PERIYAR UNIVERSITY

(NAAC 'A++' Grade with CGPA 3.61 (Cycle - 3) State University - NIRF Rank 56 -State Public University Rank 25)

SALEM - 636 011, Tamil Nadu, India.

CENTRE FOR DISTANCE AND ONLINE EDUCATION

(CDOE)

M.A ECONOMICS

SEMESTER - I

CORE III: STATISTICS FOR ECONOMISTS

(Candidates admitted from 2025 onwards)

PERIYAR UNIVERSITY

CENTRE FOR DISTANCE AND ONLINE EDUCATION (CDOE)

M.A Economics 2025 admission onwards

CORE III

STATISTICS FOR ECONOMISTS

Prepared by:

Dr.R.Asokan
Associate Professor
Dept. of Economics
Annamalai University
Annamalai Nagar- 608 002

Scrutinized & Verified by:

BOS Members,
Centre for Distance and Online Education (CDOE)
Periyar University
Salem - 636011

M A ECONOMICS

CORE – III

STATISTICS FOR ECONOMISTS

Course Objective:

1. To provide a strong foundation in statistical concepts and develop skills in data handling and research.

2. The course facilitates in inferring the intensity of relationship between multiple variables and building appropriate statistical models. The models thus formulated can be tested for their significance and can be used for forecasting.

Unit I: Probability

Probability - Addition and Multiplication Theorems - Conditional Probability - Discrete and Continuous - Random Variables - Mathematical Expectations – Bayes Theorem - Theoretical Distributions - Binomial, Poisson and Normal.

Unit II: Sampling and Hypothesis Testing

Sampling Theory - Types of Sampling - Sampling Distributions - Parameter and Statistic - Testing of Hypothesis - Level of Significance - Type I and Type II Errors - Standard Error - Properties of Estimator.

Unit III: Test of Significance

Large and Small Sample Difference between Large and Small Samples - Test of Significance for Large Samples - Test for Two Means and Standard Deviations - Proportion and Confidence Interval - Small Sample Test – t-test - Paired t- test - Chi-square Test- Test of Goodness of Fit.

Unit IV: Analysis of Variance

F test: Assumptions in F test - Analysis of Variance: Assumptions – One-Way and Two-Way Classifications.

Unit V: Statistical Decision Theory

Definitions – Concepts – Maximin - Minimax - Bayes Criterion - Expected Monetary Value - Decision Tree Analysis: Symbols - Steps - Advantages and Limitations.

Text Books

1.Gupta S.P., Statistical Methods, Sultan Chand and Sons, New Delhi, 2017.

2. Anderson, Sweeney and Williams, —Statistics for Business and Economics‖,Cengage, 2014.

References:

1. Aggarwal. Y.P (2002), —Statistics Methods – Concepts Application and Computation‖, Sterling Publishers Private Ltd., New Delhi.

2.Vittal P.R.,Mathematical Statistics, Margham Publications 3. Pillai R.S.N. andBagavathi V (2010), Statistics, Sultan & amp; Chand Sons, NewDelhi.

Web Resources

1. https://www.statista.com.
2. https://techjury.net
3. https://dss.princeton.edu/online_help/analysis/interpreting regression.htm

TABLE OF CONTENTS		
S. No	**TITLE OF THE UNITS**	**PAGE NO**

PERIYAR UNIVERSITY

(NAAC 'A++' Grade with CGPA 3.61 (Cycle - 3) State University - NIRF Rank 56 -State Public University Rank 25)

SALEM - 636 011, Tamil Nadu, India.

CENTRE FOR DISTANCE AND ONLINE EDUCATION

(CDOE)

M.A HISTORY

SEMESTER - II

CORE VI: HISTORIOGRAPHY AND HISTORICAL METHOD

(Candidates admitted from 2025 onwards)

PERIYAR UNIVERSITY

CENTRE FOR DISTANCE AND ONLINE EDUCATION (CDOE)

M.A 2025 admission onwards

CORE VI

HISTORIOGRAPHY AND HISTORICAL METHOD

Prepared by:

Centre for Distance and Online Education (CDOE)
Periyar University
Salem – 636011

SEMESTER II

CORE VI

HISTORIOGRAPHY AND HISTORICAL METHOD

Course Objectives	
1	To explain the concepts related to history and its relationship with other disciplines
2	To discuss various philosophies and interpretations of history
3	To explain the processes and procedures involved in the conduct of historical research
4	To examine the evolution of historical writing in the West
5	To examine the contribution of various historians to the development of Indian historiography

Syllabus
UNIT I: Meaning, Nature and Scope of History – Kinds of History and Allied Subjects – Lessons of History; Uses and Abuses of History – Role of Individuals, Role of Institutions and Role of Ideas in History.
UNIT II: Philosophy of History – Positivist History – Marxist Interpretation of History – Annales Paradigm – Subaltern History – Subjectivity and Need for Objectivity in History.
UNIT III: Historical Research: Pre-requisites of a Researcher – Choice of Topic – Review of Literature – Hypothesis – Sources of History– External and Internal Criticism of Sources– Collection of Data, Synthesis, Exposition and Writing – Use of Footnotes and preparation of Bibliography.
UNIT IV: Development of Historical writing in the West – Herodotus,Thucydides, St. Augustine, IbnKhaldun, L.V. Ranke, Arnold Toynbee, E.H. Carr, FernandBraudel, E.P. Thompson, Eric Hobsbawm.
UNIT V: Historians of India – V.A. Smith, D.D. Kosambi, RomilaThapar, JadunathSarkar,Bipan Chandra, RanajitGuha, K.A. NilankantaSastri, R. SathianathaAyyar, S. KrishnaswamiAyyangar, C.S. Srinivasachari,K.K. Pillai.

TABLE OF CONTENTS		
S. No	**TITLE OF THE UNITS**	**PAGE NO**

Unit- I

Probability Theorems

Probability is a measure of the likelihood that an event will occur. It's a value between 0 and 1, where 0 indicates impossibility and 1 indicates certainty. Probability is used to analyse the likelihood of various outcomes in random events.

Key Concepts:

- Event: A specific outcome or set of outcomes of a random phenomenon.
- Sample Space: The set of all possible outcomes of a random phenomenon.
- Probability of an Event: The likelihood of that event occurring, expressed as a number between 0 and 1.

For example, if you flip a fair coin, the probability of getting heads is 1/2 (or 0.5) because there's one favourable outcome (heads) and two total possible outcomes (heads or tails).

Types of Probability:

Classical Probability: Based on equally likely outcomes (e.g., rolling a die).

Empirical Probability: Based on observed frequencies in past data (e.g., the probability of rain based on historical weather patterns).

Subjective Probability: Based on personal judgment and experience (e.g., estimating the probability of success for a business venture).

Axiomatic Probability: A rigorous mathematical approach using axioms to define probability.

Probability Theorem of Complementary Events

Two events are said to be complementary events if the sum of their probability is 1. Thus, if A is an event and the probability of A is given by P(A) then this theorem states that

P(A') = 1 - P(A)

where P(A') is the probability of the complementary event of A; i.e., A'. In such cases, events A and A' are said to be mutually exhaustive also.

Example:

Consider an event A that 3 will appear on rolling a dice. Calculate the probability of not getting a 3.

Solution:

Probability of getting a 3 on dice = P(A)=16*P*(*A*)=61

The probability of A' which is the probability of not getting a 3 is calculated using the theorem of complementary events as follows:

P(A') = 1 - P(A)

P(A′)=1−16P(A′)=1−61

P(A′)=56P(A′)=65

Theorem of Addition

Theorem of Addition is used when one has to determine the probability of occurrence of two or more events. In simple terms, this theorem is used to calculate the probability of union of two or more than two events. For instance, there are two events E1 and E2 of a given sample space. By using the theorem of addition, we can determine the probability that either E1 or E2 will occur. However, to determine the probability, first of all, we have to find out whether the events are mutually exclusive or overlapping, after that only the required probability is calculated using the correct rule or formula.

Theorem of Addition has two cases:

i) When the events are Mutually Exclusive

If A and B are mutually exclusive events then according to this theorem:

P(E1 ∪ E2) = P(E1) + P(E2)

where, (E1 ∪ E2) means either E1 or E2

If there are more than two events, then

P(E1 ∪ E2 ∪ ..... En) = P(E1) + P(E2) + ... + P(En)

If these events are collectively exhaustive, then

P(E1 ∪ E2 ∪ ..... En) = P(E1) + P(E2) + ... + P(En) = 1

This rule is known as the Theorem of Addition for Mutually Exclusive Events.

Example:

If A and B are events of occurrence of 2 and occurrence of 3 on a dice respectively, then calculate the probability of occurrence of A or B; i.e., getting 2 or 3 on dice.

Solution:

Probability of getting 2=P(A)=16Probability of getting 2=P(A)=61

Probability of getting 3=P(B)=16Probability of getting 3=P(B)=61

As A and B are independent and mutually exclusive events, using the theorem of addition,

P(A∪B) = P(A) + P(B)

P(A∪B)=16+1661+61

P(A∪B)=1331

ii) When the Events are Overlapping

When the events are overlapping, the theorem of addition determines the probability that one or more events would occur in a single trial.

If E1 and E2 are overlapping events then according to this theorem:

P(E1 ∪ E2) = P(E1) + P(E2) - P(E1 ∩ E2)

where, P(E1 ∪ E2) is the probability that either E1 or E2 or both events will occur, and P(E1 ∩ E2) is the joint probability that indicates the probability of occurrence of both E1 and E2.

If there are three events E1, E2, and E3, then

P(E1 ∪ E2 ∪ E3) = P(E1) + P(E2) + P(E3) - P(E1 ∩ E2) - P(E1 ∩ E3) - P(E2 ∩ E3) + P(E1 ∩ E2 ∩ E3)

where P(E1 U E2 U E3) gives the probability of occurrence of at least one of the events; E1, E2, and E3.

Example:

Consider a class with 20 students. 10 students passed in Maths, 15 in English, and 13 in both. What is the probability that a student passed in either Math or English?

Solution:

Total students = 20

P(A) = Probability that a student passed in Maths = 10202010

P(B) = Probability that a student passed in English = 15202015

P(A∩B) = Probability that a student passed in both Math and English = 13202013

This is a case of overlapping events as some students who passed in Math may have passed in English too and some students who passed in English may have passed in Math too. Thus we need to remove these common students from the sum of students who passed in Math and English. Thus, using the theorem of addition for overlapping events, we get:

The probability that a student passed either in Maths or in English = P(AUB)

P(A∪B) = P(A) + P(B) - P(A∩B)

P(A∪B) = 1020+1520−13202010+2015−2013

P(A∪B) = 12202012

P(A∪B) = 3553

Theorem of Multiplication

When we have to determine the probability of joint occurrence or occurrence in unison of two or more than two events, the theorem of multiplication is used. **For instance, the** probability of getting the same number on two dice tossed simultaneously, drawing different coloured balls from a box having red, blue, and green balls.

There are 2 cases in theorem of multiplication:

i) When events are Independent

If E1 and E2 are independent events with P(E1) ≠ 0 and P(E2) ≠ 0, then according to this theorem of probability:

P(E1 ∩ E2) = P(E1).P(E2)

This means that the probability of intersection of two events E1 and E2 is equal to the product of the individual probabilities of events E1 and E2.

If there are more than two events; say, E1, E2, and E3, then,

P(E1 ∩ E2 ∩ E3) = P(E1).P(E2).P(E3)

Example:

A statistics problem is given to two students, say A and B. Their chances of solving it correctly are known to be 0.5 and 0.3, respectively. Find the probability that both of them solve it.

Solution:

Let E1 be the event of A solving the problem and E2 is the event of B solving the problem. Here event E1 and E2 are independent.

P(E1)=0.5P(E1)=0.5

P(E2)=0.3P(E2)=0.3

P(E1∩E2)=P(E1)×P(E2)P(E1∩E2)=P(E1)×P(E2)

P(E1∩E2)=0.5×0.3P(E1∩E2)=0.5×0.3

P(E1∩E2)=0.15P(E1∩E2)=0.15

Therefore, the chances of both A and B solving the problem is 15%.

ii) When Events are not Independent

If the events are not independent, then multiplication theorem states that the joint probability of the events E1 and E2 is given by the probability of event E1 multiplied by the probability of event E2 given that event E1 has occurred and vice-versa. Simply put, the rule uses the concept of conditional probability when the events are known to be dependent or non-independent. According to this theorem, if E1 and E2 are two

events where P(E1) ≠ 0 and P(E2) ≠ 0, and if E1 and E2 are not independent events, then:

P(E1 ∩ E2) = P(E1).P(E2/E1)

P(E1 ∩ E2) = P(E2).P(E1/E2)

Similarly, if there are three dependent events E1, E2, and E3, then,

P(E1 ∩ E2 ∩ E3) = P(E1).P(E2/E1).P(E3/E1∩E2)

Example:

A large company employs 70 engineers, of whom 36 are males and the remaining are females. Of the female engineers. 14 are under 35 years of age, 15 are between 35 and 45 years of age, and the remaining are over 45 years of age. What is the probability of randomly selected engineer who is a female and under the age of 35 years of age?

