DATA SCIENCE

UNIT - II

Statistical Inference

- Need of statistics in Data Science,
- Measures of Central Tendency: Mean, Median, Mode, Mid-range.
- Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation.
- Bayes theorem,
- Basics and need of hypothesis and hypothesis testing,
- Pearson Correlation,
- Sample Hypothesis testing,
- Chi-Square Tests, t-test.

Need of Statistics in Data Science

- Data Exploration and Descriptive Statistics:

  Statistics helps in summarizing and describing the main features of a dataset. Descriptive statistics, such as mean, median, mode, standard deviation, and quartiles, provide a comprehensive overview of the data.

- Data Cleaning and Preprocessing:

  Before applying machine learning algorithms, data scientists often need to preprocess and clean the data. Statistical methods help in handling missing values, outliers, and other anomalies, ensuring the data is suitable for analysis.

Need of Statistics in Data Science

- Inferential Statistics:

  Inferential statistics involves making predictions or inferences about a population based on a sample of data. This is critical in data science when dealing with large datasets, as it's often impractical to analyze the entire population.

- Hypothesis Testing:

  Data scientists use hypothesis testing to make inferences about a population parameter based on a sample of data. This is essential for validating assumptions and drawing meaningful conclusions from data.

Need of Statistics in Data Science

- Probability Distributions:

  Understanding probability distributions is crucial for modeling and analyzing uncertainty in data. Many machine learning algorithms, such as Naive Bayes and Gaussian Mixture Models, rely on probability distributions.

- Statistical Modeling:

  Data scientists use statistical models to identify patterns and relationships within data. Linear regression, logistic regression, and other statistical models help in predicting outcomes and making decisions based on data.

- A/B Testing:

  A/B testing is a common practice in data science for comparing two versions of a product or process. Statistical methods are employed to analyze the results and determine if there's a significant difference between the two versions.

Need of Statistics in Data Science

- Machine Learning Validation:

  In machine learning, statistical methods are used to validate models. Techniques such as cross-validation help assess the performance of a model on different subsets of data, ensuring its generalizability.

- Statistical Inference in Machine Learning:

  Statistical inference is used to draw conclusions from data, making predictions about future events. This is fundamental in machine learning for building models that can make accurate predictions on new, unseen data.

- Decision-Making and Interpretability:

  Statistical analysis help in making informed decisions based on data. It also helps in interpreting the results of machine learning models and understanding the uncertainty associated with predictions.

Measures of Central Tendency

- Measures of central tendency are statistical measures that describe the center or average of a set of values.
- They provide a single representative value around which the data tend to cluster.
- The main measures of central tendency include
  1. Mean
  2. Median
  3. Mode
  4. Mid-range.

Mean

- The mean, often referred to as the average, is calculated by summing up all the values in a dataset and then dividing the sum by the total number of values.
- The mean is sensitive to extreme values (outliers) and is commonly used when the data follows a normal distribution.

Mean = Sum of all values / Number of values

Median

- The median is the middle value in a dataset when it is sorted in ascending or descending order.
- If there is an even number of values, the median is the average of the two middle values.
- The median is less influenced by extreme values and is a robust measure of central tendency.
- The median income in the USA, as of 2014, is $53,700. That means half the people in the USA are making $53,700 or less and the other half are on the other side of that threshold.

Mode

- The mode represents the value(s) that appear most frequently in a dataset.
- A dataset can be unimodal (one mode), bimodal (two modes), trimodal (three modes), or multimodal (more than three modes).
- Unlike the mean and median, the mode can be used for both numerical and categorical data.

Mid-Range

- The mid-range is the arithmetic mean of the maximum and minimum values in a dataset.
- It provides a simple measure of central tendency but is sensitive to extreme values.
- The formula for the mid-range is:

$$\text{Mid-Range} = (\text{Maximum value} + \text{Minimum value}) / 2$$

Measures of Dispersion

- Simply looking at a central point (mean, median, or mode) may not help in understanding the actual shape of a distribution.
- Therefore, we often look at the spread, or the dispersion, of a distribution.
- The following are some of the most common quantities for measures of dispersion.
  1. Range
  2. Variance
  3. Mean Deviation
  4. Standard Deviation.

Range

- The easiest way to look at the dispersion is to take the largest score and subtract it from the smallest score. This is known as the range.
- There is, however, a disadvantage to using the range value: because it uses only the highest and lowest values, extreme scores or outliers tend to result in an inaccurate picture of the more likely range.

