UNIT-II

Introduction to Probability: Classical Probability, Relative Frequency, Sample Space, Events, Types of Probability, conditional Probability, Bayesian Rule, Relative frequency method, Random Variable, Distribution Function, Density Function

Sampling and Sampling Distribution: Random vs Non Random Sampling, Simple random sampling, cluster sampling, concept of sampling distributions, Student's t-test, Chi-square and F-distributions. Central limit theorem and its application, confidence intervals

Reference Books:

1. McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".
2. Swaroop, C. H. (2003). A Byte of Python. Python Tutorial.
3. Ken Black, sixth Editing. Business Statistics for Contemporary Decision Making. “John Wiley & Sons, Inc”.
4. Anderson Sweeney Williams (2011). Statistics for Business and Economics. “Cengage Learning”.

Probability:

- Probability is a branch of mathematics that deals with the likelihood or chance of different outcomes occurring in a particular event.
- It provides a framework for quantifying uncertainty and randomness in various situations. There are different approaches to understanding and calculating probability, and **three fundamental concepts are classical probability, relative frequency, and sample space.**
- Probability is simply how likely something is to happen. Whenever we're unsure about the outcome of an event, we can talk about the probabilities of certain outcomes—how likely they are. The analysis of events governed by probability is called statistics.
- **A strong likelihood or chance of something. The relative possibility an event will occur ...the ratio of the number of actual occurrences to the total number of possible occurrences.**

Three Types of Probability

❖ **1. Classical Probability :**

❖ Classical probability is based on the assumption of equally likely outcomes in an experiment. When all possible outcomes of an experiment are equally likely, classical probability can be calculated by dividing the number of favorable outcomes by the total number of possible outcomes. The classical probability (P(A)) of an event A is given by:

$$P(A) = \frac{\text{Number of Favorable Outcomes for Event A}}{\text{Total Number of Possible Outcomes}}$$

❖ **For example, when rolling a fair six-sided die, the probability of getting a 4 is 1/6 because there is only one favorable outcome (rolling a 4) out of the six possible outcomes (rolling a 1, 2, 3, 4, 5, or 6).**

2. Relative Frequency:

- Relative frequency probability is based on the observed outcomes of an experiment. Instead of assuming equal likelihood, it calculates probability by looking at the proportion of times an event occurs in a large number of trials. The relative frequency (P(A)) of an event A is given by

$$P(A) = \frac{\text{Number of Times Event A Occurs}}{\text{Total Number of Trials}}$$

- **For example, if you toss a coin 100 times and it comes up heads 60 times, the relative frequency of getting heads is 60/100 or 0.6.**

2. Sample Space:

- The sample space of an experiment is the set of all possible outcomes. It is denoted by S and is fundamental to understanding probability. The sample space includes every possible result, and events are subsets of the sample space.
- For example, when rolling a six-sided die, the sample space (S) is {1, 2, 3, 4, 5, 6}. Events, such as getting an even number or rolling a 3, are subsets of this sample space.
- **Events:**
- An event is a subset of the sample space, representing a collection of outcomes.
- Events can be simple (a single outcome) or compound (multiple outcomes).
- The occurrence of an event is an observable result of the experiment.
- Events are often denoted by capital letters (e.g., A, B).

Terms Related to Probability

Experiment:

- An experiment is a type of action with unknown outcomes. There are a few positive outcomes and a few negative consequences in every experiment.
- Scientists will make thousands of unsuccessful attempts before they could make a successful attempt to make any invention.

Random Experiment:

- A random experiment is one in which the set of possible outcomes is known. Still, the specific outcome in a given experiment cannot be predicted before the experiment is carried out.
- Example: Rolling a die, tossing a coin

Trial:

- Trials are the various tries made during an experiment. In other words, a trial is any particular outcome of a random experiment.
- Example: Tossing a coin
- **Event:**
- A trial with a clearly defined outcome is an event. For example, getting a tail when tossing a coin is termed an event.
- **Random Event:**
- A random event cannot be easily foreseen. The chance value for such situations is extremely low. The appearance of a rainbow in the rain is a completely random occurrence.
- **Outcome:**
- The outcome of an event is a collection of all possible outcomes.
- Example: There are two different results when a sportsperson hits a ball towards the goal post. He has a chance to score or miss the goal.

Types of Probability.

- Probability can be categorized into various types, each with its own approach to defining and calculating the likelihood of events. The three main types of probability are Classical Probability, Empirical (or Relative Frequency) Probability, and Subjective Probability.

