THOMAS ADEWUNMI UNIVERSITY, OKO, KWARA STATE
Faculty of Management and Social Sciences
Department of Economics

RAIN SEMESTER LECTURE NOTE 2023/2024 Session

COURSE INFO:

Course code:	**ECO 212**
Course title:	**Applied Statistics II**
Credit unit:	**2**

LECTURER INFO:

Lecturer's name:	**Mr Akinbode Damilola**
Department:	**Economics**
E-mail:	**damilola.akinbode@tau.edu.ng**

TOPIC 1: HYPOTHESIS TESTING

Course Description:

This course note covers the fundamental concepts of hypothesis testing in statistics. It provides an understanding of the steps involved in conducting hypothesis tests, the types of errors, and the interpretation of results. The course also discusses various statistical tests used in hypothesis testing.

Objectives:

- Understand the basic concepts and terminology of hypothesis testing.
- Learn the steps involved in conducting a hypothesis test.
- Differentiate between types of errors in hypothesis testing.
- Apply various hypothesis tests to real-world data.

Course Outline:

1. **Introduction to Hypothesis Testing**
   - Definition and Purpose
   - Null and Alternative Hypotheses
2. **Steps in Hypothesis Testing**
   - Formulating Hypotheses
   - Choosing the Significance Level (α)
   - Selecting the Appropriate Test Statistic
   - Decision Rule and P-value
   - Drawing Conclusions
3. **Types of Errors in Hypothesis Testing**
   - Type I Error (False Positive)
   - Type II Error (False Negative)
   - Power of a Test
4. **Common Hypothesis Tests**
   - Z-test
   - T-test (one-sample, two-sample, paired)
   - Chi-square Test
   - ANOVA (Analysis of Variance)
5. **Interpreting Hypothesis Test Results**
   - P-values and Confidence Intervals
   - Practical vs. Statistical Significance
6. **Assumptions and Conditions**
   - Assumptions of Different Tests
   - Checking for Normality, Independence, and Homogeneity
7. **Case Studies and Applications**
   - Real-world Examples of Hypothesis Testing
   - Data Analysis and Interpretation

Detailed Course Notes:

1. Introduction to Hypothesis Testing

- **Definition and Purpose:**
  - **Hypothesis Testing:** A statistical method used to make decisions or inferences about population parameters based on sample data.
  - Purpose: To determine whether there is enough evidence to reject a null hypothesis in favor of an alternative hypothesis.
- **Null and Alternative Hypotheses:**
  - **Null Hypothesis (H0):** A statement of no effect or no difference, serving as the default assumption.
  - **Alternative Hypothesis (H1 or Ha):** A statement that contradicts the null hypothesis, representing the effect or difference being tested.

2. Steps in Hypothesis Testing

- **Formulating Hypotheses:**
  - Define H0 and Ha clearly.
  - Example: H0: $\mu = \mu 0$ (no difference), Ha: $\mu \neq \mu 0$ (difference exists).
- **Choosing the Significance Level (α):**
  - Commonly used levels: 0.05, 0.01.
  - Represents the probability of committing a Type I error.
- **Selecting the Appropriate Test Statistic:**
  - Depends on the sample size, data distribution, and type of hypothesis.
  - Examples: Z, t, chi-square, F.
- **Decision Rule and P-value:**
  - **P-value:** The probability of observing the test statistic as extreme as, or more extreme than, the value observed under H0.
  - Compare p-value with α: if p-value $\leq \alpha$, reject H0; otherwise, fail to reject H0.
- **Drawing Conclusions:**
  - Based on the comparison of the p-value and the significance level.
  - State the conclusion in the context of the research question.

3. Types of Errors in Hypothesis Testing

- **Type I Error (False Positive):**

  - Rejecting H0 when it is true.
  - Probability of Type I error = α.
- **Type II Error (False Negative):**
  - Failing to reject H0 when it is false.
  - Probability of Type II error = β.
- **Power of a Test:**
  - The probability of correctly rejecting H0 when it is false (1 - β).
  - Higher power is desirable and can be increased by increasing the sample size or effect size.

4. Common Hypothesis Tests

- **Z-test:**
  - Used for large sample sizes (n > 30) when the population variance is known.
  - Example: Testing the population mean.
- **T-test:**
  - **One-sample T-test:** Tests if the sample mean is significantly different from a known population mean.
  - **Two-sample T-test:** Tests if the means of two independent samples are significantly different.
  - **Paired T-test:** Tests if the means of two related groups are significantly different.
- **Chi-square Test:**
  - Used for categorical data to test the association between variables or the goodness of fit.
- **ANOVA (Analysis of Variance):**
  - Used to compare the means of three or more groups.
  - **One-way ANOVA:** Tests for differences among group means.
  - **Two-way ANOVA:** Tests for the interaction effect between two independent variables.

5. Interpreting Hypothesis Test Results

- **P-values and Confidence Intervals:**
  - P-value indicates the strength of evidence against H0.

    - Confidence intervals provide a range of values within which the true population parameter is likely to fall.
- **Practical vs. Statistical Significance:**
    - Statistical significance does not always imply practical importance.
    - Consider the effect size and its real-world implications.