Solution:

Let E1 represent the event that an engineer selected at random is a female and E2 is the event that a female engineer selected is under 35 years of age.

Since there are 36 males out of 70 engineers, it means that the number of female engineers is 34.

P(E1)
= Total Female EngineersTotal EngineersTotal EngineersTotal Female Engineers

P(E1) = 34707034

P(E2E1)=1434P(E1E2)=3414

Therefore, P(E1∩E2) = P(E1)×P(E2E1)P(E1)×P(E1E2)

P(E1∩E2) = 3470×14347034×3414

P(E1∩E2) = 1551

Statistical Independence

If joint probability of two events E1 and E2 is equal to the product of marginal probability of E1 and E2, then E1 and E2 are said to be statistically independent. Mathematically, two events E1 and E2 are *statistically independent* if:

P(E1∩E2) = P(E1).P(E2)

If this relationship does not hold true, events E1 and E2 are statistically not independent.

Example:

A number is selected randomly from the first n natural numbers. Let E1 be the event that it is divisible by 2, and E2 be the event that it is divisible by 3. Show that the events are statistically independent if n=96.

Solution:

When n=96,

P(E1)=4896=12*P*(*E*1)=9648=21

P(E2)=3296=13*P*(*E*2)=9632=31

and P(E1∩E2)=1696=16*P*(*E*1∩*E*2)=9616=61

Here, P(E1∩E2)=P(E1)×P(E2)*P*(*E*1∩*E*2)=*P*(*E*1)×*P*(*E*2)

as 12×13=1621×31=61

Hence, events E1 and E2 are statistically independent.

Theorem of Total Probability

The theorem of total probability, also known as Theorem of Elimination, is employed to calculate the probability of an event whose occurrence is dependent on the occurrence (or non-occurrence) of some intermediate events in the experimental process. If an event E is associated with other intermediate, mutually exclusive events H1, H2, ..., Hn, then the total probability of its occurrence will be,

P(E) = P(H1 ∩ E) + P(H2 ∩ E) + ... + P(Hn ∩ E)

P(E) = P(H1) x P(E/H1) + P(H2) x P(E/H2) + ... + P(Hn) x P(E/Hn)

Example:

Suppose that the accounting manager of a company wants to introduce a new policy of assessment of employees of the company for their promotion and the policy is subject to clearance from the general manager (GM). At present, the post of GM is vacant and is likely to be filled soon by appointing one of the three deputy GMs. The chances of policy in question being implemented are dependent on who is appointed the GM. Now, the chances of managers A, B, and C being appointed GM are 0.4, 0.5, and 0.1 respectively. While the likelihood of policy implementation is 0.2, 0.6, and 0.4, respectively, with the three. What is the probability that the policy will eventually be implemented?

Solution:

Let's say H1, H2, and H3 represent the respective events that A, B, and C are promoted, and E be the event that policy is implemented. Now,

P(H1) = 0.4

P(H2) = 0.5

P(H3) = 0.1

P(E/H1) = 0.2

P(E/H2) = 0.6

P(E/H3) = 0.4

Now, $P(E) = P(H1 \cap E) + P(H2 \cap E) + P(H3 \cap E)$

$P(E) = P(H1).P(E/H1) + P(H2).P(E/H2) + P(H3).P(E/H3)$

$P(E) = (0.4)(0.2) + (0.5)(0.6) + (0.1)(0.4)$

$P(E) = 0.08 + 0.3 + 0.04$

$P(E) = 0.42$

Therefore, the overall chances that the policy will be introduced are 42%.

Conditional probability

It is the likelihood of an event occurring, given that another event has already happened. It's represented as P(A|B), which means "the probability of event A happening given that event B has already occurred". The formula for calculating conditional probability is P(A|B) = P(A ∩ B) / P(B), where P(B) must be greater than 0.

Here's a breakdown:

- **Event A:** The event we're interested in finding the probability for.
- **Event B:** The event that has already occurred, and whose occurrence influences the probability of event A.
- **P(A ∩ B):** The probability of both events A and B occurring.
- **P(B):** The probability of event B occurring.
- **P(A|B):** The conditional probability of A given B.

In simpler terms: Imagine a situation where you're drawing a card from a deck. If someone tells you they drew a king, and then asks what the probability of it being a heart is, that's conditional probability. You're no longer working with the entire deck, but with the reduced sample space of only the kings. The conditional probability would be the number of kings that are also hearts, divided by the total number of kings.

Conditional Probability Formula

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

Or:

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

Where:

- **P = Probability**
- **A = Event A**
- **B = Event B**

Examples of Conditional Probability

Example 1: Ceramic plates in a Bag

An example of conditional probability using ceramic plates is illustrated below. The steps are as follows:

Step i): Understand the scenario

Initially, you're given a bag with six red ceramic plates, three blue ceramic plates, and one green ceramic plates. Thus, there are 10 ceramic plates in the bag.

Step ii): Identify the events

Two events are defined:

- Event A: Drawing a red ceramic plates from the bag
- Event B: Drawing a ceramic plates that is not green

Step iii): Calculate the probability of event B: P(B)

Event B is drawing a ceramic plates that is not green. There are 10 ceramic plates altogether, nine of which are not green: the six red and three blue ceramic plates.

P(B)=(Number of ceramic plates that are not green)/(Total number of ceramic plates)=9/10P(B)=(Number of ceramic plates that are not green)/(Total number of ceramic plates)=9/10

Step iv): Identify the intersection of events A and B: P(A∩B)

The intersection of events A and B involves drawing a red ceramic plates that is also not green. Since all red ceramic plates are not green, the intersection is simple: the event of drawing a red ceramic plates.

Step v): Calculate the probability of the intersection of events A and B: P(A∩B)

P(A∩B)=(Number of red ceramic plates)/(Total number of ceramic plates)=6/10=3/5P(A∩B)=(Number of red ceramic plates)/(Total number of ceramic plates)=6/10=3/5

Step vi): Calculate the conditional probability: P(A|B)

Using the conditional probability formula, P(A|B), that is, the probability of drawing a red ceramic plates given that the ceramic plates drawn is not green, the probability is calculated.

P(A|B)=P(A∩B)/P(B)=(3/5)/(9/10)=2/3*P*(*A*|*B*)=*P*(*A*∩*B*)/*P*(*B*)=(3/5)/(9/10)=2/3

Result: The conditional probability of drawing a red ceramic plates given that the ceramic plates drawn is not green, is 2/3.

Example 2: Rolling a Fair Die

Let's consider another example of conditional probability using a fair die. The steps are as follows:

Step i): Understand the scenario

You have a fair six-sided die. You want to determine the probability of rolling an even number, given that the number rolled is greater than four.

Step ii): Identify the events

The possible outcomes (sample space) for a six-sided die are the numbers one through six. From this list, you can define the two events:

- Event A: Rolling an even number. Event A would mean rolling {2,4,6}.
- Event B: Rolling a number greater than four. Event B would mean rolling {5,6}.

Step iii): Calculate the probability of each event

The probability of each event can be calculated by dividing the number of favorable outcomes (the ones you're looking for) by the total number of outcomes in the sample space.

P(A) is the probability of rolling an even number. There are three even numbers {2,4,6} out of the six possible outcomes. Thus, P(A) = 3/6 = 1/2.

P(B) is the probability of rolling a number greater than four. Two numbers are greater than four {5,6} out of the six possible outcomes. Thus, P(B) = 2/6 = 1/3.

Step iv): Identify the intersection of events A and B

The intersection of events A and B includes the outcomes that satisfy both conditions simultaneously. In this case, that means rolling a number that is even and also greater than four. The only outcome that does both is rolling a six.

Step v): Calculate the probability of the intersection of events A and B

We'll spell this out, even if it's easy, given the above, because other examples might prove more difficult: P(A∩B) is the probability of rolling six, since six is the only outcome that is both even and greater than four. There is one outcome out of six possibilities. Therefore, P(A∩B) = 1/6.

Step vi): Calculate the conditional probability: P(B|A)

The formula for conditional probability is as follows:

$$P(B|A)=P(A\cap B)/P(A) \quad P(B|A)=P(A\cap B)/P(A)$$

When the values are substituted into the formula, here is the result:

$$P(B|A)=(1/6)/(1/2)=1/3 \quad P(B|A)=(1/6)/(1/2)=1/3$$

Result: This means that given the die rolled is even, the probability that this number is also greater than four is 1/3.

Example 3: Multiple Conditional Probabilities

Another scenario involves a student applying for admission to a college who hopes to get a scholarship and a stipend for books, meals, and housing. The steps to determine the conditional probability of getting a stipend and the scholarship are as follows:

Step i) : Understand the scenario

First, the student wants to know the likelihood of being accepted to the university. Then, if accepted, the student would like to receive an academic scholarship. Moreover, if possible, the student would also like to receive a stipend for books, meals, and housing if they get the scholarship.

Step ii): Identify the events

There are three events:

- Event A: Being accepted to the university.
- Event B: Receiving a scholarship upon acceptance
- Event C: Receiving a stipend for books, meals, and housing upon receiving a scholarship

Step iii): Calculate the probability of being accepted (event A)

The university accepts 100 out of every 1,000 applicants who have applications similar to the students. Thus, the probability of a student being accepted is P(A) = 100/1000 = 0.10 or 10%.

Step iv): Determine the probability of receiving a scholarship once accepted: P(B|A)

It's known that out of the students accepted, 10 out of every 500 receive a scholarship. Thus the probability of receiving a scholarship given acceptance is as follows:

Step v): Calculate the probability of being accepted and receiving a scholarship

To calculate the probability of being accepted and also receiving a scholarship, the likelihood of acceptance is multiplied by the conditional probability of receiving a scholarship given acceptance.

Step vi): Determine the probability of receiving a stipend after receiving a scholarship: P(C|B)

It's also known that among the scholarship recipients, 50% receive a stipend for books, meals, and housing. Thus, P(C|B) = 0.5 = 50%.

Step vii): Calculate the probability of being accepted, receiving a scholarship, and receiving a stipend

To calculate the probability of a student being accepted, receiving a scholarship, and then also receiving a stipend, the probabilities of the events are multiplied.

This step-by-step breakdown illustrates how the probabilities for each scenario are calculated using basic probability formulas and conditional probability.

Conditional Probability vs. Joint Probability and Marginal Probability

Let's now differentiate calculating conditional probability from other kinds of probability.

Conditional Probability

The example this time is a regular deck of cards. Two events are defined:

- Event A: Drawing a four
- Event B: Drawing a red card

A standard deck has 52 cards divided into four suits (hearts, diamonds, clubs, and spades). Hearts and diamonds are red, and clubs and spades are black. Each suit has 13 cards: Ace, then two through 10, and then the face cards Jack, Queen, and King.

The deck contains 26 red cards, 13 hearts, and 13 diamonds. Thus, the probability of drawing a red card is P(B) = 26/52 = 1/2.

Within the red cards are a four of hearts and a four of diamonds. Therefore, if a red card has to be drawn, a subset of the deck that includes only these 26 red cards needs to be considered.

Given that a red card has been drawn, the probability of it being a four is calculated as follows:

Marginal Probability

The marginal probability, P(A), is the probability of an event A happening on its own. It does not consider the occurrence of any other event.

Since event A is drawing a four, P(A) is calculated by dividing the number of fours by the total number of cards in the deck.

Joint Probability

Joint probability is the likelihood of two or more events happening at the same time. This is denoted as P(A∩B), the probability of events A and B occurring.

Assuming that the previous events are the same, that is, event A is the occurrence of drawing a card that is a four and event B is drawing a red card, we can find the joint probability of drawing a card that is both a four and red.

There are two cards that meet both criteria, the four of hearts and the four of diamonds. Thus, the joint probability of drawing a card that is both a four and red is calculated as follows:

Discrete and continuous probability distributions

Discrete and continuous probability distributions describe the probabilities of different outcomes for random variables. Discrete distributions deal with variables that can only take on a finite number of distinct values, while continuous distributions handle variables that can take on any value within a range.

Discrete Probability Distributions:

- **Definition:**

A probability distribution where the random variable can only take on a finite number of distinct values, or a countably infinite number of values.

- **Examples:**
    - The number of heads when flipping a coin three times (can be 0, 1, 2, or 3).
    - The number of cars passing a certain point in an hour.
    - The number of customers visiting a restaurant in a day.
- **Key Features:**
    - Probabilities are assigned to each distinct value.
    - The sum of probabilities for all possible values is equal to 1.
    - Often represented using a probability mass function (PMF).

