medicine you think might work.
2. A way of teaching you think might be better.
3. A possible location of new species.
4. A fairer way to administer standardized tests.

Some interesting hypotheses:

Bankers assumed high-income earners are more profitable than low-income earners. Old clients were more likely to diminish CD balances by large amounts compared to younger clients. This was non-intrusive because conventional wisdom suggested that older clients have a larger portfolio of assets and seek less risky investments.

6.1 Hypothesis Statement

- If you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this:
  - "If I…(do this to an independent variable)….then (this will happen to the dependent variable)."
- For example:
  - If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
  - If I (give patients counseling in addition to medication) then (their overall depression scale will decrease).
  - If I (give exams at noon instead of 7) then (student test scores will improve).
  - If I (look in this certain location) then (I am more likely to find new species).

A good hypothesis statement should:

- Include an "if" and "then" statement.
- Include both the independent and dependent variables.
- Be testable by experiment, survey or other scientifically sound technique.
- Be based on information in prior research (either yours or someone else's).
- Have design criteria (for engineering or programming projects).

6.2 Criteria for Hypothesis Construction

A hypothesis should meet the following criteria:

- It should be empirically testable, whether it is right or wrong.

- It should be specific and precise.
- The statements in the hypothesis should not be contradictory.
- It should specify variables between which the relationship is to be established.
- It should describe one issue only.

The slide lists the criteria that a hypothesis statement should meet.

6.3 Steps in Hypothesis Testing

The following are the steps for constructing a hypothesis test.

1. Select the type of hypothesis
   a) Null Hypothesis (H0)
   b) Alternative Hypothesis (Ha or H1)
2. Establish Critical or Rejection region
3. Select the Suitable Test of significance or Test Statistic
4. Check whether the test involves one sample, two samples, or multiple samples?
5. Check whether two or more samples used are independent or related?
6. Is the measurement scale nominal, ordinal, interval, or ratio?
7. The choice of a probability distribution of a sample statistics is guided but the sample size n and the value of population standard deviation.
8. Formulate a Decision Rule to Accept Null Hypothesis.
   a) Accept H0 if the test statistic value falls within the area of acceptance.
9. Reject otherwise.

The steps involved in a hypothesis test are given in the slide. Hypothesis testing can be one of the most confusing aspects for students, mostly because before you can even perform a test, you have to know what your null hypothesis is. Often, those tricky word problems that you are faced with can be difficult to decipher. But it's easier than you think; all you need to do is:

- Figure out your null hypothesis,
- State your null hypothesis,
- Choose what kind of test you need to perform,
- Either support or reject the null hypothesis.

The first step is to specify the null hypothesis. For a two-tailed test, the null hypothesis is typical that a parameter equals zero although there are exceptions. A typical null hypothesis is $\mu1 - \mu2 = 0$ which is equivalent to $\mu1 = \mu2$. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that μ1 is larger than μ2, then the null hypothesis (the reverse of the prediction) is $\mu2 - \mu1 \geq 0$. This is equivalent to $\mu1 \leq \mu2$.

The second step is to specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.

The third step is to compute the probability value (also known as the p-value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.

Finally, compare the probability value with the α level. If the probability value is lower then you reject the null hypothesis. Rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

6.4 Null Hypothesis

- If you trace back the history of science, the null hypothesis is always the accepted fact.
- Simple examples of null hypotheses that are generally accepted as being true are:
    - DNA is shaped like a double helix.
    - There are 8 planets in the solar system (excluding Pluto).

The statistical procedure for testing a hypothesis requires some understanding of the null hypothesis. Think of the outcome (dependent variable). From a statistical (and sampling) perspective, the null hypothesis asserts that the samples being compared or contrasted are drawn from the same population with regard to the outcome variable. This means that

- any observed differences in the dependent variable (outcome) must be due to sampling error (chance)
- the independent (predictor) variable does not make a difference

The symbol H0 is the abbreviation for the null hypothesis, the small zero stands for null.

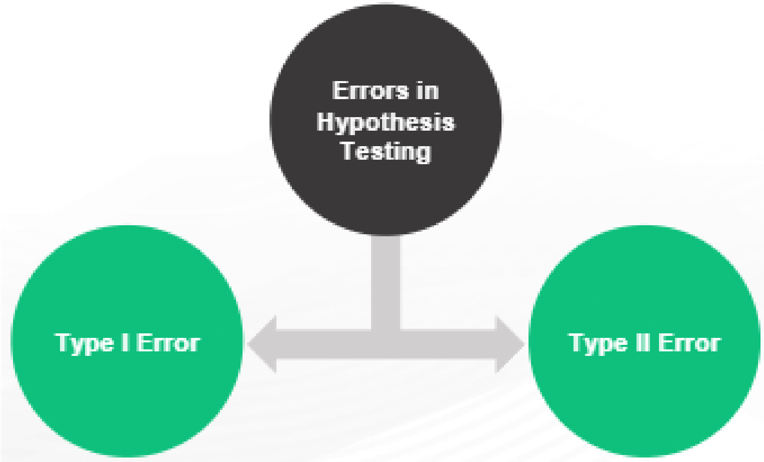
Stating a null hypothesis:

You won't be required to actually perform a real experiment or survey in elementary statistics (or even disprove a fact like "Pluto is a planet"!), so you'll be given word problems from real-life situations.