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Variance

- The variance is a measure used to indicate how spread out the data points are.
- To measure the variance, the common method is to pick a center of the distribution, typically the mean, then measure how far each data point is from the center.
- If the individual observations vary greatly from the group mean, the variance is big; and vice versa.
- Here, it is important to distinguish between the variance of a population and the variance of a sample.
- They have different notations, and they are computed differently.
- The variance of a population is denoted by σ^2; and the variance of a sample by s^2.

Variance

- The variance of a population is defined by the following formula:

$$\sigma^2 = \frac{\sum (X_i - X)^2}{N}$$

- where σ^2 is the population variance, X is the population mean, X_i is the ith element from the population, and N is the number of elements in the population.

Variance

- The variance of a sample is defined by a slightly different formula:

$$s^2 = \frac{\sum (x_i - x)^2}{(n-1)}$$

- where s^2 is the sample variance, x is the sample mean, x_i is the ith element from the sample, and n is the number of elements in the sample.
- Using this formula, the variance of the sample is an unbiased estimate of the variance of the population.

Standard Deviation

- There is one issue with the variance as a measure.
- It gives us the measure of spread in units squared.
- So, for example, if we measure the variance of age (measured in years) of all the students in a class, the measure we will get will be in $years^2$.
- However, practically, it would make more sense if we got the measure in years (not years squared).
- For this reason, we often take the square root of the variance, which ensures the measure of average spread is in the same units as the original measure.
- This measure is known as the standard deviation

Standard Deviation

- The formula to compute the standard deviation of a sample is

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

Mean Deviation

- Mean deviation is the average of the absolute differences between each data point and the mean.
- It gives an idea of the average distance of the data points from the mean.

Mean Deviation

For a population mean deviation:

$$\text{Population Mean Deviation} = \frac{\sum_{i=1}^{n} |X_i - \mu|}{n}$$

For a sample mean deviation:

$$\text{Sample Mean Deviation} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

Bayes Theorem

- Bayes' Theorem is a fundamental concept in probability theory and statistics.
- In simple terms, it provides a way to update our beliefs or probabilities about an event based on new evidence or information.
- Let's say you have an initial belief or probability (called the prior probability) of something happening.
- As you gather new information, Bayes' Theorem helps you adjust or update that belief to reflect the new evidence.

Bayes Theorem

- The formula for Bayes' Theorem is:

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

  - $P(A|B)$ is the probability of event A occurring given that event B has occurred. This is the updated or posterior probability.
  - $P(B|A)$ is the probability of event B occurring given that event A has occurred. This is the likelihood of the new evidence.
  - $P(A)$ is the prior probability of event A, your initial belief.
  - $P(B)$ is the probability of event B occurring, the total probability of the new evidence.

Outlook	Temperature	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	Yes
Rainy	Mild	High	True	No
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

Example

- Let us say that we need to decide if one should go out to play when the weather is mild based on the dataset.
- We can solve it using the method of posterior probability. Using Bayes' theorem:

$$P(\text{Yes}|\text{Mild}) = \frac{P(\text{Mild}|\text{Yes}) \times P(\text{Yes})}{P(\text{Mild})}$$

Here we have:

$$P(\text{Mild}|\text{Yes}) = 4/9 = 0.44,$$
$$P(\text{Mild}) = 6/14 = 0.43,$$

Now,

$$P(\text{Yes}|\text{Mild}) = (0.44 \times 0.64)/0.43 = 0.65.$$

Example

- In other words, we have derived that the probability of playing when the weather is mild is 65%, and if we wanted to turn that into a Yes–No decision, we can see that this probability is higher than the mid-point, that is, 50%.
- Thus, we can declare “Yes” for our answer.

Basics of Hypothesis

- A hypothesis is a tentative assumption made in order to draw out and test its logical or empirical consequences.
- It's essentially an educated guess or prediction about the relationship between variables or the outcome of a scientific study.

Types of Hypotheses

- Null Hypothesis (H0): This hypothesis proposes that there is no significant difference or relationship between variables or that there is no effect of a treatment.
- Alternative Hypothesis (H1 or Ha): This hypothesis contradicts the null hypothesis, suggesting that there is a significant difference, relationship, or effect.

Basics of Hypothesis Testing

- Hypothesis testing is a statistical method used to make inferences about a population parameter based on sample data.
- It helps determine whether the observed data supports or contradicts the hypothesis.

Steps

1. **Formulate Hypotheses:** Clearly state the null and alternative hypotheses.
2. **Choose a Significance Level:** This is the threshold for rejecting the null hypothesis. Common levels include 0.05 and 0.01.
3. **Collect Data:** Gather relevant data through experiments, surveys, or observations.
4. **Statistical Analysis:** Use appropriate statistical tests to analyze the data and calculate test statistics.
5. **Make a Decision:** Compare the test statistic to the critical value or p-value to determine whether to reject or fail to reject the null hypothesis.
6. **Draw Conclusions:** Based on the decision, draw conclusions about the hypothesis and its implications.