Continuous Probability Distributions:

- **Definition:**

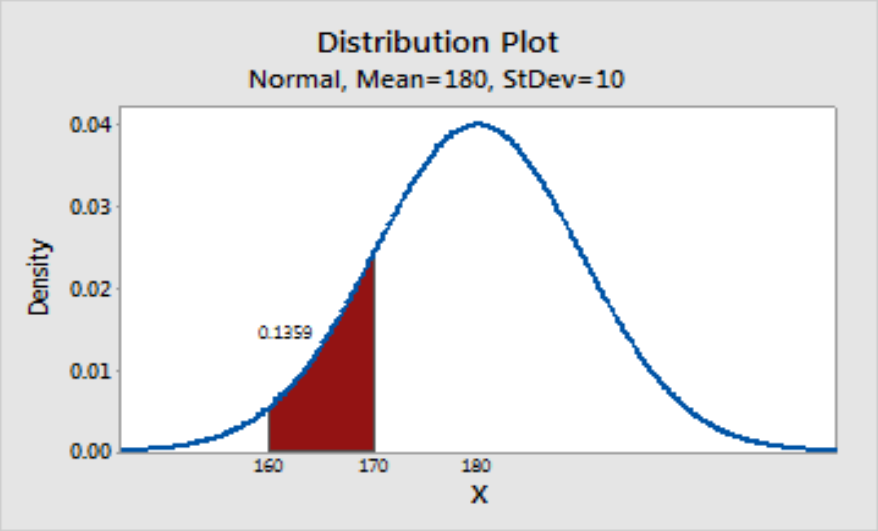
A probability distribution where the random variable can take on any value within a specified range.

- **Examples:**
    - The height of a person.
    - The temperature at a given location.
    - The amount of rainfall in a given region.

- **Key Features:**
  - The probability of a specific value is typically zero.
  - Probabilities are described using a probability density function (PDF), which represents the probability of the variable falling within a certain range.
  - Often represented by a curve or line on a graph.

Example of the distribution of weights

The **continuous normal distribution** can describe the distribution of weight of adult males. For example, you can calculate the probability that a man weighs between 160 and 170 pounds.



- **Distribution plot of the weight of adult males**
- The shaded region under the curve in this example represents the range from 160 and 170 pounds. The area of this range is 0.136; therefore, the probability that a randomly selected man weighs between 160 and 170 pounds is 13.6%. The entire area under the curve equals 1.0.
- However, the probability that X is exactly equal to some value is always zero because the area under the curve at a single point, which has no width, is zero. For example, the probability that a man weighs exactly 190 pounds to infinite precision is zero. You could calculate a nonzero probability that a man weighs more than 190 pounds, or less than 190 pounds, or between 189.9 and 190.1 pounds, but the probability that he weighs exactly 190 pounds is zero.

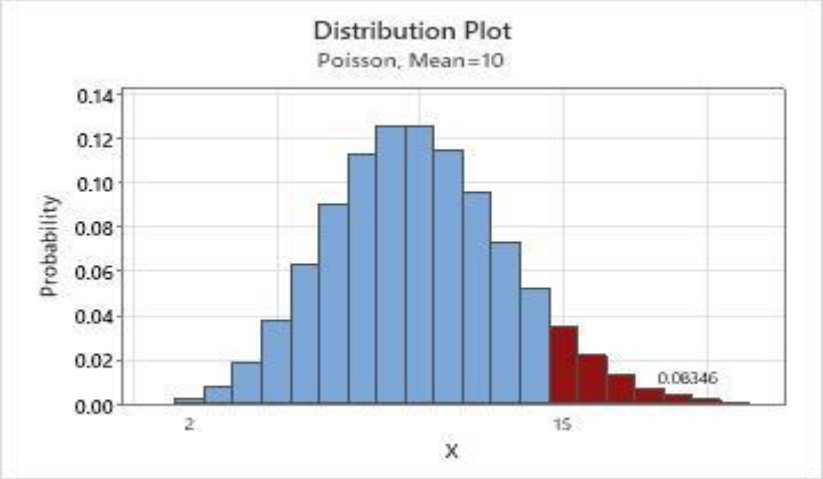
- A discrete distribution describes the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that has countable values, such as a list of non-negative integers.

  With a discrete probability distribution, each possible value of the discrete random variable can be associated with a non-zero probability. Thus, a discrete probability distribution is often presented in tabular form.

- **Example of the number of customer complaints**

  With a discrete distribution, unlike with a continuous distribution, you can calculate the probability that X is exactly equal to some value. For example, you can use the discrete Poisson distribution to describe the number of customer complaints within a day. Suppose the average number of complaints per day is 10 and you want to know the probability of receiving 5, 10, and 15 customer complaints in a day.

x	P (X = x)
5	0.037833
10	0.125110
15	0.034718



- You can also view a **discrete distribution** on a distribution plot to see the probabilities between ranges.
- **Distribution plot of the number of customer complaints**

  The shaded bars in this example represents the number of occurrences when the daily customer complaints is 15 or more. The height of the bars sums to 0.08346; therefore, the probability that the number of calls per day is 15 or more is 8.35%.

The difference between continuous distribution and discrete uniform distribution can be understood from the table given below.

Basis	Discrete Uniform Distribution	Continuous Distribution
Nature of Outcomes	Finite and countable set of outcomes	Infinite and uncountable range of outcomes
Probability Function	Probability Mass Function (PMF): P(X=x)= 1/n	Probability Density Function (PDF): f(x) = 1/(b-a)

Basis	Discrete Uniform Distribution	Continuous Distribution
Range of Values	Specific discrete values $x1$,$x2$,...,xn	Continuous range of values between a and b
Probability Calculation	Equal probability for each outcome: P(X=x)= 1/n	Equal density across the interval: f(x)= 1/(b-a)
Cumulative Distribution	CDF increases stepwise with each outcome and is defined by F(x) = P(X ≤ x).	CDF is a linear function within the interval defined by F(x) = (x – a) / (b – a) for a ≤ x ≤ b
Support	Specific values within a finite set	Continuous interval [a,b]
Real-World Application	Games of chance, like dice rolls or card draws	Random selection within a time interval, length measurement, etc.
Example	Rolling a fair six-sided die (outcomes: 1, 2, 3, 4, 5, 6)	Selecting a random point on a line segment from 1 to 10

The difference between continuous and discrete uniform distributions lies in their fundamental approach to representing data. Continuous uniform distributions encompass outcomes across a continuous range, ideal for scenarios where variables can take any value within a specified interval. On the other hand, discrete uniform distributions involve outcomes that are distinct and separate, suited for scenarios where variables can only take on a finite set of values with equal probability.

Random variable

A random variable is a variable whose value is a numerical outcome of a random phenomenon. It can be discrete or continuous, with discrete variables having a countable number of possible values and continuous variables having values within an interval.

Here's a more detailed explanation:

i) Definition:

A random variable is a function that maps each outcome of a random experiment to a real number. In simpler terms, it's a variable whose value is a numerical outcome of a random process.

ii) Types of Random Variables:

- **Discrete:**

These variables can only take on a countable number of values, often integers. Examples include the number of heads in a series of coin flips or the number of customers arriving at a store in an hour.

- **Continuous:**

These variables can take on any value within a specified range. Examples include height, weight, or temperature.

iii) Examples:

- **Discrete:**
    - The number of heads when flipping a coin three times.
    - The number of cars passing a certain point on a highway in an hour.
    - The number of defective items in a batch of 100 items.
- **Continuous:**
    - The height of a student.
    - The weight of a newborn baby.
    - The temperature of water in a bathtub.

iv) Key Concepts:

- **Probability Distribution:** A probability distribution describes the probabilities associated with the different possible values of a random variable.
- **Expected Value:** The expected value (or mean) of a random variable is the average value you would expect to get if you repeated the experiment many times.
- **Variance:** Variance measures the spread of the possible values of a random variable.

v)Importance:

Random variables are fundamental in statistics and probability because they provide a way to quantify and analyze random phenomena. They are used in a wide range of applications, including:

- **Risk assessment:** Estimating the potential losses or gains associated with a random event.
- **Data analysis:** Summarizing and interpreting data.
- **Scientific research:** Testing hypotheses about random phenomena.

Mathematical Expectations

The mathematical expectation, or expected value, of a random variable is a weighted average of all possible values the variable can take, where the weights are the probabilities of those values occurring. It essentially represents the long-term average of the random variable over many trials.

Key Concepts:

- **Discrete Random Variable:**

If X is a discrete random variable taking values x1, x2, ..., xn with probabilities p1, p2, ..., pn, the mathematical expectation E(X) is calculated as: E(X) = x1p1 + x2p2 + ... + xnpn.

- **Continuous Random Variable:**

If X is a continuous random variable with a probability density function f(x), the mathematical expectation E(X) is calculated as: $E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$.

- **Properties:**
    - **Linearity:** E(aX + b) = aE(X) + b, where a and b are constants.
    - **Sum of Independent Variables:** E(X + Y) = E(X) + E(Y), and E(XY) = E(X)E(Y) if X and Y are independent.
    - **Variance:** The variance of a random variable is the expected value of the square of the difference between the variable and its expected value: Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2.
- **Interpretation:**

The expected value provides a central tendency measure for a random variable's distribution and represents its long-term average.

The probability of the happening of the certain event is known as the probability of success (i.e. p) and the probability of non-happening of a certain even is known as the probability of failure (i.e. q). We always get p+q=1 }. The mathematical expectation is the events which are either happening or non-happening a certain event in the experiment. }. Probability of a non-happening event is zero, which is possible only if the numerator is 0. Probability of happening of a certain event is 1 which is possible only if the numerator and denominator are equal.

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

If x is a random variable which can assume any one of the values x1 , x2 , x3 , ….xn with respective probabilities p1 , p2 , p3 , ….pn then the mathematical expectation of x and denoted by E(x), is defined as: } E(x)= x1 p1+ x2 p2+….+ xnpn, Where ∑ pi = p1+ p2+…..+ pn=1

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

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 with 1/6. Thus, the expected value of the experiment will be **{1(1/6)+2(1/6)+3(1/6)+4(1/6)+5(1/6)+6(1/6)}=21/6= 3.5.**

If x is a discrete random variable and f(x) is the value of its probability distribution at x, the expected value of x is

E(X)pectation of a Discrete Random Variable

$$E(X) = \sum_x xP(x)$$

E(X)pectation of a Continuous Random Variable

$$E(X) = \int_{-\infty}^{\infty} xP(x)dx$$

Properties of Mathematical Expectations

- E(c) = c, where c is a constant
- E(cX) = c E(x), where c is a constant
- E(aX+b)=aE(X)+b, where a and b are constants
- **Addition rule of Mathematical expectation:** 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)
- **Multiplication rule of Mathematical expectation:** The mathematical expectation of the product of the two random variables will be the product of the mathematical expectation of those two variables, 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) Where X and Y are independent random variables

- The variance of a random variable is simple the expectation of its square minus the square of its expectation. Var(X) = E[X - E[X]] 2 = E[X 2] - E[X] 2 } if E[X] = 0 then Var[X] = E[X2].
- If a and b and are constants, then Var[aX]=a2Var[X] Var[X + b] = Var[X]
- The covariance in terms of expectations of between two independent variables is: Cov(X, Y) = E[(X - E[X])(Y - E[Y])]. It can be shown that Cov(X, Y) = E[XY]-E[X]E[Y].

Example 1:

If you tosses two unbiased coins. You will win Rs 10 if 2 heads appear, Rs 5 if one head appears and Rs 2 if no head appears. Find the expected value of the amount won by you.

Solution: In tossing two unbiased coins, the sample space, is

S = {HH, HT, TH, TT}. P[2 heads] = 1/4 , P (one head) = 2/4 , P(no head) = ¼ } Let x be the amount in rupees won by him. x can take the values 10, 5 and 2 with P[x = 10] = P(2heads) = 1/4 P[x = 5] = P[1Head] = 2/4 , and P[x = 2] = P[no Head] = 1/4

Probability distribution of x is x: 10 5 2 P(x): ¼, 2/4 ,1/4 Expected value of x is given as E(x)= x1 p1+ x2 p2+….+ xnpn,

E(x)=10(1/4)+5(2/4)+2(1/4)=10/4+10/4+2/4=22/4=5.5

Thus, the expected value of amount won by you is Rs 5.5.

Bayes' Theorem

Bayes' Theorem is a mathematical formula that helps determine the **conditional probability** of an event based on prior knowledge and new evidence.It adjusts probabilities when new information comes in and helps make better decisions in uncertain situations.

Bayes Theorem and Conditional Probability

Bayes' theorem (also known as the Bayes Rule or Bayes Law) is used to determine the conditional probability of event A when event B has already occurred.

The general statement of Bayes' theorem is "The conditional probability of an event A, given the occurrence of another event B, is equal to the product of the event of B, given A, and the probability of A divided by the probability of event B." i.e.

Bayes Theorem Formula

For any two events A and B, **Bayes's** formula for the Bayes theorem is given by:

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

- **P(A)** and **P(B)** are the probabilities of events A and B, also, P(B) is never equal to zero.
- **P(A|B)** is the probability of event A when event B happens,
- **P(B|A)** is the probability of event B when A happens.

Applications of Bayes' Theorem

- **Machine Learning:** Used for classification tasks, spam detection, and building predictive models.
- **Medical Diagnosis:** Helps in determining the probability of a disease given certain symptoms.
- **Risk Assessment:** Used to evaluate the likelihood of events based on prior knowledge.