6. Assumptions and Conditions

- **Assumptions of Different Tests:**
    - Normality: Data should follow a normal distribution for Z and T tests.
    - Independence: Observations should be independent.
    - Homogeneity of Variance: Variances across groups should be equal for ANOVA.
- **Checking for Normality, Independence, and Homogeneity:**
    - Use graphical methods (histograms, Q-Q plots) and statistical tests (Shapiro-Wilk test) to check normality.
    - Check study design and sampling methods for independence.
    - Use Levene's test for homogeneity of variance.

7. Case Studies and Applications

- **Real-world Examples of Hypothesis Testing:**
    - Clinical trials comparing the effectiveness of treatments.
    - A/B testing in marketing to compare conversion rates.
- **Data Analysis and Interpretation:**
    - Apply hypothesis tests to datasets using statistical software.
    - Interpret results in the context of the research question and draw actionable insights.

Recommended Readings:

- "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne
- "Introduction to the Practice of Statistics" by David S. Moore, George P. McCabe, and Bruce A. Craig
- "Statistical Methods for the Social Sciences" by Alan Agresti and Barbara Finlay

TOPIC 2: Z-TEST HYPOTHESIS TESTING

Course Description:

This course note focuses on Z-test hypothesis testing, a statistical method used to determine if there is a significant difference between sample data and a population parameter. It covers the conditions under which Z-tests are applicable, the steps involved in performing a Z-test, and the interpretation of results.

Objectives:

- Understand the concept and purpose of Z-tests in hypothesis testing.
- Learn the conditions under which Z-tests are appropriate.
- Apply Z-tests to various hypothesis testing scenarios.
- Interpret the results of Z-tests accurately.

Course Outline:

1. **Introduction to Z-Tests**
    - Definition and Purpose
    - Types of Z-Tests
2. **Conditions for Using Z-Tests**
    - Large Sample Sizes
    - Known Population Variance
3. **Steps in Conducting a Z-Test**
    - Formulating Hypotheses
    - Choosing the Significance Level (α)
    - Calculating the Z-Statistic
    - Decision Rule and P-value
    - Drawing Conclusions
4. **One-Sample Z-Test**
    - Hypothesis Testing for Population Mean
    - Worked Examples
5. **Two-Sample Z-Test**
    - Hypothesis Testing for the Difference Between Two Means

- Worked Examples

6. **Interpreting Z-Test Results**
   - P-values and Confidence Intervals
   - Practical vs. Statistical Significance
7. **Common Applications of Z-Tests**
   - Real-world Examples
   - Case Studies

Detailed Course Notes:

1. Introduction to Z-Tests

- **Definition and Purpose:**
  - **Z-Test:** A statistical test used to determine whether there is a significant difference between sample data and a population parameter, assuming the data follows a normal distribution.
  - Purpose: To test hypotheses about population means or proportions using sample data.
- **Types of Z-Tests:**
  - **One-Sample Z-Test:** Tests if the sample mean is significantly different from a known population mean.
  - **Two-Sample Z-Test:** Tests if the means of two independent samples are significantly different.

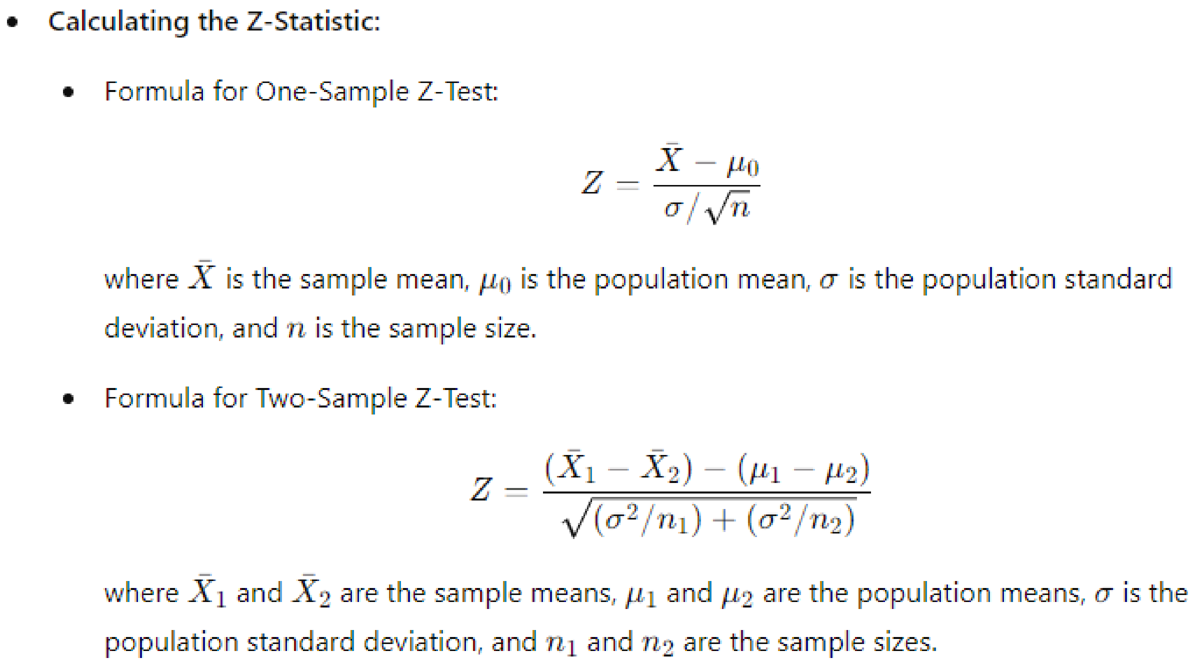
2. Conditions for Using Z-Tests

- **Large Sample Sizes:**
  - Generally, $n > 30$ is considered sufficient for the Central Limit Theorem to apply, allowing the sample mean to be approximately normally distributed.
- **Known Population Variance:**
  - The population variance (σ^2) or standard deviation (σ) must be known. If not, a T-test is more appropriate.