Example: Does a New Diet Pill Affect Weight Loss?

1. Formulate Hypotheses:
   - Null Hypothesis (H0): The new diet pill has no effect on weight loss.
   - Alternative Hypothesis (H1): The new diet pill leads to a significant increase in weight loss.
2. Choose a Significance Level:
   - Let's choose a common significance level, $\alpha = 0.05$. This means we're willing to accept a 5% chance of making a Type I error (rejecting a true null hypothesis).
3. Collect Data:
   - Conduct a study with two groups: one group taking the new diet pill and another group taking a placebo (inactive substance).
   - Measure the weight loss for each participant after a specified period.

Example: Does a New Diet Pill Affect Weight Loss?

4. Statistical Analysis:
   - Use a statistical test (e.g., t-test) to compare the average weight loss between the two groups.
   - Calculate the test statistic and the p-value.
5. Make a Decision:
   - If the p-value is less than or equal to 0.05, you reject the null hypothesis.
   - If the p-value is greater than 0.05, you fail to reject the null hypothesis.
6. Draw Conclusions:
   - If you reject the null hypothesis: Conclude that there is evidence to suggest that the new diet pill has a significant effect on weight loss.
   - If you fail to reject the null hypothesis: Conclude that there is not enough evidence to suggest that the new diet pill leads to a significant increase in weight loss.

Need for Hypothesis and Hypothesis Testing

1. Guidance for Research:

   Hypotheses set the direction for research, specifying what the researcher expects to find and guiding the design of experiments or studies.

1. Framework for Testing:

   Hypothesis testing provides a systematic method using statistical tools to evaluate hypotheses, ensuring decisions are evidence-based rather than intuitive.

1. Control of Type I Error:

   By setting a significance level, hypothesis testing minimizes the risk of mistakenly rejecting a true null hypothesis (Type I error).

Need for Hypothesis and Hypothesis Testing

4. Basis for Inference:

   Hypothesis testing allows researchers to draw broader conclusions about entire populations based on sample data, especially relevant in fields like medicine, social sciences, and economics.

4. Facilitates Scientific Progress:

   Testing hypotheses is essential for advancing scientific knowledge by confirming or refuting theories, contributing to the cumulative growth of scientific understanding.

Type I Error

- Definition:
  - Type I error occurs when we reject a null hypothesis that is actually true.
  - In other words, it's a false positive.
- Symbolically:
  - α (alpha) is commonly used to represent the significance level, which is the probability of making a Type I error.
- Example:
  - Suppose a medical test incorrectly indicates that a healthy person has a disease (false alarm).