- **Spam Filtering:** In spam detection, it helps determine the probability that an email is spam given certain words or phrases it contains.
- **Updating Beliefs:** Allows for adjusting initial beliefs or probabilities when new evidence becomes available.

The difference between Conditional Probability and Bayes's**.** The theorem can be understood with the help of the table given below.

Bayes Theorem	Conditional Probability
Bayes's Theorem is derived using the definition of conditional probability. It is used to find the reverse probability.	Conditional Probability is the probability of event A when event B has already occurred.
Formula: P(A|B) = [P(B|A)P(A)] / P(B)	**Formula:** P(A|B) = P(A∩B) / P(B)
Purpose: To update the probability of an event based on new evidence.	**Purpose**: To find the probability of one event based on the occurrence of another.
Focus: Uses prior knowledge and evidence to compute a revised probability.	**Focus**: Direct relationship between two events.

Binomial, Poisson, and Normal distributions

Binomial, Poisson, and Normal distributions are fundamental probability distributions used in statistics. The binomial distribution models the probability of a specific number of successes in a fixed number of independent trials. The Poisson

distribution models the probability of a certain number of events occurring within a fixed interval of time or space, given a known average rate. The normal distribution, also known as the bell curve, is a continuous distribution often used to model natural phenomena.

Binomial Distribution:

- **Discrete:** Deals with a fixed number of trials.
- **Two Outcomes:** Each trial results in one of two possibilities (success or failure).
- **Independent:** Trials are independent of each other.
- **Constant Probability:** The probability of success is the same for each trial.
- **Example:** Flipping a coin multiple times and counting the number of heads.
- **Parameters:** n (number of trials) and p (probability of success).

Poisson Distribution:

- **Discrete:** Deals with the number of events in a fixed interval.
- **Rare Events:** The probability of an event is very small.
- **Independent:** Events occur independently of each other.
- **Constant Rate:** Events occur at a constant average rate.
- **Example:** Counting the number of customers arriving at a store in an hour.
- **Parameter:** λ (average number of events in the interval).

Normal Distribution:

- **Continuous:** Deals with continuous data.
- **Bell-shaped:** Symmetrical around the mean.
- **Mean and Standard Deviation:** Determined by the mean (average) and standard deviation (spread).
- **Example:** Modeling human height or weight.
- **Parameters:** μ (mean) and σ (standard deviation).

Relationship between the Distributions:

- The binomial distribution can be approximated by the Poisson distribution when the number of trials (n) is large and the probability of success (p) is small, such that the product of n and p (np) is a moderate value.
- The normal distribution can be used to approximate the binomial distribution when n is large, and it can also be used to approximate the Poisson distribution when λ is large.

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

Example i): If a coin is tossed 5 times, using binomial distribution 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

$P(x=2) = {}^5C2\ p^2\ q^{5-2} = 5!\ /\ 2!\ 3! \times (½)^2 \times (½)^3$

P(x=2) = 5/16

(b) For at least four heads,

x ≥ 4, P(x ≥ 4) = P(x = 4) + P(x=5)

Hence,

$P(x = 4) = {}^5C4\ p^4\ q^{5-4} = 5!/4!\ 1! \times (½)^4 \times (½)^1 = 5/32$

$P(x = 5) = {}^5C5\ p^5\ q^{5-5} = (½)^5 = 1/32$

Answer: Therefore, P(x ≥ 4) = 5/32 + 1/32 = 6/32 = 3/16

Example ii): For the same question given above, find the probability of getting at most 2 heads.

Solution:

Solution: P(at most 2 heads) = P(X ≤ 2) = P (X = 0) + P (X = 1)

$P(X = 0) = (½)^5 = 1/32$

$P(X=1) = 5C1\ (½)^5.= 5/32$

Answer: Therefore, P(X ≤ 2) = 1/32 + 5/32 = 3/16

Poisson Distribution Formula

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

where

x = 0, 1, 2, 3, ...

λ = mean number of occurrences in the interval

e = Euler's constant ≈ 2.71828

Example i): In a café shop, the customer arrives at a mean rate of 2 per min. Find the probability of arrival of 5 customers in 1 minute using the Poisson distribution formula.

Solution:

Given: $\lambda = 2$, and $x = 5$.

Using the Poisson distribution formula:

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

$$P(X = 5) = (e^{-2} 2^5)/5!$$

$$P(X = 6) = 0.036$$

Answer: The probability of arrival of 5 customers per minute is 3.6%.

Example 2: Find the mass probability of function at $x = 6$, if the value of the mean is 3.4.

Solution:

Given: $\lambda = 3.4$, and $x = 6$.

Using the Poisson distribution formula:

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

$$P(X = 6) = (e^{-3.4} 3.4^6)/6!$$

$$P(X = 6) = 0.072$$

Answer: The probability of function is 7.2%.

Normal Distribution Formula

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

$\mu =$ mean of x
$\sigma =$ standard deviation of x
$\pi \approx 3.14159 \ldots$
$e \approx 2.71828 \ldots$

Example i): Find the probability density function of the normal distribution of the following data. x = 2, μ = 3 and σ = 4.

Solution:

Given,

- *Variable (x) = 2*
- *Mean = 3*
- *Standard Deviation = 4*

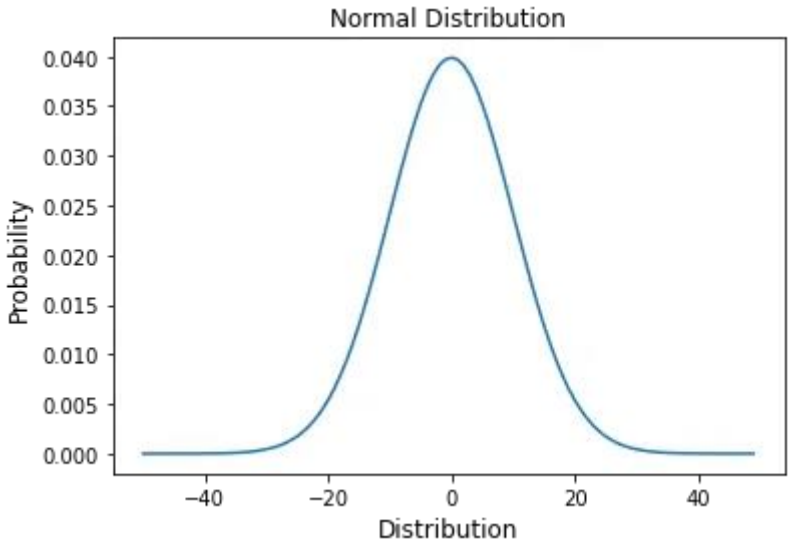
Using formula of probability density of normal distribution

Simplifying,

f(2, 3, 4) = 0.09666703

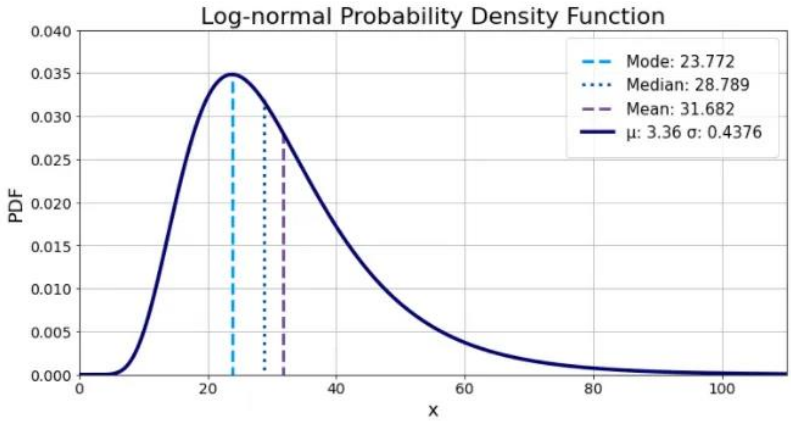
Normal (Gaussian) distribution:Images

Normal distribution is one of the most common distributions. As the name suggests, most situations behave according to a normal distribution, which is why it is termed as such. Its widespread use is due, in part, to the fact that the huge sum of (small) random variables frequently turns out to be regularly distributed.



Binomial distribution:

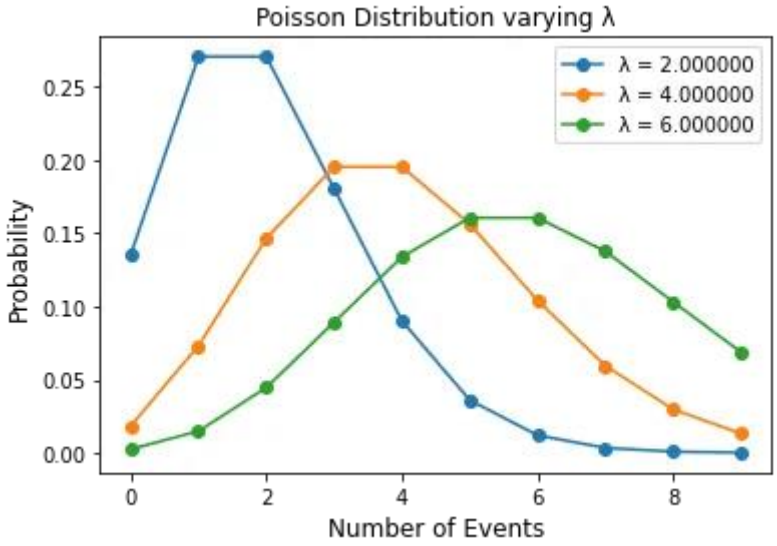
When an experiment or survey has exactly two possible outcomes (e.g., success or failure, True or False, 1 or 0), it has a binomial distribution. The prefix "bi" implies "two" or "twice," and, therefore, the term "binomial" refers to a distribution type with two probable outcomes. A coin flip, for instance, can only result in one of two outcomes — heads or tails, and the results of a test might result in either a passing or failing grade.



The Bernoulli distribution and the binomial distribution share many similarities. The number of successes in Bernoulli trials has a binomial distribution if each trial is independent. The Bernoulli distribution, on the other hand, is the Binomial distribution with n=1.

Poisson distribution:

A Poisson distribution refers to a discrete frequency distribution that gives the probability of a number of independent events occurring within a fixed time. In other words, Poisson distributions can be used to forecast the frequently of independent events over a specific time span.

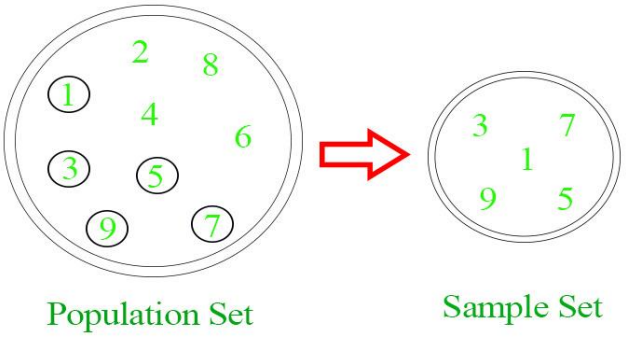


Probability distribution plays a very critical role in machine learning model building. Probability distribution models can help us compute each outcome's probability, long-term average outcomes, as well as estimate the variability in results of random variables without having to know all of the actual results of the random variables we're interested in.

Unit-II

Sampling Theory

Sampling is a statistical technique for efficiently analyzing large datasets by selecting a representative subset. Rather than analyzing an entire dataset, sampling analyzes a small portion so researchers can make conclusions about a larger population. This allows for informed decision-making without exhaustive data collection.



Population is the whole set of variables, elements, entities which are considered for a statistical study. It is also known as the universal set from where actual inferences are drawn. Population set consists of all the attributes of individuals or elements under consideration, but doing estimations on a Population is very exhausting resources as well as time-wise alike. *Example*: Consider the mean weight of all men on Earth. This here, is considered a hypothetical population because it includes all men that have ever lived on earth which includes people who will exist in the future and also people who have lived earlier before us. But there comes an anomaly, while doing such measurement which is not all men in the population tray are observable (consider men, who will exist in the future and also men, who have lived before and doesn't exist right now). Also, performing statistics on the population sample (if hypothetically possible) would require a great deal of time as well as resources, which will be exhaustive and inefficient as well. Thus what is perform instead is to take a subset from the available population and perform statistics on them and

interpolate inferences about the entire population. Taking out a subset, makes the task easier as the time required to scrutinize the subset is lesser than the time required to scrutinize the whole set of Population. Statistics is performed on the sample set to draw conclusions about the entire population tray. Calculations are considered to be a conclusion of the population set because it doesn't measure with the actual data of the population set and is not free from errors. This is obvious as sample set is used as a medium frame, having fewer members and thus some information is lost. (which results in errors).

The sampling process can be done by conducting the following steps:

1. **Define the population:** Identify the group from which the sample will be taken, such as customers, transactions, or employees.
2. **Choose a sampling method:** Different methods can be used depending on the study's objectives. For example, random sampling focuses on fairness, whereas systematic sampling uses regular intervals.
3. **Determine the sample size:** The sample size needs to be manageable and large enough to provide reliable results.
4. **Collect data from the sample:** Collecting data can be done in various ways depending on the study, such as surveys, interviews, or records.
5. **Analyze and interpret the data:** Once the data is retrieved, it needs to be interpreted using statistical tools and methods to come up with conclusions.