1. Classical Probability:

- This type of probability is based on a priori knowledge and assumes that all outcomes in the sample space are equally likely.
- It is most applicable to situations where each outcome is equally likely, such as flipping a fair coin or rolling a fair die.
- The classical probability of an event A is calculated as the ratio of the number of favorable outcomes to the total number of possible outcomes.

$$P(A) = \frac{\text{Number of Favorable Outcomes for Event A}}{\text{Total Number of Possible Outcomes}}$$

2. Empirical (Relative Frequency) Probability:

- ❖ Empirical probability is based on observed frequencies from past events or experiments.
- ❖ It involves conducting experiments or observations and calculating the ratio of the number of times an event occurs to the total number of trials.
- ❖ As the number of trials increases, the relative frequency converges to the actual probability of the event.

$$P(A) = \frac{\text{Number of Times Event A Occurs}}{\text{Total Number of Trials}}$$

3. Subjective Probability:

- ❖ Subjective probability is based on an individual's personal judgment or belief about the likelihood of an event occurring.
- ❖ It is subjective in nature and may vary from person to person, depending on their knowledge, experience, and perception.
- ❖ Subjective probabilities are often used in decision-making and situations where objective data is unusual.
- ❖ There is no specific formula for subjective probability; individuals assign probabilities based on their perception, experience, or other subjective factors.

Conditional Probability:

- Conditional probability is one of the types of probability in probability theory, where the probability of one event is dependent on the other event already happened.
- As this type of event is very common in real life, conditional probability is often used to determine the probability of such cases.
- It is denoted by P(A|B), where A and B are events, and it reads as "the probability of A given B.“
- The formula for conditional probability is defined as:
  - $P(A|B)=P(A \cap B) / P(B)$
- P(A|B) is the probability of event A happening, given that event B has already happened.
- P(A∩B) is the probability of both events A and B happening.
- P(B) is the probability of event B happening

Key points about conditional probability:

Interpretation:

- Conditional probability represents the updated probability of an event A occurring based on the knowledge that event B has already occurred.
- It provides a refined probability estimate, taking into account the additional information provided by event B.

Calculation:

- The formula $P(A|B) = P(B)P(A \cap B)$ calculates the conditional probability by dividing the probability of both events A and B occurring by the probability of event B.

Independent Events:

- If events A and B are independent, then $P(A|B)=P(A)$, meaning that the occurrence of event B does not affect the probability of event A.

Dependent Events:

- If events A and B are dependent, the conditional probability , $P(A|B)$ may differ from the unconditional probability P(A).
- **Multiplication Rule:**
- The multiplication rule for probability is related to conditional probability and states that $P(A \cap B)=P(B) \times P(A|B)$.

Bayesian Rule

- Bayesian Rule, also known as Bayes' Theorem or Bayes' Rule, is a fundamental concept in probability theory.
- It provides a way to update the probability of a hypothesis based on new evidence or information. In data analytics, Bayes' Theorem is often used for statistical inference, machine learning, and decision-making.
- The formula for Bayes' Theorem is expressed as follows:

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

Where:

- P(A|B) is the probability of hypothesis A given the evidence B (posterior probability).
- P(B|A) is the probability of evidence B given the hypothesis A (likelihood).
- P(A) is the prior probability of hypothesis A (prior probability).
- P(B) is the probability of evidence B occurring (normalizing constant).

Bayes' Theorem has numerous applications in various fields. Here are a few examples:

Medical Diagnosis:

- Scenario: Suppose you have a rare medical condition (A) that affects only 1% of the population. There is a diagnostic test (B) for this condition, but it is not perfect and produces false positives.
- Application: Bayes' Theorem can be used to calculate the probability that you have the condition given a positive test result, taking into account the test's sensitivity and specificity.

Spam Filtering:

- Scenario: You receive an email, and your email system uses a spam filter. The filter looks at certain characteristics of emails (features) to determine whether an email is spam (A) or not.
- Application: Bayes' Theorem can be applied to update the probability of an email being spam based on observed features, improving the accuracy of the spam filter over time.

Quality Control:

- Scenario: A factory produces items, and there is a quality control test (B) to check whether an item is defective (A) or not.
- Application: Bayes' Theorem can help in updating the probability of an item being defective given a failed quality control test, incorporating information about the overall defect rate and the test's accuracy.

Legal Decision Making:

- Scenario: In a court trial, evidence (B) is presented, and the goal is to determine the guilt or innocence of the accused (A).
- Application: Bayes' Theorem can be used to update the probability of guilt or innocence based on the presented evidence, considering the likelihood of the evidence under both scenarios.