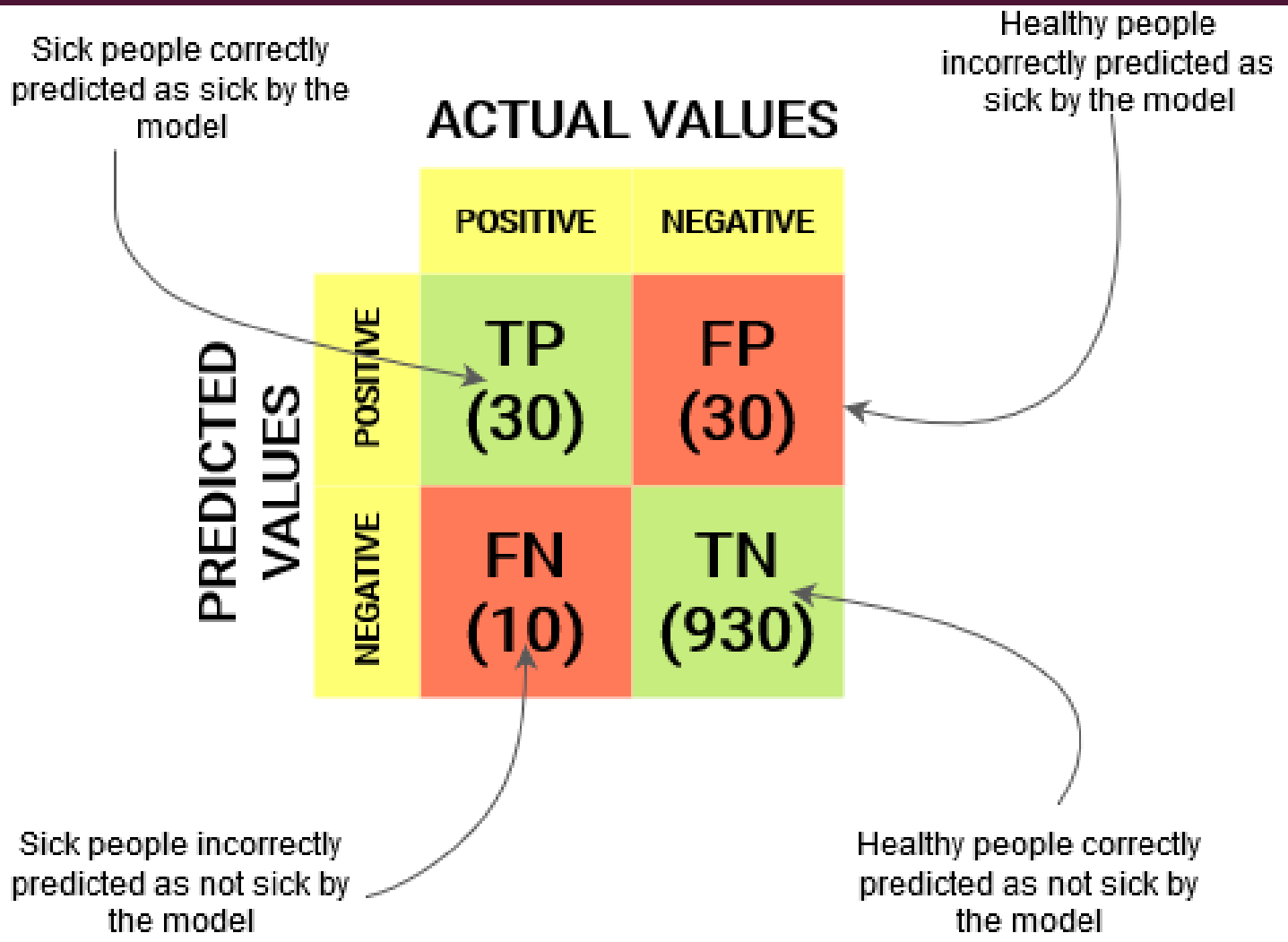
Type II Error

- Definition:
  - Type II error occurs when we fail to reject a null hypothesis that is actually false.
  - It's a false negative.
- Symbolically:
  - β (beta) is commonly used to represent the probability of making a Type II error.
- Example:
  - A medical test fails to detect a disease in a person who actually has the disease (missed diagnosis).

Confusion Matrix

- A confusion matrix is a performance measurement tool used in machine learning, particularly in the field of classification, to evaluate the accuracy of a model.
- It is a table that describes the performance of a classification algorithm on a set of test data for which the true values are known.
- The confusion matrix is composed of four different metrics:
- True Positive (TP): The number of instances correctly predicted as positive.
- True Negative (TN): The number of instances correctly predicted as negative.
- False Positive (FP): The number of instances incorrectly predicted as positive (Type I error).
- False Negative (FN): The number of instances incorrectly predicted as negative (Type II error).

Confusion Matrix



Confusion Matrix

- Accuracy: (TP + TN) / (TP + FP + FN + TN)
- Precision (Positive Predictive Value): TP / (TP + FP)
- Recall (Sensitivity, True Positive Rate): TP / (TP + FN)
- Specificity (True Negative Rate): TN / (TN + FP)
- F1 Score:
  - F1-score is used to evaluate the overall performance of a classification model. It is the harmonic mean of precision and recall,
  - 2 * (Precision * Recall) / (Precision + Recall)

Error Rate

- Error rate is simply one minus the accuracy.
- If the accuracy of a model is 90%, the error rate would be 10%.
- It is calculated as:
  - Error Rate = (FP + FN) / (TP + TN + FP + FN)
  - Or, more simply:
  - Error Rate = 1 - Accuracy

Correlation

- Correlation is a statistical analysis that is used to measure and describe the strength and direction of the relationship between two variables.
- Strength indicates how closely two variables are related to each other, and direction indicates how one variable would change its value as the value of the other variable changes.
- Correlation is a simple statistical measure that examines how two variables change together over time.

Correlation

- Take, for example, “umbrella” and “rain.”
- If someone who grew up in a place where it never rained saw rain for the first time, this person would observe that, whenever it rains, people use umbrellas.
- They may also notice that, on dry days, people do not carry umbrellas.
- By definition, “rain” and “umbrella” are said to be correlated!
- More specifically, this relationship is strong and positive.

Correlation

- A correlation coefficient close to 1 indicates a strong positive correlation, meaning that as one variable increases, the other variable tends to increase as well.
- A correlation coefficient close to -1 indicates a strong negative correlation, implying that as one variable increases, the other variable tends to decrease.
- A correlation coefficient close to 0 suggests a weak or no linear correlation between the variables.