Types of Sampling

Different sampling techniques can be used in different scenarios, depending on the parameters and goals of the study. The different methods of sampling are as follows:

Random Sampling

Random sampling is often used in surveys and market research, ensuring that every constituent of a population has an equal chance of being selected. For example, a bank might randomly choose 1,000 customers to assess spending

habits. Randomly choosing these customers helps reduce bias and is great for getting general results.

Stratified Sampling

Stratified sampling breaks a population into different subgroups (strata) based on specific, shared characteristics. Samples are then chosen from each group. Stratified sampling is best used when a population is diverse.

For example, if a company wanted to determine employee satisfaction, it would make no sense to randomly choose employees because every job function is different, which will directly impact job satisfaction. Stratified sampling would first divide workers based on department and then pick samples from each. This ensures that the subgroups are properly represented.

Cluster Sampling

Cluster sampling selects entire groups rather than individuals. For example, a consulting agency evaluating a bank's different branch performances would select entire branches rather than the individuals in each branch.

While cluster sampling might sound similar to stratified sampling, there are differences. Cluster sampling randomly selects entire groups while stratified sampling selects a few individuals from all groups.

The individuals in cluster groups are different whereas the individuals in stratified sampling subgroups are the same due to a shared characteristic. The primary goal of cluster sampling is to make it easier to gather data.

Systematic Sampling

In systematic sampling, every nth item is chosen from a population at regular intervals. For example, if a company wanted to analyze 2,000 invoices from a total of 20,000, it would choose every 10th invoice for review after a random starting point.

Systematic sampling is similar to random sampling but is more structured and ensures even coverage of the larger population. However, systematic patterns in the data could lead to unintended bias. For example, if a retail company selects every

seventh day and that consistently falls on a weekend, it may overrepresent high sales days, skewing the results.

Convenience Sampling

Convenience sampling is just that, convenient. It is cost-effective but has a high chance of introducing bias and may not truly be representative. For example, if a retail store only selects customers who come in during lunch, they miss out on the insights from evaluating morning and evening shoppers.

Importance of Sampling in Business and Finance

As discussed, sampling has many uses. In business in finance, it is applied in varied ways:

- **Market research:** Companies use sampling to understand consumer preferences and predict demand. By analyzing a sample of their target audiences, companies can determine product fit, gauge the interest in new items, and refine marketing strategies.
- **Financial auditing:** Auditors perform a detailed analysis of company financials and transactions. They can choose transaction samples to identify errors and fraud without having to check every single company transaction. A sample would still allow auditors to identify inconsistencies or patterns of inaccurate reporting.
- **Quality control in manufacturing:** To ensure product quality, manufacturers use sampling without having to inspect every item produced. If defects are found in the sample, fixes can be made before the entire batch is sent out. This helps ensure customer satisfaction and avoid costly recalls.

Sampling distribution

Sampling distribution is essential in various aspects of real life. Sampling distributions are important for inferential statistics. A sampling distribution represents the distribution of a statistic, like the mean or standard deviation, which is calculated from multiple samples of a population. It shows how these statistics vary across different samples drawn from the same population.

Sampling distribution is also known as a **finite-sample distribution**. Sampling distribution is the probability distribution of a statistic based on random samples of a given population. It represents the distribution of frequencies on how spread apart various outcomes will be for a specific population.

Since population is too large to analyze, you can select a smaller group and repeatedly sample or analyze them. The gathered data, or statistic, is used to calculate the likely occurrence, or probability, of an event.

Types of Distributions

There are 3 main types of sampling distributions are:

- Sampling Distribution of Mean
- Sampling Distribution of Proportion
- T-Distribution

Sampling Distribution of Mean

Mean is the most common type of sampling distribution.

It focuses on calculating the mean or rather the average of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

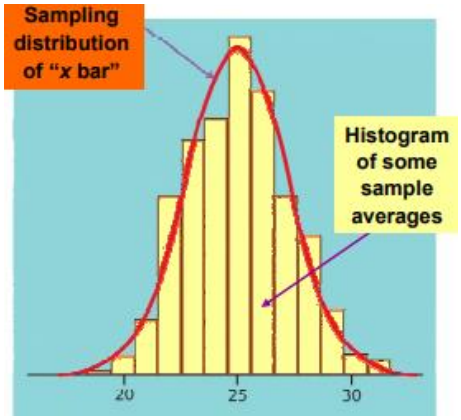
We take many random samples of a given size n from a population with mean μ and standard deviation σ. Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.

For any population with mean μ and standard deviation σ:

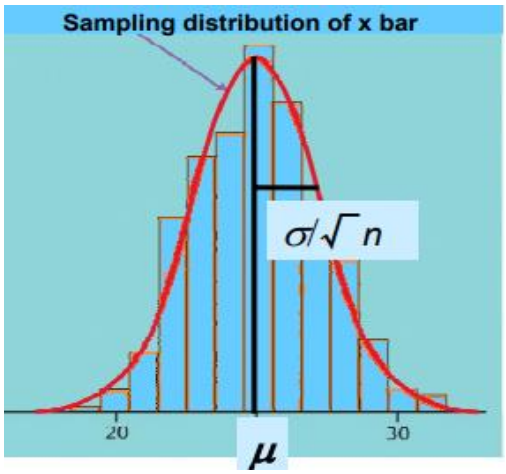
- Mean, or center of the sampling distribution of $\bar{x}$, is equal to the population mean, μ.

$$\mu_{\bar{x}} = \mu$$

There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution is an unbiased estimate of the population mean μ.

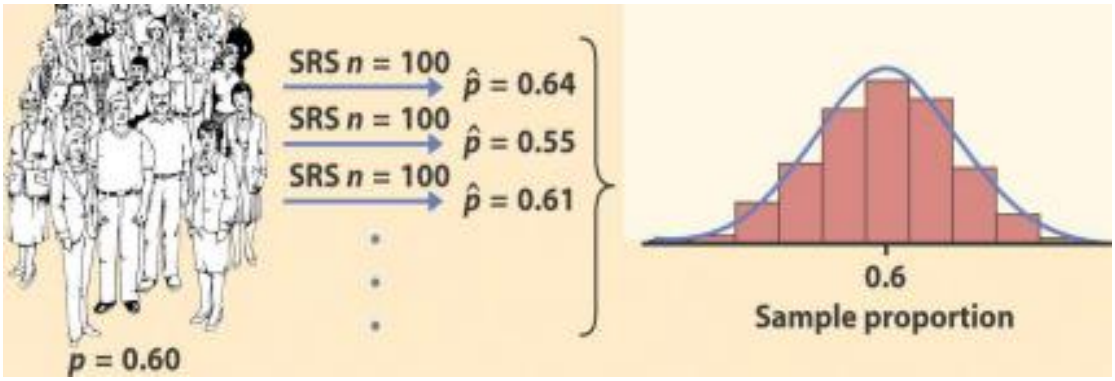


Standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of √n. Averages are less variable than individual observations.



Sampling Distribution of Proportion

Sampling distribution of proportion focuses on proportions in a population. Here, you select samples and calculate their corresponding proportions. The means of the sample proportions from each group represent the proportion of the entire population.

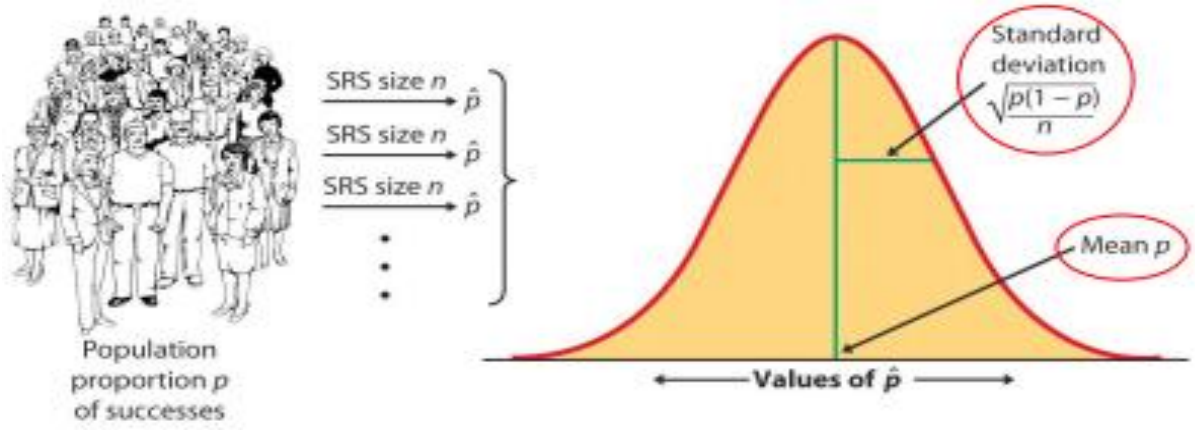


ormula for the sampling distribution of a proportion (often denoted as p̂) is: $\hat{p} = x/n$

where:

- **p̂** is Sample Proportion
- **x** is Number of "successes" or occurrences of Event of Interest in Sample
- **n** is Sample Size

This formula calculates the proportion of occurrences of a certain event (e.g., success, positive outcome) within a sample.



T-Distribution

Sampling distribution involves a small population or a population about which you don't know much. It is used to estimate the mean of the population and other statistics such as confidence intervals, statistical differences and linear regression. T-

distribution uses a t-score to evaluate data that wouldn't be appropriate for a normal distribution.

Formula for the t-score, denoted as t, is: $t = [x - \mu] / [s /\sqrt{(n)}]$

where:

- **x** is Sample Mean
- **μ** is Population Mean (or an estimate of it)
- **s** is Sample Standard Deviation
- **n** is Sample Size

This formula calculates the difference between the sample mean and the population mean, scaled by the standard error of the sample mean. The t-score helps to assess whether the observed difference between the sample and population means is statistically significant.

Z-test

According to the central limit theorem, if X1, X2, ..., Xn is a random sample of size n taken from a population with mean μ and variance σ2 then the sampling distribution of the sample mean tends to normal distribution with mean μ and variance σ2/n as sample size tends to large.

This formula indicates that as the sample size increases, the spread of the sample means around the population mean decreases, with the standard deviation of the sample means shrinking proportionally to the square root of the sample size, and the variate Z,

$$Z = (x - \mu)/(\sigma/\sqrt{n})$$

where,

- **z** is z-score
- **x** is Value being Standardized (either an individual data point or the sample mean)

- **μ** is Population Mean
- **σ** is Population Standard Deviation
- **n** is Sample Size

This formula quantifies how many standard deviations a data point (or sample mean) is away from the population mean. Positive z-scores indicate values above the mean, while negative z-scores indicate values below the mean. Follows the normal distribution with mean 0 and variance unity, that is, the variate Z follows standard normal distribution.

According to the central limit theorem, the sampling distribution of the sample means tends to normal distribution as sample size tends to large (n > 30).

The sampling distribution is the mechanism that helps us to make data-driven insights about a large population using a manageable subset of that population - our sample. Whether we're investigating means, proportions, variances, or differences therein, sampling distributions guide us, helping turn seemingly abstract numbers into meaningful knowledge. While sampling distribution might seem like a challenging concept to get your head around at first, with time and practice, it reveals itself as an indispensable ally in your data analysis toolkit. This comprehension transforms intimidating arrays of data into a rich tapestry of insights, fuelling confident decision-making based on robust statistical evidence.

Parameter vs Statistic

A **parameter** is a number describing a whole population (e.g., population mean), while a **statistic** is a number describing a sample (e.g., sample mean).

The goal of quantitative research is to understand characteristics of populations by finding parameters. In practice, it's often too difficult, time-consuming or unfeasible to collect data from every member of a population. Instead, data is collected from samples.

With inferential statistics, we can use sample statistics to make educated guesses about population parameters.

Population vs sample

In research, a **population** is the entire group that you're interested in studying. This may be a group of people (e.g., all adults in the US or all employees of a company), but it can also mean a group containing other kinds of elements: objects, events, organizations, countries, species, organisms, etc.

A **sample** is a smaller group taken from the population. The sample is the group of elements that you will actually collect data from

Common examples of parameters include:

- **Population Mean:** We use population mean when we want to find the average of the population. Population mean is denoted by μ.
- **Population standard deviation:** It is denoted by σ. It is used to find the dispersion of data in the population.
- **Population proportion:** We use p to denote population proportion. It is a parameter that is used to calculate percentage of the data points that follow some specific pattern.

Statistics

Statistics is a numerical quantity that is derived from a sample. Since sample vary in nature, the Statistics are also variable in nature. Since statistical values refer to the sample, using these values, we can derive the parameter values.