Customer Relationship Management (CRM):

- Scenario: A company wants to predict the likelihood of a customer (A) making a purchase based on their previous behavior and interactions (B).
- Application: Bayes' Theorem is used to update the probability of a customer making a purchase given recent interactions, helping tailor marketing strategies.

Weather Forecasting:

- Scenario: Meteorologists use historical weather data (A) and current observations to predict future weather conditions (B).
- Application: Bayes' Theorem can help in adjusting the probability of certain weather events based on new observations, providing more accurate and updated weather forecasts.

Machine Learning:

- Scenario: In a machine learning classification problem, you want to predict whether an image contains a specific object (A) based on features extracted from the image (B).
- Application: Bayes' Theorem is used in Bayesian classifiers to update the probability of an image containing the object given observed features.

Traffic Flow Prediction:

- Scenario: Transportation planners want to predict the probability of traffic congestion (A) based on historical traffic data and current conditions (B).
- Application: Bayes' Theorem helps in updating the probability of traffic congestion given recent observations, improving traffic management strategies.

Relative frequency method:

- The relative frequency method is a statistical approach used in data analytics to analyze and summarize data based on the observed frequencies or proportions of different outcomes.
- It involves calculating the relative frequency of each event, which is the proportion of times that event occurs relative to the total number of observations.
- This method is particularly useful for exploring and describing the distribution of categorical or discrete data.
- Here are the key steps involved in the relative frequency method:

Collect Data:

- Gather the relevant data, typically categorical or discrete, where observations fall into distinct categories or classes.

Count Frequencies:

- Count the number of occurrences of each category in the dataset. This creates a frequency distribution, showing how often each category appears.

Calculate Relative Frequencies:

- Calculate the relative frequency for each category by dividing the frequency of that category by the total number of observations. Mathematically, it can be expressed as:

$$\text{Relative Frequency of Category} = \frac{\text{Frequency of Category}}{\text{Total Number of Observations}}$$

Express as Percentages:

- Optionally, the relative frequencies can be expressed as percentages by multiplying them by 100.
- This helps in providing a more intuitive understanding of the distribution.

Visualize the Distribution:

- Represent the relative frequencies graphically using charts or graphs such as bar charts, pie charts, or histograms.
- Visualization aids in better understanding the patterns and trends in the data.

❖ Let's consider a simple example where you have collected data on the favorite colors of 100 people:

Blue: 30 people

Red: 20 people

Green: 15 people

Yellow: 10 people

Other: 25 people

Calculations:

$$\text{Relative Frequency of Blue} = \frac{30}{100} = 0.30$$

$$\text{Relative Frequency of Red} = \frac{20}{100} = 0.20$$

$$\text{Relative Frequency of Green} = \frac{15}{100} = 0.15$$

$$\text{Relative Frequency of Yellow} = \frac{10}{100} = 0.10$$

$$\text{Relative Frequency of Other} = \frac{25}{100} = 0.25$$

Random Variable

- A random variable is a mathematical concept used in probability theory and statistics to describe numerical outcomes that result from a random experiment, process, or phenomenon. In simpler terms, it is a variable whose values are determined by chance. Random variables are a key component in the study of probability and are used to model uncertain situations.
- There are two main types of random variables:

Discrete Random Variable:

- A discrete random variable is one that takes on a countable number of distinct values. These values are often isolated points on the number line.
- Examples of discrete random variables include the number of heads obtained when flipping a coin, the number of cars passing through a toll booth in an hour, or the number of emails received in a day.
- The probability distribution of a discrete random variable is often described using a probability mass function (PMF), which gives the probability of each possible value of the random variable.

Continuous Random Variable:

- A continuous random variable is one that can take on any value within a specified range or interval.
- Continuous random variables are associated with continuous phenomena and have an infinite number of possible values.
- Examples include the height of a person, the temperature in a room, or the time it takes for a computer to process a task.
- The probability distribution of a continuous random variable is described using a probability density function (PDF).
- Unlike the PMF for discrete random variables, the PDF does not directly give probabilities for specific values but instead provides the probability density over intervals.