Example

- Positive Correlation:
  - Example: The number of hours a student spends studying and their exam scores.
  - If there's a positive correlation, it means that as the number of study hours increases, the exam scores also tend to increase.
- Negative Correlation:
  - Example: The amount of exercise a person engages in and their body weight.
  - If there's a negative correlation, it means that as the amount of exercise increases, body weight tends to decrease.

Example

- No Correlation:
  - Example: The shoe size of individuals and their IQ scores.
  - If there's no correlation, it implies that there's no systematic relationship between these two variables.
- Strong Correlation:
  - Example: The annual income of individuals and the price of the houses they own.
  - A strong positive correlation might indicate that as income increases, the value of the houses owned also tends to increase.

Example

- Weak Correlation:
  - Example: The amount of rainfall and the sales of sunglasses.
  - A weak correlation might suggest a minor or inconsistent relationship between these two variables.

Pearson Correlation

- An important statistic, the Pearson's r correlation, is widely used to measure the degree of the relationship between linear related variables.
- When examining the stock market, for example, the Pearson's r correlation can measure the degree to which two commodities are related.

Pearson Correlation

- The following formula is used to calculate the Pearson's r correlation:

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

where

r = Pearson's r correlation coefficient,
N = number of values in each dataset,
$\sum xy$ = sum of the products of paired scores,
$\sum x$ = sum of x scores,
$\sum y$ = sum of y scores,

Example

Height	Weight
64.5	118
73.3	143
68.8	172
65	147
69	146
64.5	138
66	175
66.3	134
68.8	172
64.5	118

- N = number of values in each dataset = 10
- $\sum xy$ = sum of the products of paired scores = 98,335.30
- $\sum x$ = sum of x scores = 670.70
- $\sum y$ = sum of y scores = 1463
- $\sum x^2$ = sum of squared x scores = 45,058.21
- $\sum y^2$ = sum of squared y scores = 218,015

Example

- Plugging these values into the Pearson's r correlation formula gives us 0.39 (approximated to two decimal places) as the correlation coefficient.
- This indicates two things:

  (1) "height" and "weight" are positively related, which means that, as one goes up, so does the other; and

  (2) the strength of their relation is medium.

Chi-Square Tests

- Chi-square tests are a family of statistical tests that are used to examine the association or independence between categorical variables.
- There are different types of chi-square tests, each suited for different types of data and research questions.
- Here are two common types:
  - Chi-Square Test for Independence
  - Chi-Square Goodness-of-Fit Test

Chi-Square Test for Independence

- The Chi-Square Test for Independence is a statistical test used to determine if there is a significant association between two categorical variables.
- It assesses whether the observed distribution of frequencies in a contingency table is different from what would be expected under the assumption of independence between the variables.

Example

Table of Observed Values

Qualification / Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

1. Formulate Hypotheses
2. Set Significance Level (α)

Null hypothesis: There is no relation between the marital status and educational qualification.

Alternate Hypothesis: There is significant relation between the marital status and educational qualification.

$$\text{Significance level } (\propto) = 0.05$$

3. Calculate Expected Frequencies

Table of Expected Values

Qualification / Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D
Never Married	$\frac{90 \times 39}{300} = 11.7$	$\frac{90 \times 90}{300} = 27$	25.2	16.2	9.9
Married	19.5	45	42	27	16.5
Divorced	3.9	9	8.4	5.4	3.3
Widowed	3.9	9	8.4	5.4	3.3

Chi-square Equation

$$\frac{(Observed\ value - Expected\ value)^2}{Expected\ value}$$

4. Compute the Chi-Square Statistic

Calculation of χ^2

Observed Values (O)	Expected Values (E)	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
18	11.7	6.3	39.69	3.39
36	27	9	81	3
21	25.2	-4.2	17.64	0.7
9	16.2	-7.2	51.84	3.2
6	9.9	-3.9	15.21	1.53
12	19.5	-7.5	56.25	2.88
36	45	-9	81	1.8
. . .	. . .	. . .	. . .	. . .
3	3.3	-0.3	0.09	0.02
				$\sum \frac{(O-E)^2}{E}$ $\chi^2 = 23.57$

$$\chi^2_{calculated} = 23.57$$

5. Determine Degrees of Freedom

Degrees of freedom= (columns – 1) (rows-1)

=(5-1) (4-1) =4 × 3 =**12**

6. Find Critical Value or P-value

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14

7. Make a Decision

Significance level $(\propto) = 0.05$

$$X^2_{tabular} = 21.03$$

$$\chi^2_{calculated} = 23.57$$

$$X^2_{calculated} > X^2_{tabular} \left(or\ called\ as\ X^2_{critical}\right)$$

$\therefore$ *we reject Null hypothesis, and accept alternate hypothesis*

8. Interpret Results

Alternate Hypothesis: There is significant relation between the marital status and educational qualification.