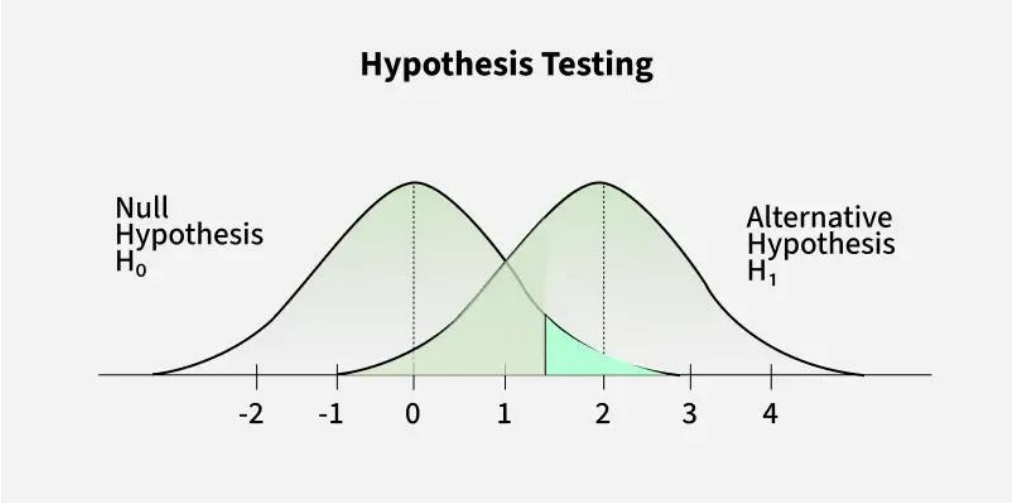
	Parameter	**Statistics**
Definition	A numerical characteristic that describes an entire population.	A numerical measure calculated from a sample drawn from the population.
Nature	Fixed and constant	Variable and can change from sample to sample
Value	Unknown (as it pertains to the entire population)	Known (as it is derived from the sample data)

Representation	Represents the entire population	Represents a sample of the population
Usage	Used in theoretical concepts and hypothesis formulation	Used in practical data analysis and research
Accuracy	More accurate as it represents the entire population	It may vary in accuracy as it depends on the sample size and selection
Calculation	It cannot be calculated exactly in most real-world scenarios	Can be calculated using sample data
Purpose	To describe the characteristics of the population	To estimate the population parameters based on sample data
Example	The actual average height of all adults in a country	The average height calculated from a sample group of adults in a country

- **Sample mean ($\bar{x}$):** is used to find the average of the sample.
- **Sample standard deviation:** It is denoted by s. This is used to find the standard deviation of the sample data.
- **Sample proportion ($\hat{p}$):** Often read as p-hat, this is used to find the fraction of the data points that possess some characteristic.

Hypothesis testing

Hypothesis testing compares two opposite ideas about a group of people or things and uses data from a small part of that group (a sample) to decide which idea is more likely true. We collect and study the sample data to check if the claim is correct.



For example, if a company says its website gets 50 visitors each day on average, we use hypothesis testing to look at past visitor data and see if this claim is true or if the actual number is different.

Defining Hypotheses

- **Null Hypothesis (H_0):** The starting assumption. For example, "The average visits are 50."
- **Alternative Hypothesis (H_1):** The opposite, saying there is a difference. For example, "The average visits are not 50."

Key Terms of Hypothesis Testing

To understand the Hypothesis testing firstly we need to understand the key terms which are given below:

- **Significance Level (α):** How sure we want to be before saying the claim is false. Usually, we choose 0.05 (5%).
- **p-value:** The chance of seeing the data if the null hypothesis is true. If this is less than α, we say the claim is probably false.
- **Test Statistic:** A number that helps us decide if the data supports or rejects the claim.
- **Critical Value:** The cutoff point to compare with the test statistic.

- **Degrees of freedom:** A number that depends on the data size and helps find the critical value.

Types of Hypothesis Testing

It involves basically two types of testing:

One-Tailed Test

Used when we expect a change in only one direction either up or down, but not both. For example, if testing whether a new algorithm improves accuracy, we only check if accuracy increases.

There are two types of one-tailed test:

- **Left-Tailed (Left-Sided) Test:** Checks if the value is less than expected. Example: H0:μ≥50μ≥50 and H1: μ<50μ<50
- **Right-Tailed (Right-Sided) Test**: Checks if the value is greater than expected. Example: H0 : μ≤50μ≤50 and H1:μ>50μ>50

2. Two-Tailed Test

Used when we want to see if there is a difference in either direction higher or lower. For example, testing if a marketing strategy affects sales, whether it goes up or down

Example: H0: μ=μ= 50 and H1: μ≠50μ□=50

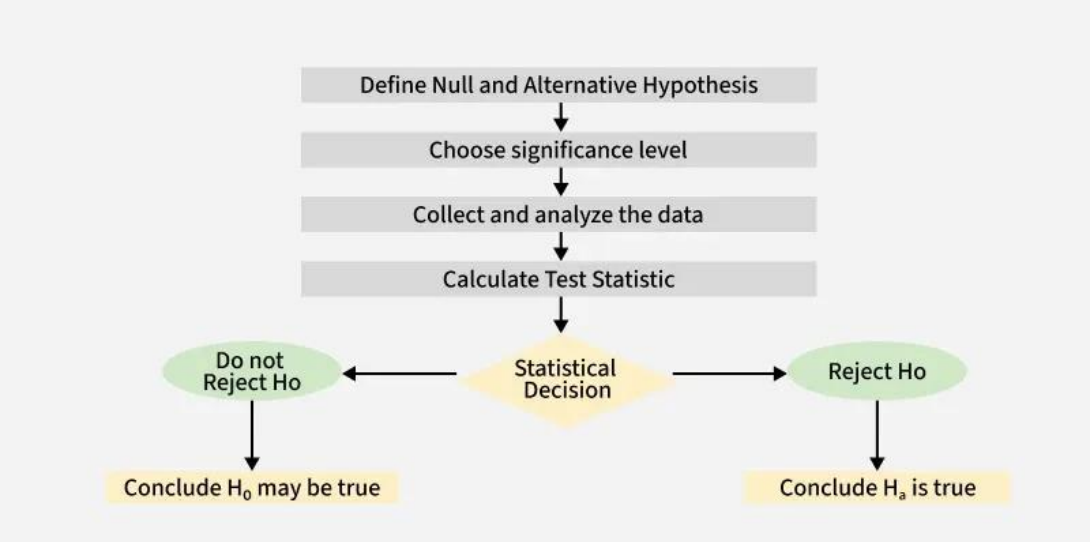
Type 1 and Type 2 errors in Hypothesis Testing

In hypothesis testing Type I and Type II errors are two possible errors that can happen when we are finding conclusions about a population based on a sample of data. These errors are associated with the decisions we made regarding the null hypothesis and the alternative hypothesis.

- **Type I error:** When we reject the null hypothesis although that hypothesis was true. Type I error is denoted by alpha(αα).
- **Type II errors:** When we accept the null hypothesis but it is false. Type II errors are denoted by beta(β).

	Null hypothesis is TRUE	Null hypothesis 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)

Working of Hypothesis testing involves various steps:



Step 1: Define Hypotheses:

- **Null hypothesis (H_0):** Assumes no effect or difference.
- **Alternative hypothesis (H_1):** Assumes there is an effect or difference.

Example: Test if a new algorithm improves user engagement.

*Note: In this we assume that our data is **normally distributed**.*

Step 2: Choose significance level

We select a significance level (usually 0.05). This is the maximum chance we accept of wrongly rejecting the null hypothesis (Type I error). It also sets the confidence needed to accept results.

Step 3: Collect and Analyze data.

- Now we gather data this could come from user observations or an experiment. Once collected we analyze the data using appropriate statistical methods to calculate the **test statistic**.
- **Example**: We collect data on user engagement before and after implementing the algorithm. We can also find the mean engagement scores for each group.

Step 4: Calculate Test Statistic

The test statistic measures how much the sample data deviates from what we did expect if the null hypothesis were true. Different tests use different statistics:

- **Z-test:** Used when population variance is known and sample size is large.
- **T-test:** Used when sample size is small or population variance unknown.
- **Chi-square test:** Used for categorical data to compare observed vs. expected counts.

Step 5: Make a Decision

We compare the test statistic to a critical value from a statistical table or use the p-value:

- **Using Critical Value:**
  - If test statistic > critical value → reject H0.
  - If test statistic ≤ critical value → fail to reject H0.
- **Using P-value:**
  - If p-value ≤ α → reject H0.
  - If p-value > α → fail to reject H0.

Example: If p-value is 0.03 and α is 0.05, we reject the null hypothesis because 0.03 < 0.05.

Step 6: Interpret the Results

Based on the decision, we conclude whether there is enough evidence to support the alternative hypothesis or if we should keep the null hypothesis.

Limitations of Hypothesis Testing

Although hypothesis testing is a useful technique but it have some limitations as well:

- **Limited Scope**: Hypothesis testing focuses on specific questions or assumptions and not capture the complexity of the problem being studied.
- **Data Quality Dependence**: The accuracy of the results depends on the quality of the data. Poor-quality or inaccurate data can led to incorrect conclusions.
- **Missed Patterns**: By focusing only on testing specific hypotheses important patterns or relationships in the data might be missed.
- **Context Limitations**: It doesn't always consider the bigger picture which can oversimplify results and led to incomplete insights.

- **Need for Additional Methods**: To get a better understanding of the data hypothesis testing should be combined with other analytical methods such as data visualization or machine learning techniques which we study later in upcoming articles.

Standard Error

The standard error of the mean, or simply **standard error**, indicates how different the population mean is likely to be from a sample mean. It tells you how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

The standard error of the mean (SE or SEM) is the most commonly reported type of standard error. But you can also find the standard error for other statistics,

like medians or proportions. The standard error is a common measure of sampling error—the difference between a population parameter and a sample statistic.

Why standard error matters

In statistics, data from samples is used to understand larger populations. Standard error matters because it helps you estimate how well your sample data represents the whole population.

With probability sampling, where elements of a sample are randomly selected, you can collect data that is likely to be representative of the population. However, even with probability samples, some sampling error will remain. That's because a sample will never perfectly match the population it comes from in terms of measures like means and standard deviations.

By calculating standard error, you can estimate how representative your sample is of your population and make valid conclusions.

A high standard error shows that sample means are widely spread around the population mean—your sample may not closely represent your population. A low standard error shows that sample means are closely distributed around the population mean—your sample is representative of your population.

You can decrease standard error by increasing sample size. Using a large, random sample is the best way to minimize sampling bias.

Standard error formula

The standard error of the mean is calculated using the standard deviation and the sample size.

From the formula, you'll see that the sample size is inversely proportional to the standard error. This means that the larger the sample, the smaller the standard error, because the sample statistic will be closer to approaching the population parameter.

Different formulas are used depending on whether the population standard deviation is known. These formulas work for samples with more than 20 elements ($n > 20$).

When population parameters are known

When the population standard deviation is known, you can use it in the below formula to calculate standard error precisely.

Formula	Explanation
$SE = \frac{\sigma}{\sqrt{n}}$	• SE is standard error • σ is population standard deviation • n is the number of elements in the sample

When population parameters are unknown

When the population standard deviation is unknown, you can use the below formula to only estimate standard error. This formula takes the sample standard deviation as a point estimate for the population standard deviation.

Formula	Explanation
$SE = \frac{s}{\sqrt{n}}$	• SE is standard error • s is sample standard deviation • n is the number of elements in the sample

Importance of Standard Error

When a sample of observations is extracted from a population and the sample mean is calculated, it serves as an estimate of the population mean. Almost certainly, the sample mean will vary from the actual population mean. It will aid the statistician's research to identify the extent of the variation. It is where the standard error of the mean comes into play.

When several random samples are extracted from a population, the standard error of the mean is essentially the standard deviation of different sample means from the population mean.

However, multiple samples may not always be available to the statistician. Fortunately, the standard error of the mean can be calculated from a single sample

itself. It is calculated by dividing the standard deviation of the observations in the sample by the square root of the sample size.

The standard error is not the only measure of dispersion and accuracy of the sample statistic. It is, however, an important indicator of how reliable an estimate of the population parameter the sample statistic is. Taken together with such measures as effect size, p-value and sample size, the effect size can be a very useful tool to the researcher who seeks to understand the reliability and accuracy of statistics calculated on random samples.

Properties of Estimators

In statistics, an **estimator** is a rule for calculating an estimate of a given quantity based on observed data: thus the rule (the estimator), the quantity of interest (the estimand) and its result (the estimate) are distinguished. For example, the sample mean is a commonly used estimator of the population mean.

A good estimator in statistics should be unbiased, consistent, efficient, and sufficient. Unbiasedness means the estimator's expected value equals the true parameter being estimated. Consistency means the estimator approaches the true parameter as the sample size increases. Efficiency refers to having the smallest variance among competing estimators. Sufficiency means the estimator captures all the relevant information about the parameter from the sample.

- **Unbiasedness:**

An estimator is unbiased if its expected value equals the true value of the parameter it's estimating. In simpler terms, on average, the estimator's estimates don't systematically overestimate or underestimate the true parameter.