The Distribution Function

- In the theoretical discussion on Random Variables and Probability, we note that the probability distribution induced by a random variable X is determined uniquely by a consistent assignment of mass to semi-infinite intervals of the form $(-\infty,t]$ for each real t.
- This suggests that a natural description is provided by the following.
- **Definition**
- The distribution function FX for random variable X is given by

$$F_X(t) P(X \leq t) = P(X \in (-\infty, t]) \ \forall t \in R$$

- In terms of the mass distribution on the line, this is the probability mass at or to the left of the point t. As a consequence

Density Function

- A density function, also known as a probability density function (PDF), is a function that describes the likelihood of a random variable appearing as a certain value.
- The function's value at any given sample in the sample space can be interpreted as the relative likelihood that the value of the random variable would be equal to that sample

Sample Definition

- How frequently do researchers hunt for the appropriate survey participants for a market research study or an already conducted survey in the field?
- The sample or respondents for this study may be chosen from a group of **known or unknowing consumers or customers.**
- Often times, even though you are aware of the typical respondent profile, you cannot complete your research project without the respondents.
- In these circumstances, researchers and research teams get in touch with specialized organizations to use their respondent panel or purchase respondents from them to finish research studies and surveys.
- They could be respondents from the general population who meet the demographic requirements or those who meet certain criteria.
- The success of research investigations depends on these responders. The many sample types, sampling techniques, and representative instances are covered in length in this page. Additionally, it describes how to compute the size, provides information about an online sample, and highlights the benefits of employing them.

What is a Sample?

- A sample is a summarized set of information that a **researcher selects or picks from a broader population using a predetermined technique of selection**. These components are referred to as **observations, sampling units, or sample points.**
- Developing a sample is a productive way to carry out research. The entire population must frequently be studied, which is difficult, expensive, and time consuming.
- As a result, studying the sample offers information the researcher can use to understand the complete population.
- For instance, a cell phone manufacturer might want to interview students at American universities about certain features.
- If the researcher wants to find features that students utilize, features they would want to see, and the price they are prepared to pay, an extensive research study must be carried out.
- It is crucial to complete this step in order to comprehend the features that need to be developed, those that need to be upgraded, the device's price, and the go-to-market plan.

- There were 24.7 million students enrolled in American universities in 2016-17 alone.
- It is impossible to investigate all of these pupils; the time and money required to do so would render the study meaningless and the new technology unnecessary.
- A sufficient sample of students for study can be obtained by selecting universities based on their geographic location and then selecting a subset of their students.
- The population for market research is typically rather large. The entire population cannot be counted, in all likelihood.
- Typically, the sample represents a sizeable portion of this population.
- Surveys, polls, and questionnaires are then used by researchers to gather data from these samples, and this data analysis is extrapolated to the larger community.

Types of Samples: selection methodologies with examples

- A sampling method is the procedure used to get a sample.
- While this technique generates the quantitative and qualitative data that can be collected as part of a research study, sampling plays a crucial role in the research design.
- Probability sampling and non-probability sampling are two separate approaches to sampling techniques.

Examples of probability sampling techniques

- The process of obtaining a sample through probability sampling involves choosing the objects from a population according to probability theory.
- Everyone in the population is included in this technique, and everyone has an equal chance of getting chosen. Hence, there is absolutely no bias in this kind of sample. The research can then involve every member of the population.
- The selection criteria are chosen at the beginning of the market research study and are a crucial part of the investigation.
- **Four different types of samples can be used in probability sampling. They are:**


1
Simple Random Sampling
2
Cluster Sampling
3
Systematic Sampling
4
Stratified Random Sampling

1. Simple random sampling

- This method of sample selection is the easiest to understand. Each participant has an equal chance of taking part in the study using this strategy. Each member of this sample population has an equal chance of being chosen at random to become an object.
- For instance, if a university dean wanted to gather input from students about how they felt about the professors and their level of education, this sample could include all 1000 students at the university. To create this sample, 100 students can be chosen at random from any class.

2. Cluster sampling

- A sample technique called cluster sampling divides the respondent population into equal clusters. Based on defining demographic factors like age, location, gender, etc., clusters are found and included in a sample. This makes it incredibly simple for a survey developer to draw useful conclusions from the responses.
- For instance, if the FDA (Food and Drug Administration) wishes to gather information on negative drug side effects, it can divide the US mainland into distinct clusters, such as states. Respondents in these clusters are subsequently given research surveys. Using this method of sample generation, comprehensive data collection and easily actionable information are provided.

3. Systematic sampling

- By using systematic sampling, a population is sampled by randomly selecting respondents at equal intervals. Picking a beginning point and then choosing responders at a predetermined sample interval is the method for choosing the sample.
- As an diagram, when choosing 1,000 volunteers for the Olympics from a list of 10,000 applicants, each applicant is assigned a count between 1 and 10,000. Then a sample of 1,000 volunteers can be obtained by counting backwards from 1 and selecting each respondent with a 10-second interval.

4. Stratified random sampling

- In the research design phase, stratified random sampling is a technique for segmenting the respondent population into discrete but pre-defined parameters. The responders in this method don't overlap; rather, they speak for the entire population as a whole.
- For instance, a researcher examining persons from various socioeconomic backgrounds can identify respondents based on their yearly earnings. Afterward, some of the objects from these samples can be employed for the research study. This creates smaller groups of persons or samples.