Chi-Square Goodness-of-Fit Test

- The Chi-Square Goodness-of-Fit Test is a statistical test used to assess whether the observed frequencies in a categorical dataset match the expected frequencies under a specified distribution.
- This test is typically employed when you have one categorical variable with multiple categories and you want to evaluate if the observed data fits a theoretical or expected distribution.

Example

- Example Scenario:
  - Suppose you work for a company that produces packs of colored candies, and the company claims that the distribution of colors in their packs is uniform, meaning each color should appear with equal probability.
- Hypotheses:
- Null Hypothesis
  - The observed distribution of candy colors is not uniform.
- Alternative Hypothesis
  - The observed distribution of candy colors is uniform.
- Data:
  - You randomly sample 200 candies from different packs and record the following counts:

Example

Color	Observed Count
Red	45
Blue	55
Green	50
Yellow	50

- Expected Distribution (Assuming Uniform):
- If the distribution is uniform, each color should occur 200/4=50 times.

Test Procedure

1. **Set Significance Level (α):**
   - Choose a significance level, e.g., 0.05.
2. **Organize Data:**
   - Set up observed and expected frequencies.
3. **Calculate Chi-Square Statistic:**
   - Use the formula: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, where O_i is the observed frequency, E_i is the expected frequency for each category.

$$\chi^2 = \frac{(45-50)^2}{50} + \frac{(55-50)^2}{50} + \frac{(50-50)^2}{50} + \frac{(50-50)^2}{50}$$

Test Procedure

4. **Degrees of Freedom:**
   - For a goodness-of-fit test, the degrees of freedom (df) is $k - 1$, where k is the number of categories.

   $df = 4 - 1 = 3$
5. **Critical Value or P-value:**
   - Compare the calculated chi-square statistic to a critical value from the chi-square distribution table with the determined degrees of freedom.
   - Alternatively, use statistical software to find the p-value associated with the test statistic.
   - If using a significance level of 0.05, the critical value for $df = 3$ is approximately 7.815.

Test Procedure

6. **Decision:**
   - If the calculated chi-square statistic is greater than the critical value or if the p-value is less than the chosen significance level, reject the null hypothesis.

t-test

- A t-test is a statistical method used to determine if there is a significant difference between the means of two groups.
- It is commonly employed when you have a small sample size and are comparing the means of two independent groups or assessing the difference between the means of a single group at two different time points

Independent Samples t-test

- Used when comparing the means of two independent groups.
- Assesses whether the means of two groups are statistically different from each other.
- Example: Comparing the exam scores of two different groups of students who were taught using different teaching methods.

Independent Samples t-test

Formula for Equal Variances:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Formula for Unequal Variances (Welch's t-test):

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

Where:

- $\bar{X}_1, \bar{X}_2$ are the sample means of the two groups.
- s_1^2, s_2^2 are the sample variances of the two groups.
- n_1, n_2 are the sample sizes of the two groups.
- s_p is the pooled standard deviation for equal variances.

Paired Samples t-test (Dependent Samples t-test)

- Used when comparing the means of two related groups (paired observations).
- Assesses whether there is a significant difference between the means of two sets of measurements taken on the same group.
- Example: Comparing the scores of students before and after receiving a specific treatment.

Paired Samples t-test

$$t = \frac{\bar{X}_d}{\frac{s_d}{\sqrt{n}}}$$

Where:

- $\bar{X}_d$ is the mean of the differences between paired observations.
- s_d is the standard deviation of the differences.
- n is the number of paired observations.

One-Sample t-test

- Used when comparing the mean of a single sample to a known or hypothesized population mean.
- Assesses whether the mean of a sample is significantly different from a specified population mean.
- Example: Testing if the average weight of a sample of individuals is significantly different from the known population average weight.

One-Sample t-test

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where:

- $\bar{X}$ is the sample mean.
- μ_0 is the hypothesized population mean.
- s is the sample standard deviation.
- n is the sample size.

Example: Paired Samples t-test (Dependent Samples t-test)

Certainly! Let's consider an example of a paired samples t-test and go through all the calculations step by step.

Example:

Scenario:

Suppose you are investigating whether a new training program has a significant impact on employees' productivity. You collect productivity scores from the same group of employees before and after the training.

Data:

- Before Training: Productivity scores - 78, 82, 75, 80, 79
- After Training: Productivity scores - 85, 88, 82, 87, 89

Hypotheses:

- Null Hypothesis (H_0): There is no significant difference in the mean productivity scores before and after training.
- Alternative Hypothesis (H_1): There is a significant difference in the mean productivity scores before and after training.