- **Consistency:**

A consistent estimator's estimates converge to the true parameter as the sample size increases. This means that with more data, the estimator becomes increasingly accurate.

- **Efficiency:**

Efficiency refers to the variance of the estimator. An efficient estimator has the smallest variance among all unbiased estimators. A smaller variance implies that the estimator's estimates are more tightly clustered around the true parameter.

- **Sufficiency:**

A sufficient estimator utilizes all the relevant information about the parameter contained in the sample. This means that no additional information is gained by considering other statistics or functions of the sample data.

Unit III

Test of Significance Large and Small Sample

In statistics, the primary difference between large and small samples lies in the number of observations included. A large sample typically has 30 or more observations, while a small sample has fewer than 30. This distinction impacts the statistical methods used for analysis and the interpretation of results. Large samples generally provide more reliable estimates and increase the power of statistical tests, while small samples may require different statistical approaches and can lead to less precise results.

Large Samples (n ≥ 30):

- **Accuracy and Precision:**

Larger samples tend to provide more accurate and precise estimates of population parameters (like the mean or proportion). The margin of error is generally smaller, meaning the sample results are likely closer to the true population value.

- **Normal Distribution:**

As sample size increases, the sampling distribution of many statistics (like the sample mean) approaches a normal distribution, even if the underlying population distribution is not normal. This allows for the use of statistical tests based on the normal distribution (like the z-test).

- **Statistical Power:**

Larger samples increase the statistical power of a study, making it easier to detect true differences or relationships between variables.

- **Generalizability:**

Results from larger samples are often more generalizable to the larger population from which the sample was drawn.

Small Samples (n < 30):

- **Less Precision:**

Small samples may lead to less precise estimates and a wider margin of error.

- **Non-Normal Distributions:**

If the population is not normally distributed, and the sample size is small, the sampling distribution may not be normal.

- **Different Statistical Tests:**

For small samples, statistical tests that assume a test, which are designed for smaller sample sizes, are used.

- **Increased Risk of Error:**

Small samples can be more susceptible to random variation, leading to results that may not accurately reflect the population.

In summary, significance tests for large samples, particularly the Z-test, are powerful tools for analyzing data and drawing meaningful conclusions about populations based on sample observations.

Differences between two means and standard deviations,

To test for differences between two means and standard deviations, you can use a combination of t-tests and F-tests. The t-test is used to compare means, while the F-test is used to compare variances (and thus standard deviations).

Testing for Differences in Means (t-test):

- Independent Samples t-test: This test is used when comparing the means of two independent groups. The formula for the test statistic depends on whether the population variances are assumed to be equal or unequal.
    - Equal Variances (Pooled t-test): If the variances are assumed to be equal, a pooled variance is calculated to estimate the common population variance. The t-statistic is calculated as:

formula $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s_p^2 * (1/n_1 + 1/n_2))}$

where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means

- s_p^2 is the pooled variance
- n_1 and n_2 are the sample sizes
- Unequal Variances (Welch's t-test): If the variances are not assumed to be equal, a more complex formula for the t-statistic is used, and the degrees of freedom are also adjusted.
- Paired t-test: This test is used when comparing the means of two related groups (e.g., before and after measurements on the same individuals). The test focuses on the differences between the paired observations. The t-statistic is calculated as:

Code

t = d̄ / (s_d / √n)

where:

- d̄ is the mean of the differences
- s_d is the standard deviation of the differences
- n is the number of pairs

2. Testing for Differences in Standard Deviations (F-test):

- F-test for Equality of Variances: This test compares the variances (or standard deviations) of two populations. The test statistic is calculated as the ratio of the two sample variances:

formula $F = s_1^2 / s_2^2$

where s_1^2 and s_2^2 are the sample variances. The larger variance is typically placed in the numerator.

Hypotheses:

The null hypothesis (H_0) is that the population variances are equal ($\sigma_1^2 = \sigma_2^2$), and the alternative hypothesis (H_a) is that they are not equal ($\sigma_1^2 \neq \sigma_2^2$), or one is greater than the other ($\sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$).

Interpretation:

The calculated F-statistic is compared to a critical value from the F-distribution (based on the chosen significance level and degrees of freedom). If the calculated F-statistic exceeds the critical value, the null hypothesis is rejected, indicating a significant difference in the variances.

Proportion and Confidence Interval

A confidence interval for a proportion provides a range of values within which the true population proportion is likely to fall, given a certain level of confidence. It's a crucial concept in statistics, helping to estimate population characteristics from sample data with a measure of uncertainty.

Key Concepts:

- **Population Proportion (p):**

This is the unknown proportion of a population that possesses a specific characteristic. For example, the proportion of all registered voters who plan to vote for a particular candidate.

- **Sample Proportion (p̂):**

This is the proportion of a sample (a subset of the population) that exhibits the characteristic of interest. For instance, if you survey 100 voters and 60 say they will vote for the candidate, the sample proportion is 0.60.

- **Confidence Level:**

This indicates the probability that the confidence interval will contain the true population proportion. Common confidence levels are 90%, 95%, and 99%.

- **Margin of Error:**

This quantifies the uncertainty in the sample estimate and is added and subtracted from the sample proportion to create the confidence interval.

Formula and Calculation:

The confidence interval for a population proportion is calculated as:

$\hat{p} \pm z^* * \sqrt{((\hat{p}(1-\hat{p}))/n)}$

Where:

- **p̂**: is the sample proportion.
- **z* is the critical value**: from the standard normal distribution corresponding to the desired confidence level (e.g., 1.96 for 95% confidence).
- **n**: is the sample size.
- **√((p̂(1-p̂))/n)**: is the standard error of the proportion.

Interpretation:

A 95% confidence interval for the proportion of voters who will vote for a candidate means that if you were to repeat the sampling process many times, 95% of the resulting confidence intervals would contain the true population proportion of voters.

Example:

If a poll of 1000 voters shows 52% support a particular candidate, with a 95% confidence level, the margin of error might be 3%. The confidence interval would be: 0.52 ± 0.03, or 0.49 to 0.55. This means the pollster is 95% confident that the true proportion of voters who support the candidate is between 49% and 55%.

Important Considerations:

- The confidence interval is only an estimate, and there's always a chance it doesn't contain the true population proportion.
- The wider the confidence interval, the less precise the estimate, and the lower the confidence level, the wider the interval.
- Sample size plays a crucial role. Larger sample sizes lead to narrower confidence intervals and more precise estimates.

SMALL SAMPLE TESTS

Student's t

Let $x_1, x_2,...x_n$ be a random sample of size n from a normal population with mean μ and variance σ^2 .

The student's t test is defined in the statistics as

$$t = \frac{\bar{x} - \mu}{\left(\frac{S}{\sqrt{n}}\right)}, \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ , is the}$$

sample mean and $S^2 = \frac{1}{n-1}\sum_{n=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimate of the

population variance σ^2 and t follows student's t-distribution with v = n – 1 degrees of freedom with probability density function

Total area under the curve is 1.

3. t-distribution is symmetrical about t = 0 and has a mean zero.

4. The variance of t-distribution is greater than 1, but tends to 1 as $n \to \infty$. As $v \to \infty$, t - distribution becomes normal.

5. Variance = v / v – 2 if v > 2 and $\mu_2 > 1$ always.

Working Procedure :

For the small samples (n < 30), σ known, decision is based on the t-distribution with v = n - 1 degrees of freedom. di be bellso a

1. Null hypothesis $H_0 : \mu = \mu_0$

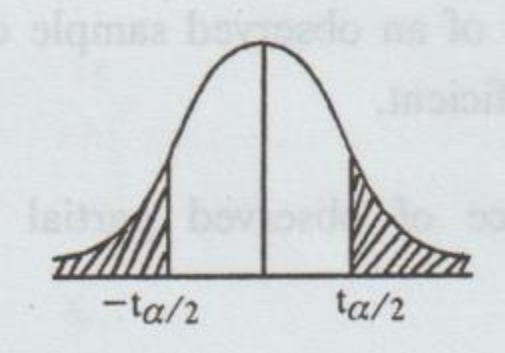
2. Alternative hypothesis $H_1: \mu \neq \mu_0$ (or) $\mu > \mu_0$ (or) $\mu < \mu_0$

3. Level of significance: α

4. Critical region

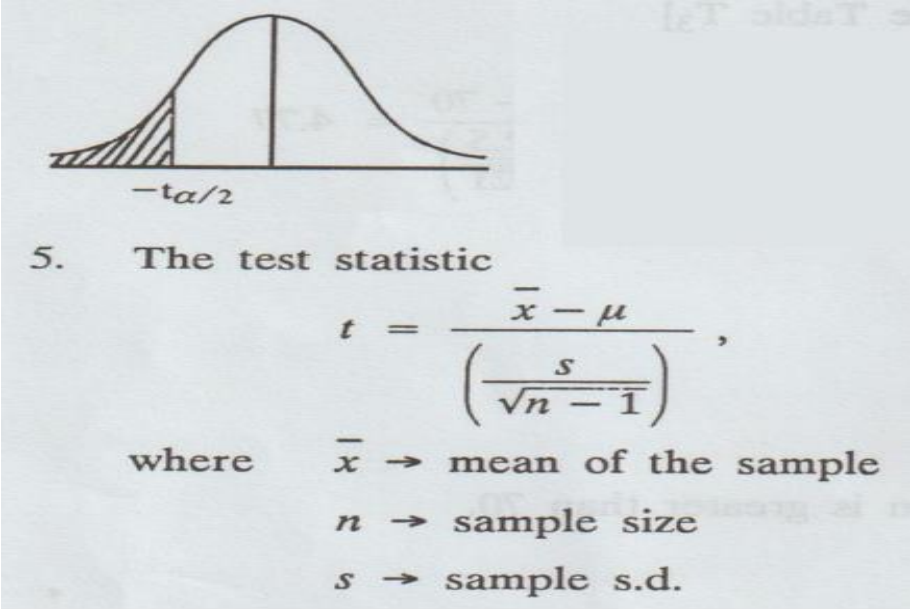
(a) If $\mu \neq \mu_0$, then the test is two-tailed test for the given α.

The critical values are $-t_{\alpha/2}$ and $t_{\alpha/2}$ from the $t_{\alpha/2}$ distribution table with d.f. = n - 1



(b) If $\mu > \mu_0$, then the test is one-tail test (right) for the given α

The critical values is t_α with d.f. = n - 1.



Conclusion :

(a) If $-t_{\alpha/2} < t < t_{\alpha/2}$ then we accept H_0; otherwise, add we reject H_0

(b) If $t < t_\alpha$, then accept H_0, otherwise, we reject H_0

(c) If $-ta < t_\alpha$ then accept H_0, otherwise, we reject H_0.

Example 1

Given a sample mean of 83, a sample standard deviation of 12.5 and a sample size of 22, test the hypothesis that the value of the population mean is 70 against the alternative that it is more than 70. Use the 0.025 significance level.

Solution :

Solution :

Given : $n = 22, \mu = 70, s = 12.5, \bar{x} = 83$

$$\alpha = 0.025 = \frac{25}{1000} \times 100 = 2.5\%$$

1. $H_0 : \mu = 70$ (i.e., $\bar{x} = \mu$)
2. $H_1 : \mu > 70$ (i.e., $\bar{x} > \mu$)

[Use one-tailed test (right)]

3. $\alpha = 0.025 = 2.5\%$, d.f. $= 22 - 1 = 21$
4. Table value $|t| = 2.080$ [See Table T_3]
5. Calculate :

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{83 - 70}{\left(\frac{12.5}{\sqrt{21}}\right)} = 4.77$$

Conclusion :

Here, Cal t > table t

i.e., 4.77 > 2.08

So, we reject H_0

Mean value of the population is greater than 70.

Paired T-Test

The paired sample *t*-test, sometimes called the dependent sample *t*-test, is a statistical procedure. Specifically, it determines whether the mean difference between two sets of observations is zero. In a paired sample t-test, one should measure each subject or entity twice, resulting in pairs of observations. Common

applications of the paired sample *t*-test include case-control studies or repeated-measures designs. Suppose if you want to evaluate the effectiveness of a company training program, you might follow the following approach. One approach you might consider would be to measure the performance of a sample of employees before and after completing the program, and analyze the differences using a paired sample *t*-test.

Hypotheses

Like many statistical procedures, the paired sample *t*-test has two competing hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis assumes that the true mean difference between the paired samples is zero. Under this model, all observable differences are explained by random variation. Conversely, the alternative hypothesis assumes that the true mean difference between the paired samples is not equal to zero. As a result, the alternative hypothesis can take one of several forms depending on the expected outcome. If the direction of the difference does not matter, one should use a two-tailed hypothesis. Otherwise, the power of the test increases by an upper-tailed or lower-tailed hypothesis. The null hypothesis remains the same for each type of alternative hypothesis. The paired sample *t*-test hypotheses are formally defined below:

• The true mean difference (μdμd) poses to be equal to zero in the null hypothesis (H0H0).

• A two-tailed alternative hypothesis (H1H1) believes that the true mean difference μdμd is not equal to zero.