Non-probability sampling methodologies with examples

- The researcher's judgment is used to choose a sample in the non-probability sampling technique.
- This kind of sample is primarily determined by the researcher's or mathematician's capacity to access it.
- When conducting preliminary research, this kind of sampling is employed since the main goal is to generate a hypothesis regarding the research issue.
- Here, each participant does not have an equal probability of being in the sample population, and the sample is only made aware of these parameters after it has been chosen.
- Non-probability sampling can be divided into **four different kinds of samples**. They are:


1
Convenience Sampling
2
Purposive Sampling
3
Snowball Sampling
4
Quota Sampling

1. Convenience sampling

- In plain English, convenience sampling refers to the ease with which a researcher can contact a respondent.
- The method used to create this sample is not scientific.
- The selection of the sample components is done only on the basis of proximity, not representativeness, and the researchers have almost no control over it.
- When there are time and financial constraints on gathering feedback, this non-probability sampling method is used.
- As an Example, consider researchers who are conducting a mall-intercept study to determine the likelihood that people will use a **perfume** produced by a perfume business.
- Based on their closeness to the survey desk and desire to engage in the study, the sample respondents in this sampling technique are selected.

2. Judgmental/purposive sampling

- The judgmental or purposive sampling approach is a way of selecting a sample based only on the researcher's judgment and understanding of the target audience, the nature of the study, and other relevant factors. Only those individuals who meet the research criteria and end goals are chosen using this sample technique, while the rest are excluded. If the research question is "Would you like to do your Masters?" and the only acceptable response is "Yes," then everyone else is not included in the study. As an illustration, if the research question is "What University do you prefer as a student for Masters?"

3. Snowball sampling

- A non-probability sampling method where the samples have uncommon characteristics is known as snowball sampling or chain-referral sampling. This sampling method uses recommendations from current participants to find the sample populations needed for a study. For instance, when asked for feedback on a touchy subject like AIDS, respondents are reticent to provide details. In this situation, the researcher can enlist individuals who have expertise or understanding of such individuals and ask them to gather information on behalf of the researcher.

4. Quota sampling

- With quota sampling, the researcher is free to choose the sample they want to use based on their stratification. This method's main characteristic is that two persons cannot coexist in two different environments.
- For instance, a shoe producer could want to comprehend how millennial view the brand in relation to other factors like comfort, cost, etc. For this study, it solely chooses female millennial because the goal is to gather opinions on women's shoes.

Understanding T-values and P-values

- Every T-value contains a P-value to work with it.
- A **P-value** is referred to as the **probability that the outcomes from the sample data** happened coincidentally. **A p-value, or probability value, is a number describing how likely it is that your data would have occurred under the null hypothesis of your statistical test.**
- P-values have values starting **from 0% to 100**%. They are generally written as a decimal.
- For instance, a P-value of 10% is 0.1.
- It is good to have low P-values. Lower P-values indicate that the data **did not happen coincidentally**.
- For instance, a P-value of 0.1 indicates that there is only a 1% probability that the experiment's outcomes occurred coincidentally.
- Generally, in many cases, a P-value of 5%, that is 0.05, is accepted to mean the data is said to be valid.

A p-value is a statistical measurement used to validate a hypothesis against observed data. A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference

CHI-SQUARE TEST

- A chi-square test is a statistical test that is used to compare observed and expected results.
- A chi-square test is a statistical hypothesis test that examines whether two categorical variables are independent in influencing the test statistic.
- It's used to compare observed results with expected results to determine if a difference is due to chance or a relationship between the variables.
- **The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration**.
- As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.
- A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable.
- Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.

- It is used to calculate the difference between two categorical variables, which are:
  - As a result of chance
  - Because of the relationship
- Formula For Chi-Square Test

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Where

c = Degrees of freedom

O = Observed Value

E = Expected Value

- The degrees of freedom in a statistical calculation represent the number of variables that can vary in a calculation. The degrees of freedom can be calculated to ensure that chi-square tests are statistically valid. These tests are frequently used to compare observed data with data that would be expected to be obtained if a particular hypothesis were true.
- The Observed values are those you gather yourselves.
- The expected values are the frequencies expected, based on the null hypothesis.