Step-by-Step Calculations:

1. **Calculate Differences:**
   - Differences = After Training - Before Training
     - Differences = $85 - 78, 88 - 82, 82 - 75, 87 - 80, 89 - 79$
     - Differences = $7, 6, 7, 7, 10$

2. **Calculate Mean of Differences ($\bar{X}_d$):**
   - $\bar{X}_d = \frac{7+6+7+7+10}{5} = 7.4$
3. **Calculate Standard Deviation of Differences (s_d):**
   - $s_d = \sqrt{\frac{\sum (X_d - \bar{X}_d)^2}{n-1}}$
   - $s_d = \sqrt{\frac{(7-7.4)^2+(6-7.4)^2+\ldots+(10-7.4)^2}{4}} \approx 1.58$
4. **Calculate Standard Error of the Mean (SE):**
   - $SE = \frac{s_d}{\sqrt{n}}$, where n is the number of paired observations.
   - $SE = \frac{1.58}{\sqrt{5}} \approx 0.71$
5. **Calculate t-statistic:**
   - $t = \frac{\bar{X}_d}{SE}$
   - $t = \frac{7.4}{0.71} \approx 10.42$

6. **Degrees of Freedom (df):**
   - $df = n - 1 = 5 - 1 = 4$
7. **Critical Value (for a two-tailed test at a 5% significance level):**
   - Obtain the critical t-value from a t-table or use statistical software.
8. **Compare t-statistic with Critical Value:**
   - If the absolute value of the calculated t-statistic is greater than the critical value, you reject the null hypothesis.

Conclusion:

Compare the calculated t-statistic with the critical value. If $|t| > t_{\text{critical}}$, reject the null hypothesis. In this example, the large t-statistic suggests a significant difference, and you may conclude that the training program had a significant impact on productivity.

Example: One-Sample t-test

Certainly! Let's consider an example where the null hypothesis is rejected in a one-sample t-test:

Example:

Scenario:

Suppose you are investigating whether a new teaching method has a significant impact on students' test scores. You collect test scores from a sample of students who were taught using the new method.

Data:

Sample Test Scores - 72, 78, 75, 80, 77

Population Parameter (Hypothesized Mean):

Expected Average Test Score (μ_0) = 70

Hypotheses:

- Null Hypothesis (H_0): The average test score of students taught using the new method is equal to the expected average test score (no difference).
- Alternative Hypothesis (H_1): The average test score of students taught using the new method is significantly different from the expected average test score.

Step-by-Step Calculations:

1. **Calculate Sample Mean ($\bar{X}$):**
   - $\bar{X} = \frac{72+78+75+80+77}{5} = 76.4$
2. **Calculate Sample Standard Deviation (s):**
   - $s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$
   - $s = \sqrt{\frac{(72-76.4)^2+(78-76.4)^2+\ldots+(77-76.4)^2}{4}} \approx 2.17$

3. **Calculate Standard Error of the Mean (SE):**
   - $SE = \frac{s}{\sqrt{n}}$, where n is the sample size.
   - $SE = \frac{2.17}{\sqrt{5}} \approx 0.97$
4. **Calculate t-statistic:**
   - $t = \frac{\bar{X} - \mu_0}{SE}$
   - $t = \frac{76.4 - 70}{0.97} \approx 6.7$
5. **Degrees of Freedom (df):**
   - $df = n - 1 = 5 - 1 = 4$
6. **Critical Value (for a two-tailed test at a 5% significance level):**
   - Obtain the critical t-value from a t-table or use statistical software.
7. **Compare t-statistic with Critical Value:**
   - If the absolute value of the calculated t-statistic is greater than the critical value, you reject the null hypothesis.

Conclusion:

In this example, the calculated t-statistic ($t \approx 6.7$) is much larger than the critical t-value, leading to the rejection of the null hypothesis. Therefore, you conclude that the new teaching method has a significant impact on students' test scores.

Example: Independent Samples t-test for equal variances (Pooled Variance)

Data:

- Group A: $n_1 = 20, \bar{X}_1 = 45, S_1 = 5$
- Group B: $n_2 = 25, \bar{X}_2 = 40, S_2 = 6$

Equal Variance t-test Formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where S_p is the pooled standard deviation, and it's calculated as:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Degrees of Freedom for the t-test:

$$\text{Degrees of Freedom} = n_1 + n_2 - 2$$

Calculations:

1. Calculate the pooled standard deviation (S_p):

$$S_p = \sqrt{\frac{(20-1)(5^2)+(25-1)(6^2)}{20+25-2}}$$

$$S_p = \sqrt{\frac{19\times25+24\times36}{43}}$$

$$S_p = \sqrt{\frac{475+864}{43}}$$

$$S_p = \sqrt{\frac{1339}{43}}$$

$$S_p \approx \sqrt{31.1395}$$

$$S_p \approx 5.58$$

2. Calculate the t-statistic:

$$t = \frac{45-40}{5.58\times\sqrt{\frac{1}{20}+\frac{1}{25}}}$$

$$t = \frac{5}{5.58 \times \sqrt{0.05+0.04}}$$
$$t = \frac{5}{5.58 \times \sqrt{0.09}}$$
$$t = \frac{5}{5.58 \times 0.3}$$
$$t \approx \frac{5}{1.674}$$
$$t \approx 2.983$$

3. Calculate the degrees of freedom:

$$\text{Degrees of Freedom} = 20 + 25 - 2$$
$$\text{Degrees of Freedom} = 43$$

Results:

- The calculated t-statistic is approximately 2.983.
- The critical t-value at $df = 43$ and $\alpha = 0.05$ is approximately 2.015.

- Since $|2.983| > 2.015$, you would reject the null hypothesis at the 0.05 significance level.

Thus, based on the results of the t-test, there is evidence to suggest a significant difference between the means of Group A and Group B.

Example: Independent Samples t-test for unequal variances (Welch's t-test)

Data:

- Group A: $n_1 = 20, \bar{X}_1 = 45, S_1 = 5$
- Group B: $n_2 = 25, \bar{X}_2 = 40, S_2 = 6$

Unequal Variance t-test (Welch's t-test) Formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Degrees of Freedom for Welch's t-test:

$$\text{Degrees of Freedom} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1}\right)}{n_1 - 1} + \frac{\left(\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}\right)}{n_2 - 1}}$$

Calculations:

1. Calculate the t-statistic:

$$t = \frac{45-40}{\sqrt{\frac{5^2}{20}+\frac{6^2}{25}}}$$

$$t = \frac{5}{\sqrt{\frac{25}{20}+\frac{36}{25}}}$$

$$t = \frac{5}{\sqrt{1.25+1.44}}$$

$$t = \frac{5}{\sqrt{2.69}}$$

$$t \approx \frac{5}{1.640}$$

$$t \approx 3.049$$

2. Calculate the degrees of freedom:

$$\text{Degrees of Freedom} = \frac{\left(\frac{5^2}{20}+\frac{6^2}{25}\right)^2}{\frac{\left(\frac{5^2}{20}\right)^2}{20-1}+\frac{\left(\frac{6^2}{25}\right)^2}{25-1}}$$

$$\text{Degrees of Freedom} \approx \frac{\left(\frac{25}{20}+\frac{36}{25}\right)^2}{\frac{\left(\frac{\left(\frac{25}{20}\right)^2}{19}\right)}{19}+\frac{\left(\frac{\left(\frac{36}{25}\right)^2}{24}\right)}{24}}$$

$$\text{Degrees of Freedom} \approx \frac{\left(\frac{2.25}{1.25}\right)^2}{\frac{\left(\frac{1.102}{19}\right)}{19}+\frac{\left(\frac{2.0736}{24}\right)}{24}}$$

$$\text{Degrees of Freedom} \approx \frac{\left(\frac{1.8}{1.25}\right)^2}{\frac{0.0579}{19}+\frac{0.0864}{24}}$$

$$\text{Degrees of Freedom} \approx \frac{1.44}{\frac{0.0579}{19}+\frac{0.0864}{24}}$$

$$\text{Degrees of Freedom} \approx \frac{1.44}{\frac{0.00305}{19}+\frac{0.0036}{24}}$$

$$\text{Degrees of Freedom} \approx \frac{1.44}{\frac{0.0581}{19}+\frac{0.0036}{24}}$$

$$\text{Degrees of Freedom} \approx \frac{1.44}{0.00306+0.00015}$$

$$\text{Degrees of Freedom} \approx \frac{1.44}{0.00321}$$

$$\text{Degrees of Freedom} \approx 448.48$$

Calculated t-statistic: $t \approx 3.049$

Calculated degrees of freedom: $\text{Degrees of Freedom} \approx 448.48$

Critical t-value for $df \approx 448$ and $\alpha = 0.05$: $t_{\text{critical}} \approx \pm 1.965$ (from t-distribution table)

Comparison:

- The calculated t-statistic (3.049) is greater than the critical t-value (1.965).

Conclusion:

- Since the absolute value of the calculated t-statistic is greater than the critical t-value, you would reject the null hypothesis at the 0.05 significance level.

Therefore, based on this comparison, there is evidence to suggest a significant difference between the means of Group A and Group B in the example.