• According to an upper-tailed alternative hypothesis (H1H1) the true mean difference μdμd is greater than zero.

• It is assumed that μdμd is lesser than zero in a lower-tailed alternative hypothesis (H1H1).

Need help conducting your paired t-test? Leverage our 30+ years of experience and low-cost, same-day service to complete your results today!

Schedule now using the calendar below.

Note. It is important to remember that hypotheses are never about data, they are about the processes which produce the data. In the formulas above, the value of μdμd is unknown. The goal of hypothesis testing is to determine the hypothesis (null or alternative) with which the data are more consistent.

Assumptions

As a parametric procedure (a procedure which estimates unknown parameters), the paired sample *t*-test makes several assumptions. Although *t*-tests are quite robust, it is good practice to evaluate the degree of deviation from these assumptions in order to assess the quality of the results. In a paired sample *t*-test, the observations are defined as the differences between two sets of values. Hence, each assumption refers to these differences, and not the original data values. The paired sample *t*-test has four main assumptions:

- • The dependent variable must be continuous (interval/ratio).
- • The observations are independent of one another.
- • The dependent variable should approximately distribute normally.
- • The dependent variable should not contain any outliers.

Level of Measurement

As the paired sample t-test is based on the normal distribution, it requires the sample data to be numeric and continuous. So, the continuous data can take on any value within a range (income, height, weight, etc.). The opposite of continuous data is discrete data, which can only take on a few values (Low, Medium, High, etc.). Occasionally, one can use discrete data to approximate a continuous scale, such as with Likert-type scales.

Independence

You usually cannot test the independence of observations, but you can reasonably assume it if the data collection process was random and without replacement. In our example, it is reasonable to assume that the participating employees are independent of one another.

Normality

To test the assumption of normality, a variety of methods are available, but the simplest is to inspect the data visually using a tool like a histogram (Figure 1). Real-world data are almost never perfectly normal, so the consideration of this assumption reasonably met if the shape looks approximately symmetric and bell-shaped. The data in the example figure below approximately follows a normal distribution.

Procedure

The procedure for a paired sample *t*-test involves four steps. The following are the symbols that should be used:

- D = D = Differences between two paired samples
- di = di = The ith ith observation in DD
- n = n = The sample size
- d¯¯¯ = d¯ = The sample mean of the differences
- σ^ = σ^ = The sample standard deviation of the differences
- T =T =The critical value of a *t*-distribution with (n − 1n − 1) degrees of freedom
- t = t = The *t*-statistic (*t*-test statistic) for a paired sample *t*-test
- p = p = The pp-value (probability value) for the *t*-statistic.

The following are the four steps:

- 1. Calculate the sample mean.
- d¯¯¯ = d1 + d2 + ⋯ + dnnd¯ = d1 + d2 + ⋯ + dnn
- 2. Calculate the sample standard deviation.
- σ^ = (d1 − d¯¯¯)2 + (d2 − d¯¯¯)2 + ⋯ + (dn − d¯¯¯)2n − 1−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−√σ^ = (d1 − d¯)2 + (d2 − d¯)2 + ⋯ + (dn − d¯)2n − 1
- 3. Calculate the test statistic.
- t = d¯¯¯ − 0σ^/n−−√t = d¯ − 0σ^/n
- 4. Calculate the probability of observing the test statistic under the null hypothesis. This value is obtained by comparing *t* to a *t*-distribution with

(n − 1n − 1) degrees of freedom. This can be done by looking up the value in a table, such as those found in many statistical textbooks, or with statistical software for more accurate results.

- $p = 2 \cdot Pr(T > |t|)p = 2 \cdot Pr(T > |t|)$ (two-tailed)
- $p = Pr(T > t)p = Pr(T > t)$ (upper-tailed)
- $p = Pr(T < t)p = Pr(T < t)$ (lower-tailed)

Determine whether the results provide sufficient evidence to reject the null hypothesis in favour of the alternative hypothesis.

Interpretation

There are two types of significance to consider when interpreting the results of a paired sample *t*-test, statistical significance and practical significance.

Statistical Significance

The *p*-value determines the Statistical significance. The *p*-value gives the probability of observing the test results under the null hypothesis. The lower the *p*-value, the lower the probability of obtaining a result like the one that was observed if the null hypothesis was true. Thus, a low *p*-value indicates decreased support for the null hypothesis. However, the possibility that the null hypothesis is true and that we simply obtained a very rare result can never be ruled out completely. The cutoff value for determining statistical significance is ultimately decided on by the researcher, but usually a value of .05 or less is chosen. This corresponds to a 5% (or less) chance of obtaining a result like the one that was observed if the null hypothesis was true.

Practical Significance

Practical significance depends on the subject matter specifically. It is not uncommon, especially with large sample sizes, to observe a result that is statistically significant but not practically significant. In most cases, both types of significance are required in order to draw meaningful conclusions.

Paired T Test Example

For example, imagine we have a training program and administer a pretest and posttest to the same sample of students. Consequently, each student has a pair of test scores. We need to determine whether the average change for the pairs of scores is different from zero.

Here is what the data look like in the datasheet. Note that the analysis does not use the subject's ID number.

SubjectID	Pretest	Posttest
1	90.563	110.642
2	94.816	101.588
3	109.56	120.607
4	90.222	83.2217
5	97.598	109.272
6	91.167	115.806
7	96.65	99.8958
8	97.616	117.94
9	88.845	106.052
10	90.817	82.8229
11	89.294	116.639
12	115.83	128.61
13	121.29	119.665
14	87.872	108.383
15	93.793	96.3738

Here's the deciding characteristic for when you should use paired t tests versus an independent samples t test. Does it make sense to assess the difference within a row? In other words, does each row correspond to one person or item? Are the samples paired with each other?

For our dataset, each row in the dataset contains the same subject in the two measurement columns. Consequently, it makes sense to find the difference between the pairs of values. Because we have paired samples, each difference in a row represents how much a subject's score changed after the training program. The paired t-test is the correct choice.

Conversely, if each row had contained different subjects, it would not make sense to subtract them. The change between the pretest for one subject and the posttest for another does not provide meaningful information. In that case, we'd need to perform an independent samples t test.

Interpreting the Results

Here's how to read and report the results for a paired t test.

Paired T-Test and CI: Pretest, Posttest

```
Paired T for Pretest - Posttest

             N    Mean  StDev  SE Mean
Pretest     15   97.06  10.31     2.66
Posttest    15  107.83  13.25     3.42
Difference  15  -10.77  11.17     2.88


95% CI for mean difference: (-16.96, -4.59)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = -3.73  P-Value = 0.002
```

The output indicates that the mean for the Pretest is 97.06, and for the Posttest it is 107.83. The average difference between the paired pretest and posttest scores is -10.77. If the p-value is less than your significance level, the difference does not equal zero.

Because our p-value (0.002) for the paired sample t-test is less than the standard significance level of 0.05, we can reject the null hypothesis. The results are statistically significant. Our sample data support the notion that the average paired difference does not equal zero. Specifically, the Posttest mean is greater than the Pretest mean.

The sample estimate of the difference (-10.77) is unlikely to equal the population difference. The confidence interval estimates that the actual population difference between the Pretest and Posttest is likely between -16.96 and -4.59.

The negative values reflect the fact that the Pretest has a lower mean than the Posttest (i.e., Pretest – Posttest < 0). The confidence interval excludes the zero (no difference between the paired samples) as a likely value, so we can conclude that the population difference does not equal zero.

If high scores are better, the paired sample t-test indicates that the Posttest scores are significantly better than the pretest scores.

chi-square test

A chi-square test is a statistical hypothesis test used to determine if there's a significant difference between observed and expected frequencies, especially when dealing with categorical data. It helps assess the independence of two categorical variables or whether a sample distribution matches an expected distribution.

- **Purpose:**

To determine if the observed differences between expected and actual frequencies are due to chance or if there's a relationship between the variables being studied.

- **Types of Chi-Square Tests:**
    - **Chi-Square Test of Independence:** Examines the relationship between two categorical variables. For example, is there a relationship between gender and preference for a certain product?
    - **Chi-Square Goodness-of-Fit Test:** Determines how well a sample distribution matches an expected distribution. For example, does a die roll distribution match the expected uniform distribution (1/6 for each side)?
- **When to Use:**

Chi-square tests are used when dealing with categorical data, where you're comparing frequencies rather than numerical scores.

- **How it works:**

The test compares the observed frequencies (actual data) with the expected frequencies (what you'd expect if there were no relationship or if the sample matched the expected distribution). The formula involves calculating a chi-square statistic, which is then compared to a critical value from the chi-square distribution to determine statistical significance.

- **Limitations:**

Chi-square tests are sensitive to sample size and cannot establish causation.

The chi-square formula

Both of Pearson's chi-square tests use the same formula to calculate the test statistic, chi-square (X^2):

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- X^2 is the chi-square test statistic
- Σ is the summation operator (it means “take the sum of”)
- *O* is the observed frequency
- *E* is the expected frequency

The larger the difference between the observations and the expectations (*O* − *E* in the equation), the bigger the chi-square will be. To decide whether the difference is big enough to be statistically significant, you compare the chi-square value to a critical value.

To clarify the calculation of the chi-squared value, we refer to the following case: for variables *one* and *two* with category *A* and *B*, an observation was made or a sample exists. Now we want to check whether the frequencies from the sample correspond to the expected frequencies from the population.

Observed frequency:

	Category A	Category B
Category A	10	13
Category B	13	14

Expected frequency:

	Category A	Category B
Category A	9	11
Category B	12	13

With the upper equation you can now calculate **chi-squared**:

$$\chi^2 = \frac{(10-9)^2}{9} + \frac{(13-11)^2}{11} + \frac{(13-12)^2}{12} + \frac{(14-13)^2}{13} = 0.635$$

After calculating chi-squared the number of degrees of freedom *df* is needed. This is given by

$$df = (p-1)(q-1) = 1$$

with

- *p*: number of lines
- *q*: number of columns

From the table of the chi-squared distribution one can now read the critical chi-squared value. For a significance level of 5 % and a *df* of 1, this results in 3.841. Since the calculated chi-squared value is smaller, there is no significant difference.

As a **prerequisite** for this test, please note that all expected frequencies must be greater than 5.

Goodness of fit test

The Chi-square goodness of fit test checks whether your sample data is likely to be from a specific theoretical distribution. We have a set of data values, and an idea about how the data values are distributed. The test gives us a way to decide if the data values have a “good enough” fit to our idea, or if our idea is questionable.

A goodness-of-fit test determines how well a theoretical distribution (like a normal or binomial distribution) aligns with observed data. It essentially assesses whether the observed frequencies in a sample match the frequencies expected under a specific probability model. The most common type is the chi-square goodness-of-fit test.

Elaboration:

- **Purpose:**

Goodness-of-fit tests are used to evaluate the agreement between a theoretical model and real-world data. This helps researchers determine if their model adequately captures the characteristics of the data.

- **Types:**
    - **Chi-square goodness-of-fit test:** This test compares the observed frequencies of data falling into different categories with the expected frequencies based on a hypothesized distribution.
    - **Kolmogorov-Smirnov test:** This test compares the empirical cumulative distribution function (CDF) of the sample data with the theoretical CDF of a hypothesized distribution.
    - **Shapiro-Wilk test:** This test specifically assesses whether a sample comes from a normal distribution.
- **Process:**
    - **Define the hypothesis:** State the null hypothesis that the observed data follows the theoretical distribution.
    - **Calculate expected frequencies:** Determine the frequencies expected under the hypothesized distribution.
    - **Calculate the test statistic:** Calculate the test statistic, which measures the discrepancy between observed and expected frequencies (e.g., chi-square statistic for the chi-square test).
    - **Determine the p-value:** Find the probability of observing a test statistic as extreme as, or more extreme than, the calculated value, assuming the null hypothesis is true.
    - **Make a decision:** Compare the p-value with the significance level (alpha) to decide whether to reject or fail to reject the null hypothesis.
- **Applications:**

- **Evaluating models:** Determining how well a model fits a dataset.
- **Predicting trends:** Using the test to identify patterns in data and predict future trends.
- **Comparing populations:** Assessing whether a sample is representative of a larger population.

Importance of Goodness-of-Fit Tests

Goodness-of-fit tests are important in statistics for many reasons. First, they provide a way to assess how well a statistical model fits a set of observed data. The main importance of running a goodness-of-fit test is to determine whether the observed data are consistent with the assumed statistical model. By extension, a goodness-of-fit test may be useful in choosing between different models which may better fit the data.

Goodness-of-fit tests can also help to identify outliers or market abnormalities that may be affecting the fit of the model. Outliers can have a large impact on the model fit and may need to be removed or dealt with separately. Sometimes, outliers are not easily identifiable until they have been integrated into an analytical model.

Goodness-of-fit tests can also provide information about the variability of the data and the estimated parameters of the model. This information can be useful for making predictions and understanding the behavior of the system being modeled. Based on the data being fed into the model, it may be necessary to refine the model specific to the dataset being tested, the residuals being calculated, and the p-value for potentially extreme data.