3. Steps in Conducting a Z-Test

- **Formulating Hypotheses:**
  - Define the null hypothesis (H0) and the alternative hypothesis (H1 or Ha).
  - Example: H0: μ = μ0 (no difference), Ha: μ ≠ μ0 (difference exists).

- **Choosing the Significance Level (α):**
  - Commonly used levels: 0.05, 0.01.
  - Represents the probability of committing a Type I error.



- Calculating the Z-Statistic:
  - Formula for One-Sample Z-Test:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

where $\bar{X}$ is the sample mean, μ_0 is the population mean, σ is the population standard deviation, and n is the sample size.

  - Formula for Two-Sample Z-Test:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma^2/n_1) + (\sigma^2/n_2)}}$$

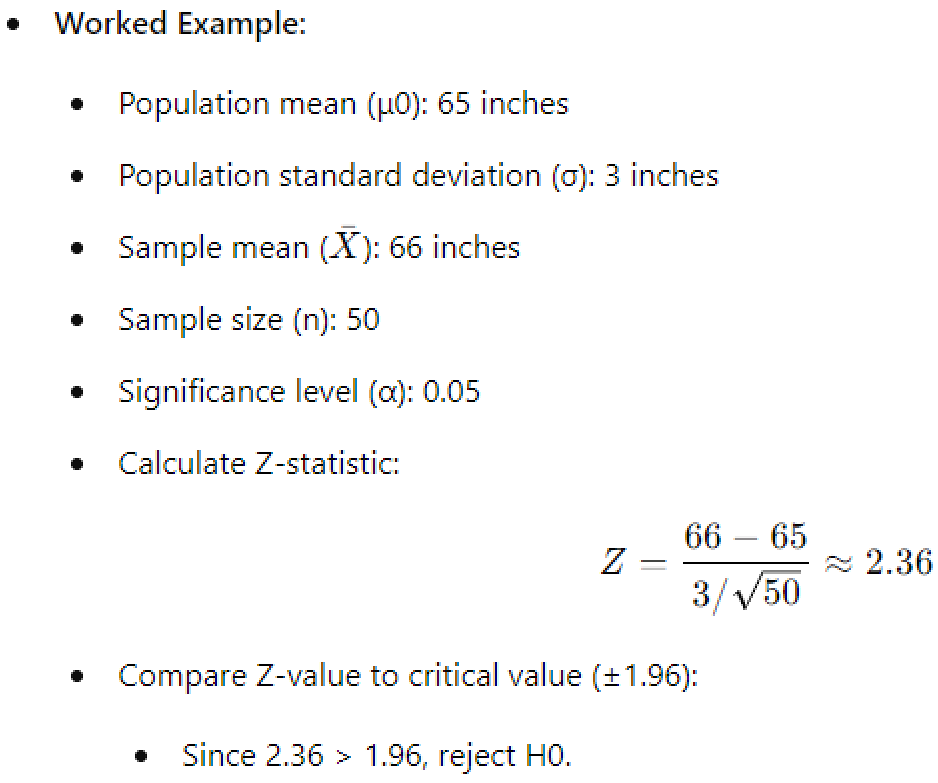
where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, μ_1 and μ_2 are the population means, σ is the population standard deviation, and n_1 and n_2 are the sample sizes.

- **Decision Rule and P-value:**
  - Compare the calculated Z-value to critical Z-values from the standard normal distribution (e.g., ±1.96 for α = 0.05 in a two-tailed test).
  - Alternatively, compare the p-value to the significance level α: if p-value ≤ α, reject H0; otherwise, fail to reject H0.

- **Drawing Conclusions:**
  - Based on the comparison of the p-value and the significance level.
  - State the conclusion in the context of the research question.

4. One-Sample Z-Test

- **Hypothesis Testing for Population Mean:**
  - Example Scenario: Testing if the average height of a sample of students is different from the known population mean height.



- Worked Example:
  - Population mean (μ0): 65 inches
  - Population standard deviation (σ): 3 inches
  - Sample mean ($\bar{X}$): 66 inches
  - Sample size (n): 50
  - Significance level (α): 0.05
  - Calculate Z-statistic:

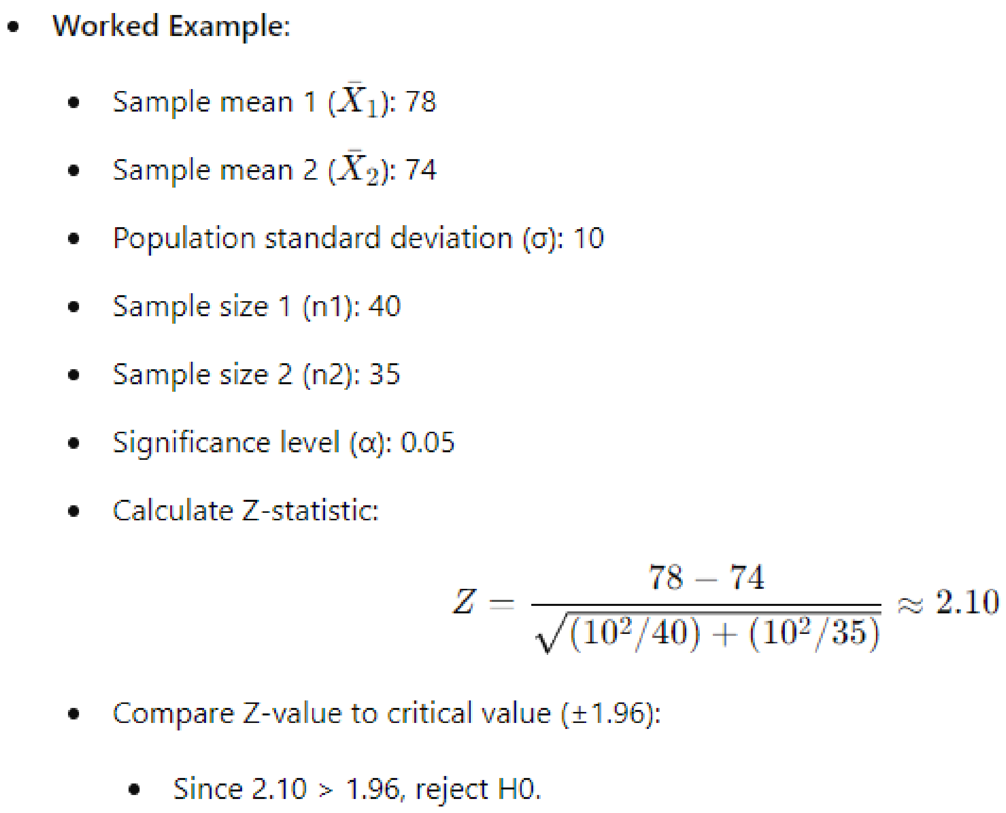
$$Z = \frac{66 - 65}{3/\sqrt{50}} \approx 2.36$$

  - Compare Z-value to critical value (±1.96):
    - Since 2.36 > 1.96, reject H0.

  - Conclusion: There is a significant difference between the sample mean and the population mean.

5. Two-Sample Z-Test

- **Hypothesis Testing for the Difference Between Two Means:**
  - Example Scenario: Testing if the average scores of two different classes are significantly different.



- Worked Example:
  - Sample mean 1 ($\bar{X}_1$): 78
  - Sample mean 2 ($\bar{X}_2$): 74
  - Population standard deviation (σ): 10
  - Sample size 1 (n1): 40
  - Sample size 2 (n2): 35
  - Significance level (α): 0.05
  - Calculate Z-statistic:

$$Z = \frac{78 - 74}{\sqrt{(10^2/40) + (10^2/35)}} \approx 2.10$$

  - Compare Z-value to critical value (±1.96):
    - Since 2.10 > 1.96, reject H0.

  - Conclusion: There is a significant difference between the means of the two classes.

6. Interpreting Z-Test Results

- **P-values and Confidence Intervals:**
  - **P-value:** The probability of observing the test statistic as extreme as, or more extreme than, the value observed under H0.
  - **Confidence Interval:** Provides a range of values within which the true population parameter is likely to fall.
- **Practical vs. Statistical Significance:**
  - Statistical significance does not always imply practical importance.
  - Consider the effect size and its real-world implications.

7. Common Applications of Z-Tests

- **Real-world Examples:**
  - Comparing the average test scores of a sample to a national average.

    - Determining if a new manufacturing process results in a different mean production time compared to the old process.

- **Case Studies:**
    - **Medical Research:** Testing the efficacy of a new drug compared to a standard treatment.
    - **Business Analytics:** Analyzing whether a new marketing strategy significantly impacts sales compared to the previous strategy.

Recommended Readings:

- "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne
- "Introduction to the Practice of Statistics" by David S. Moore, George P. McCabe, and Bruce A. Craig
- "Fundamentals of Statistics" by Michael Sullivan

TOPIC 3: T-TEST HYPOTHESIS TESTING

Course Description:

This course note provides an in-depth exploration of T-test hypothesis testing, a statistical method used to compare sample means and determine if there are significant differences. It covers various types of T-tests, the conditions under which they are used, and the interpretation of results.

Objectives:

- Understand the concept and purpose of T-tests in hypothesis testing.
- Learn the different types of T-tests and their applications.
- Conduct T-tests and interpret their results.
- Apply T-tests to real-world data scenarios.