Categorical Variables

- Categorical variables belong to a subset of variables that can be divided into discrete categories. Names or labels are the most common categories. These variables are also known as qualitative variables because they depict the variable's quality or characteristics.
- Categorical variables can be divided into two categories:
- **Nominal Variable:** A nominal variable's categories have no natural ordering. Example: Gender, Blood groups
- **Ordinal Variable:** A variable that allows the categories to be sorted is ordinal variables. Customer satisfaction (Excellent, Very Good, Good, Average, Bad, and so on) is an example.

Use of Chi-Square Test:

- Chi-square is a statistical test that examines the differences between categorical variables from a random sample in order to determine whether the expected and observed results are well-fitting.
- Here are some of the uses of the Chi-Squared test:
- The Chi-squared test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution.
- The Chi-squared test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets.
- **Karl Pearson introduced this test in 1900 for categorical data analysis and distribution. This test is also known as 'Pearson's Chi-Squared Test'.**
- Chi-Squared Tests are most commonly used in hypothesis testing. A hypothesis is an assumption that any given condition might be true, which can be tested afterwards.
- The Chi-Square test estimates the size of inconsistency between the expected results and the actual results when the size of the sample and the number of variables in the relationship is mentioned.

- These tests use degrees of freedom to determine if a particular null hypothesis can be rejected based on the total number of observations made in the experiments. Larger the sample size, more reliable is the result.
- There are two main types of Chi-Square tests namely –
  - Independence
  - Goodness-of-Fit

Independence

- The Chi-Square Test of Independence is a derivable (also known as inferential) statistical test which examines whether the two sets of variables are likely to be related with each other or not.
- This test is used when we have counts of values for two nominal or categorical variables and is considered as non-parametric test.
- A relatively large sample size and independence of observations are the required criteria for conducting this test.

For Example-

- In a movie theatre, suppose we made a list of movie genres.
- Let us consider this as the first variable. The second variable is whether or not the people who came to watch those genres of movies have bought snacks at the theatre.
- Here the null hypothesis is that the genre of the film and whether people bought snacks or not are unreliable.
- If this is true, the movie genres don't impact snack sales.

Goodness-Of-Fit

- In statistical hypothesis testing, the Chi-Square Goodness-of-Fit test determines whether a variable is likely to come from a given distribution or not.
- We must have a set of data values and the idea of the distribution of this data.
- We can use this test when we have value counts for categorical variables.
- This test demonstrates a way of deciding if the data values have a " good enough" fit for our idea or if it is a representative sample data of the entire population.

For Example-

- Suppose we have bags of balls with five different colours in each bag. The given condition is that the bag should contain an equal number of balls of each colour. The idea we would like to test here is that the proportions of the five colours of balls in each bag must be exact.

Write a Python Program to implement Chi-Square Test

```
import numpy as np
from scipy.stats import chi2_contingency
# Define your contingency table (replace with your data)
observed_data = np.array([
    [10, 20, 30],
    [15, 25, 40] ])
# Calculate the Chi-square statistic, p-value, degrees of freedom, and expected table
chi2_statistic, p_value, degrees_of_freedom, expected_data = chi2_contingency(observed_data)
# Print the results
print("Chi-Square Statistic:", chi2_statistic)
print("P-value:", p_value)
print("Degrees of Freedom:", degrees_of_freedom)
print("Expected Table:\n", expected_data)
# Interpretation
if p_value < 0.05:
    print("Reject null hypothesis: There is a statistically significant relationship between the variables.")
else:
    print("Fail to reject null hypothesis: There is no evidence of a statistically significant relationship.")
```

Output

```
Chi-Square Statistic: 0.1296296296296296
P-value: 0.9372410104578182
Degrees of Freedom: 2
Expected Table:
 [[10.71428571 19.28571429 30.        ]
 [14.28571429 25.71428571 40.        ]]
Fail to reject null hypothesis: There is no evidence of a statistically significant relationship.
```

Explanation:

1. **Import libraries:** We import numpy for numerical operations and chi2_contingency from scipy.stats for chi-square test calculations.
2. **Define observed data:** Replace observed_data with your actual contingency table containing observed counts for each category combination.
3. **Calculate Chi-square test:** chi2_contingency function takes the observed data as input and returns the chi-square statistic, p-value, degrees of freedom, and expected table.
4. **Print results:** The program prints the calculated values and interprets the results based on the p-value.
    1. If p-value < 0.05 (common significance level), we reject the null hypothesis and conclude that there is a statistically significant relationship between the variables.
    2. Otherwise, we fail to reject the null hypothesis and say there's no evidence of a significant relationship.