You'll need to figure out what your hypothesis is from the problem. This can be a little trickier than just figuring out what the accepted fact is. With word problems, you are looking to find a fact that is mollifiable (i.e. something you can reject).

6.5 Errors in Hypothesis Testing

Two types of hypothesis testing:



There are two types of errors in hypothesis testing.

1. Type I error
2. Type II error

6.5.1 Type I Error

Types I Error:

- It is also known as an error of the first kind, occurs when the null hypothesis (H0) is true, but is rejected.
- A type I error may be compared with a so called false positive.
- A Type I error occurs when we believe a falsehood.
- The rate of the type I error is called the size of the test and denoted by the Greek letter α (alpha).It usually equals the significance level of a test. If type I error is fixed at 5%, it means that there are about 5 chances in 100 that we will reject H0 when H0 is true.

More generally, a Type I error occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05, then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α (alpha) level or simply α. It is also called the significance level.

The Type I error rate is affected by the α level: the lower the α level, the lower the Type I error rate. It might seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

6.5.2 Type II Error

Types II Error:

- It is also known as an error of the second kind, occurs when the null hypothesis is false, but erroneously fails to be rejected.
- Type II error means accepting the hypothesis which should have been rejected. A type II error may be compared with a so-called False Negative.
- A Type II error is committed when we fail to believe a truth.
- A type II error occurs when one rejects the alternative hypothesis (fails to reject the null hypothesis) when the alternative hypothesis is true.
- The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test (which equals 1-β).

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a Type II error. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called β (beta). The probability of correctly rejecting a false null hypothesis equals 1- β and is called power.

6.5.3 Error Scenarios in Hypothesis Testing

The following table summarizes the error scenarios in a hypothesis testing.

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

Correct decisions:

- The null hypothesis is true and is accepted.
- The null hypothesis is false and is rejected.

Incorrect decisions:

Type I Error: The null hypothesis is true, but it is rejected. The probability of committing a type I error is denoted by α or p-value.

Type II Error: The null hypothesis is false, but it is accepted. The probability of committing type II error is denoted by β.

6.5.4 Type III Error

Types III Error:

- Many statisticians are now adopting a third type of error, type III, where the null hypothesis is rejected for the wrong reason.
- In an experiment, a researcher might assume a hypothesis and perform research. After analyzing the results, the null is rejected.
- The problem, that there may be some relationship between the variables, but it could be for a different reason than stated in the hypothesis.
- An unknown process may underlie the relationship.

A type III error is where you correctly reject the null hypothesis, but it's rejected for the wrong reason. This compares to a Type I error (incorrectly rejecting the null hypothesis) and a Type II error (not rejecting the null when you should). Type III errors are not considered serious, as they do mean you arrive at the correct decision. They usually happen because of random chance and are a rare occurrence.

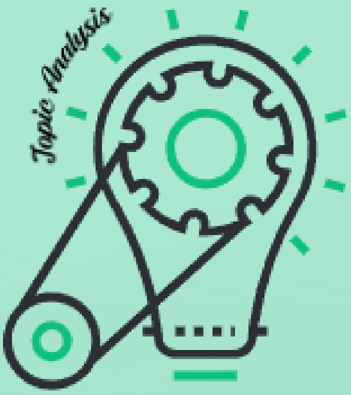
You can also think of a Type III error as giving the right answer (i.e. correctly rejecting the null) to the wrong question. Either way, you're still arriving at the correct conclusion for the wrong reason. When we say the "wrong question", that normally means you've formulated your hypotheses incorrectly. In other words, both your null and an alternative hypothesis may be poorly worded or completely incorrect.

What did You Grasp?

1. State True or False.

   A type II error is one in which a false null hypothesis is not rejected.

   A) True
   B) False



2. Fill in the blank.

   The threshold for rejecting a null hypothesis is also called ____.

   A) Confidence level
   B) Significance level
   C) Hypothesis level
   D) None of the above

In a nutshell, we learnt:

1. The principles of counting including addition rule, product rule, permutation and combination.
2. Introduction and important definitions of Probability Theory.
3. The concept of Conditional Probability.
4. Bayes Theorem and its applications.
5. Different methods of discrete probability distribution like discrete uniform distribution, Poisson distribution, Bernoulli distribution and Binomial distribution.
6. The concepts of Covariance and Correlation.
7. Different methods of Continuous Probability Distribution, including normal distribution, continuous uniform distribution, exponential distribution, confidence intervals, t-distribution and chi-square distribution.
8. Properties of Central Limit Theorem with the help of examples.
9. Hypothesis Testing and the types of errors.