Course Outline:

1. **Introduction to T-Tests**
   - Definition and Purpose
   - Types of T-Tests
2. **Conditions for Using T-Tests**
   - Small Sample Sizes
   - Unknown Population Variance
   - Normality Assumption
3. **Steps in Conducting a T-Test**
   - Formulating Hypotheses
   - Choosing the Significance Level (α)
   - Calculating the T-Statistic
   - Decision Rule and P-value
   - Drawing Conclusions
4. **Types of T-Tests**
   - One-Sample T-Test
   - Two-Sample T-Test (Independent Samples)
   - Paired Sample T-Test (Dependent Samples)

5. **One-Sample T-Test**
    - Hypothesis Testing for Population Mean
    - Worked Examples
6. **Two-Sample T-Test**
    - Hypothesis Testing for the Difference Between Two Independent Means
    - Worked Examples
7. **Paired Sample T-Test**
    - Hypothesis Testing for the Difference Between Two Related Means
    - Worked Examples
8. **Interpreting T-Test Results**
    - P-values and Confidence Intervals
    - Practical vs. Statistical Significance
9. **Common Applications of T-Tests**
    - Real-world Examples
    - Case Studies

Detailed Course Notes:

1. Introduction to T-Tests

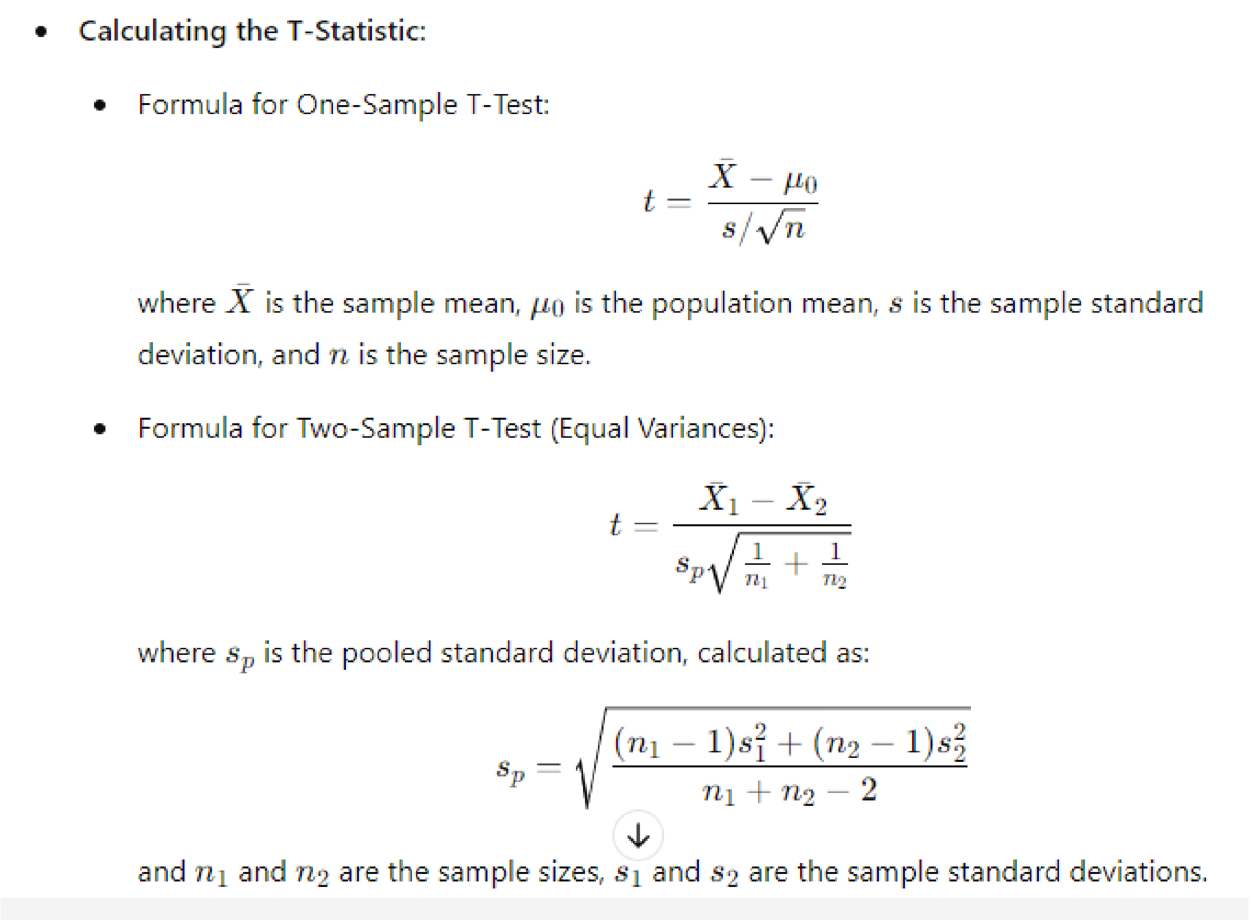
- **Definition and Purpose:**
    - **T-Test:** A statistical test used to compare the means of one or two samples and determine if they are significantly different from each other or a known value.
    - Purpose: To test hypotheses about population means using sample data, especially when the population variance is unknown and the sample size is small.
- **Types of T-Tests:**
    - **One-Sample T-Test:** Tests if the sample mean is significantly different from a known population mean.
    - **Two-Sample T-Test:** Tests if the means of two independent samples are significantly different.
    - **Paired Sample T-Test:** Tests if the means of two related groups are significantly different.

2. Conditions for Using T-Tests

- **Small Sample Sizes:**
  - T-tests are appropriate for small sample sizes (typically $n < 30$).
- **Unknown Population Variance:**
  - The population variance is unknown, and the sample standard deviation is used as an estimate.
- **Normality Assumption:**
  - The data should approximately follow a normal distribution. For small samples, this is critical; for larger samples, the T-test is robust to deviations from normality.