Student's t-Test

- In the area of statistics, a student's t-test is mentioned as a method of testing the theory about the mean of a small sample drawn from a normally distributed population where the standard deviation of the given population is unknown.
- We can define the Student t-test as a method that tells you how significant the differences can be between different groups. **A Student t-test is defined as a statistic and this is used to compare the means of two different populations.**
- It is a method that is often used in hypothesis testing to find out whether a process or whether a given treatment actually has any effect on the population of interest, or whether or not two populations are different from each other.
- You wish to know whether the mean petal length of iris flowers differs according to their distinct species.
- You find two different species of iris flowers growing in a garden and they measure 25 petals of each species.
- You can test the difference between these two groups with the help of the Student t-test.

- The null hypothesis (H0) is one that tells the true difference between these groups.
- The alternate hypothesis (Ha) is one that tells the true difference is different from zero.

Student t Test Introduction

- In the year 1908, an Englishman named **William Sealy Gosset** developed the t-test as well as t distribution.
- **William** worked at the Guinness brewery in Dublin and found which existing statistical techniques using large samples were not useful for the small sample sizes which he encountered in his work).
- The **t distribution** belonging under a family of curves in which the number of degrees of freedom specifies a particular curve.
- As the sample size (and the degrees of freedom) increases, the t distribution approaches the bell shape of the standard normal distribution. In common, for tests involving the mean of a sample of size greater than 30, then the normal distribution is applied.

Types of Student t-Test

- When choosing a Student t-test, two things need to be kept in mind: whether the groups being compared are coming from a single population or two different populations, There are different types of t-tests, but the two most common ones are.
  1. **Independent Samples T-Test**
  2. **Paired Samples T-Test**
- The **independent samples t-test**, also known as the unpaired t-test, is a statistical test that determines if there is a significant difference between the means of two unrelated groups
- The Independent Samples t Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples t Test is a parametric test. This test is also known as: Independent t Test.
- A **paired samples t-test**, also known as a dependent samples t-test, compares the means of two measurements taken from the same individual, object, or related units
- The Paired-Samples T Test procedure compares the means of two variables for a single group. The procedure computes the differences between values of the two variables for each case and tests whether the average differs from 0. The procedure also automates the t-test effect size computation

Student t-Test Formula

- We have already discussed the t-test definition. The formula for the two-sample t-test (a.k.a. the Student's t-test) is shown below.

$$\text{Student t Test Formula, } t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}}$$

- In the formula given above, t is equal to the t-value, x1 and x2 are the means of the two groups being compared, $s^2$is the pooled standard error of the two groups, and n1 and n2 are the numbers of observations in each of the groups.
- A larger t-value denotes the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups.
- You can compare your calculated t-value against the values in a critical value chart to determine whether your t-value is greater than what would be expected by chance. If so, you can reject the null hypothesis and you can conclude which two groups are in fact different.

1. Independent Samples T-Test:

- This test is used when comparing the means of two independent groups. For example, comparing the average scores of two different groups of participants in an experiment or comparing the means of two different treatment groups.
- The formula for the t-statistic in the independent samples t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Where X1 and X2 are the sample means s1 and s2 are the sample standard deviations, and n1 and n2 are the sample sizes.

Paired Samples T-Test:

- This test is used when comparing the means of two related groups. For example, comparing the scores of the same group of participants before and after a treatment.
- The formula for the t-statistic in the paired samples t-test is:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

- where d is the mean of the differences between paired observations, s d is the standard deviation of the differences, and n is the number of pairs.

Example Program: Paired t-Test Program for Dependent Samples

1. **Import the stats module from scipy:**

```
from scipy import stats
```

❖ This imports the necessary statistical functions from the scipy library.

2. **Define the paired_t_test function**

```
def paired_t_test(before, after):
    # Perform paired t-test
    t_statistic, p_value = stats.ttest_rel(before, after)

    return t_statistic, p_value
```

❖ This function takes two lists (before and after) as input, representing paired measurements before and after a treatment.

❖ It then uses stats.ttest_rel to perform a paired t-test and returns the t-statistic and p-value.

3. Example usage:

before_treatment = [28, 30, 32, 34, 36]

after_treatment = [25, 29, 31, 33, 35]

❖ These lists represent the measurements taken before and after a treatment.

4. Perform the paired t-test:

t_statistic, p_value = paired_t_test(before_treatment, after_treatment)

❖ This line calls the paired_t_test function with the provided lists and stores the results in t_statistic and p_value.

5. Print the results:

print("t-statistic:", t_statistic)

print("p-value:", p_value)

❖ These lines print the calculated t-statistic and p-value.

6. Interpret the result:

alpha = 0.05

if p_value < alpha:

print("Reject the null hypothesis. There is a significant difference before and after treatment.")