3. Steps in Conducting a T-Test

- **Formulating Hypotheses:**
  - Define the null hypothesis (H0) and the alternative hypothesis (H1 or Ha).
  - Example: H0: $\mu = \mu0$ (no difference), Ha: $\mu \neq \mu0$ (difference exists).
- **Choosing the Significance Level (α):**
  - Commonly used levels: 0.05, 0.01.
  - Represents the probability of committing a Type I error.



- **Calculating the T-Statistic:**
    - Formula for One-Sample T-Test:

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

where $\bar{X}$ is the sample mean, μ_0 is the population mean, s is the sample standard deviation, and n is the sample size.

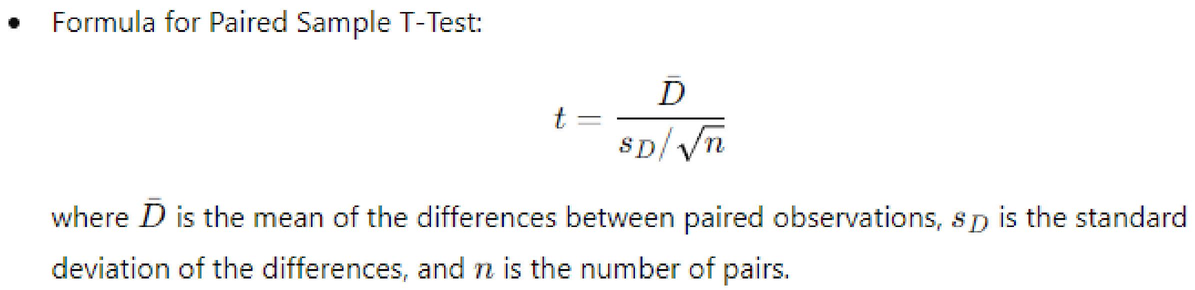
    - Formula for Two-Sample T-Test (Equal Variances):

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

where s_p is the pooled standard deviation, calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

and n_1 and n_2 are the sample sizes, s_1 and s_2 are the sample standard deviations.



- Formula for Paired Sample T-Test:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

where $\bar{D}$ 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.

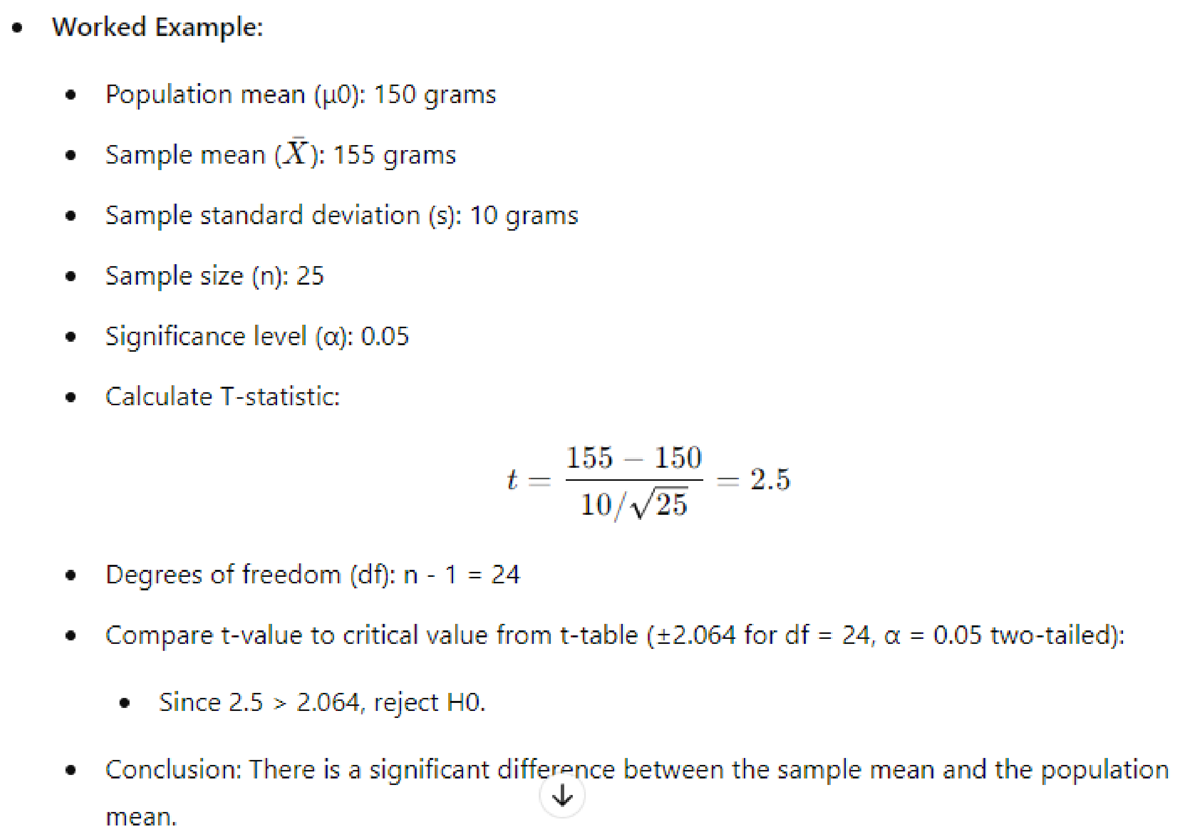
- **Decision Rule and P-value:**
    - Compare the calculated t-value to critical t-values from the t-distribution table (based on degrees of freedom and significance level).
    - Alternatively, compare the p-value to the significance level α: if p-value ≤ α, reject H0; otherwise, fail to reject H0.
- **Drawing Conclusions:**
    - Based on the comparison of the p-value and the significance level.

- State the conclusion in the context of the research question.

4. Types of T-Tests

One-Sample T-Test

- **Hypothesis Testing for Population Mean:**
    - Example Scenario: Testing if the average weight of a sample of apples is different from the known population mean weight.



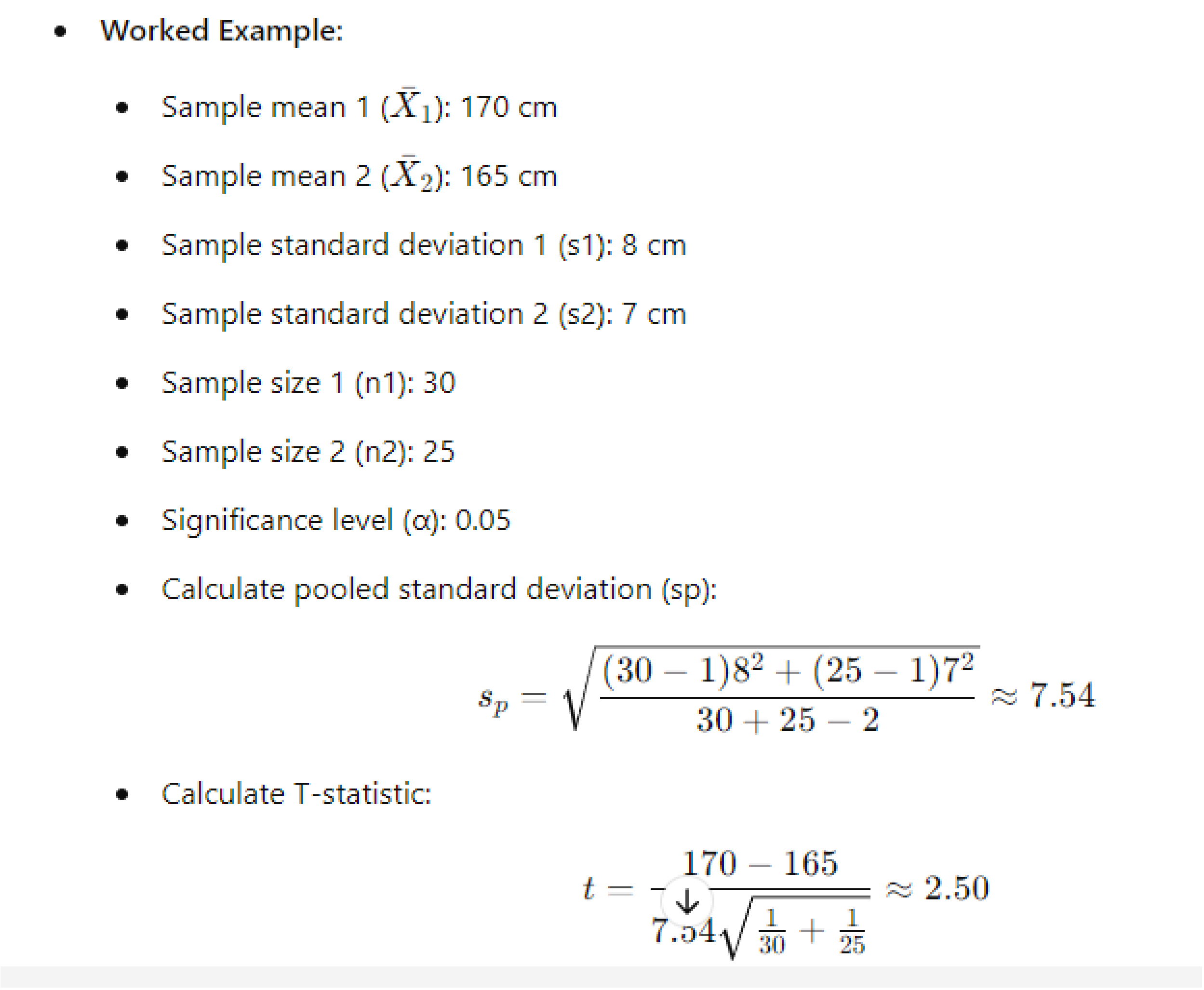
- Worked Example:
    - Population mean (μ0): 150 grams
    - Sample mean ($\bar{X}$): 155 grams
    - Sample standard deviation (s): 10 grams
    - Sample size (n): 25
    - Significance level (α): 0.05
    - Calculate T-statistic:

$$t = \frac{155 - 150}{10/\sqrt{25}} = 2.5$$

    - Degrees of freedom (df): n - 1 = 24
    - Compare t-value to critical value from t-table (±2.064 for df = 24, α = 0.05 two-tailed):
        - Since 2.5 > 2.064, reject H0.
    - Conclusion: There is a significant difference between the sample mean and the population mean.

Two-Sample T-Test

- **Hypothesis Testing for the Difference Between Two Independent Means:**
    - Example Scenario: Testing if the average heights of male and female students are significantly different.