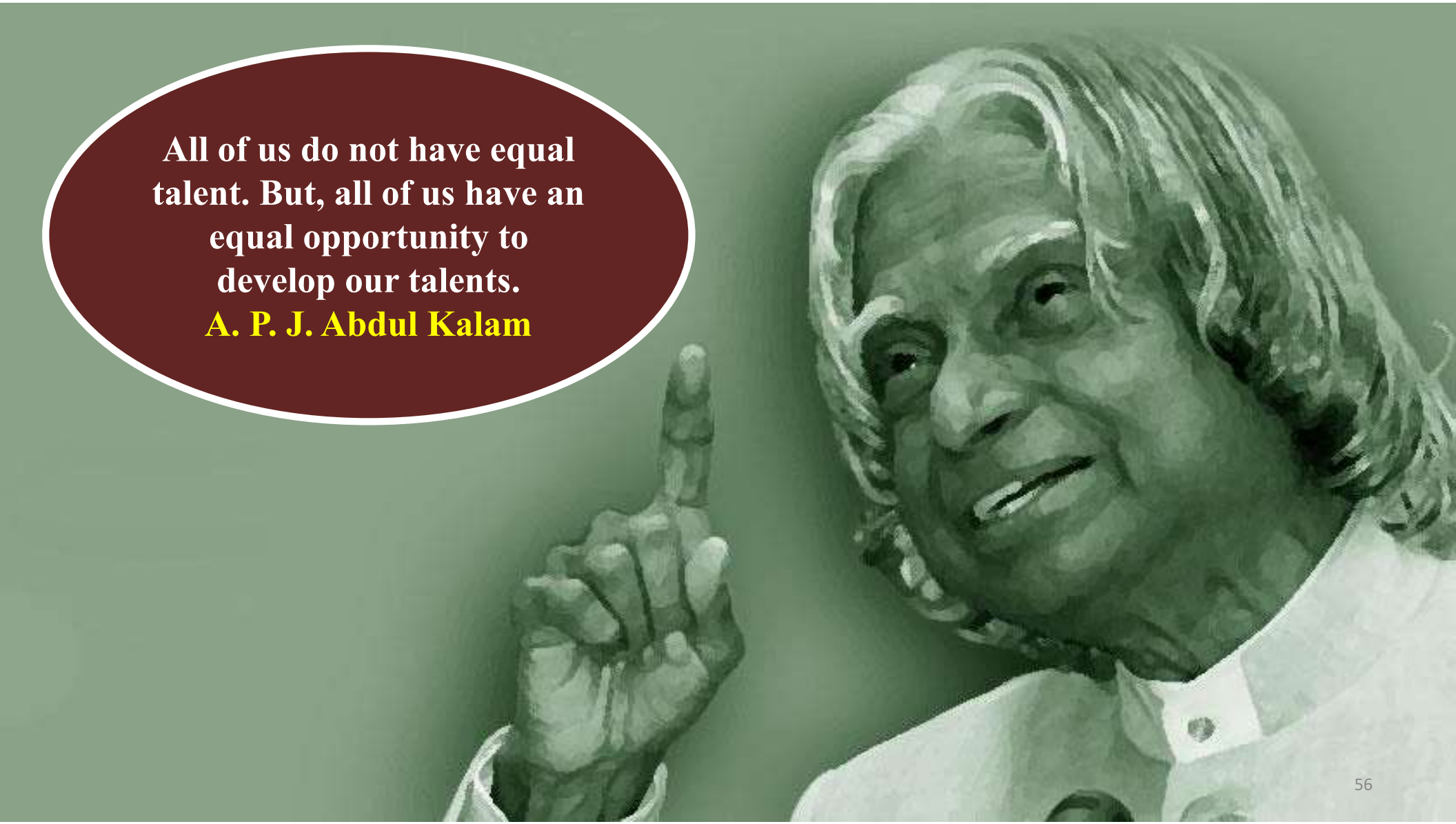
else:

print("Fail to reject the null hypothesis. There is no significant difference before and after treatment.")

❖ These lines interpret the result by comparing the p-value to a significance level (alpha). If the p-value is less than alpha, the null hypothesis is rejected, indicating a significant difference between before and after treatment. Otherwise, the null hypothesis is not rejected. Adjust the alpha level based on the desired significance threshold.

OUTPUT

```
t-statistic: 0.5741692517632145
p-value: 0.5816333668955778
Fail to reject the null hypothesis. There is no significant difference between the groups.
```


All of us do not have equal talent. But, all of us have an equal opportunity to develop our talents.
A. P. J. Abdul Kalam


Thank
You!