- **Worked Example:**
  - Sample mean 1 ($\bar{X}_1$): 170 cm
  - Sample mean 2 ($\bar{X}_2$): 165 cm
  - Sample standard deviation 1 (s1): 8 cm
  - Sample standard deviation 2 (s2): 7 cm
  - Sample size 1 (n1): 30
  - Sample size 2 (n2): 25
  - Significance level (α): 0.05
  - Calculate pooled standard deviation (sp):

$$s_p = \sqrt{\frac{(30-1)8^2 + (25-1)7^2}{30+25-2}} \approx 7.54$$

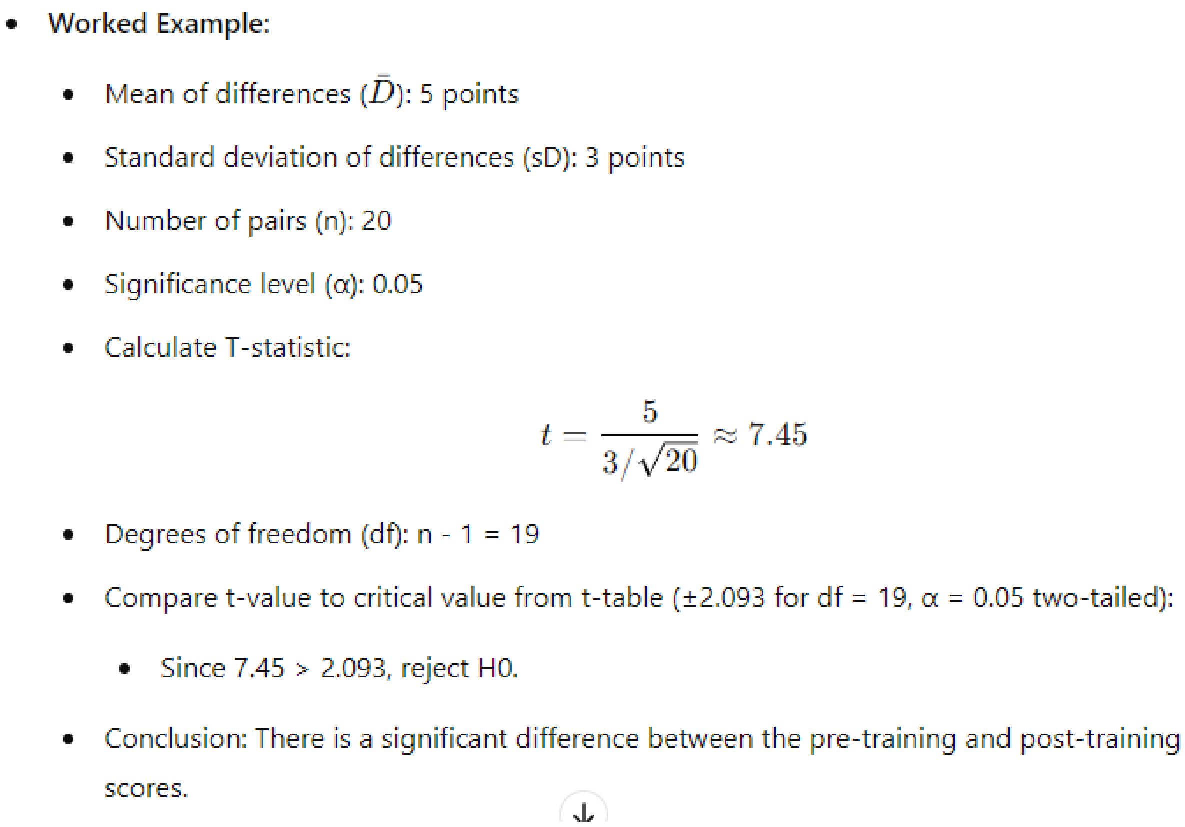
  - Calculate T-statistic:

$$t = \frac{170-165}{7.54\sqrt{\frac{1}{30}+\frac{1}{25}}} \approx 2.50$$

  - Degrees of freedom (df): n1 + n2 - 2 = 53
  - Compare t-value to critical value from t-table (±2.009 for df = 53, α = 0.05 two-tailed):
    - Since 2.50 > 2.009, reject H0.
  - Conclusion: There is a significant difference between the means of the two groups.

Paired Sample T-Test

- **Hypothesis Testing for the Difference Between Two Related Means:**
  - Example Scenario: Testing if the average test scores of students before and after a training program are significantly different.



- Worked Example:
  - Mean of differences ($\bar{D}$): 5 points
  - Standard deviation of differences (sD): 3 points
  - Number of pairs (n): 20
  - Significance level (α): 0.05
  - Calculate T-statistic:

$$t = \frac{5}{3/\sqrt{20}} \approx 7.45$$

  - Degrees of freedom (df): n - 1 = 19
  - Compare t-value to critical value from t-table (±2.093 for df = 19, α = 0.05 two-tailed):
    - Since 7.45 > 2.093, reject H0.
  - Conclusion: There is a significant difference between the pre-training and post-training scores.

8. Interpreting

T-Test Results

- **P-values and Confidence Intervals:**
  - **P-value:** The probability of observing the test statistic as extreme as, or more extreme than, the value observed under H0.
  - **Confidence Interval:** Provides a range of values within which the true population parameter is likely to fall.
- **Practical vs. Statistical Significance:**
  - Statistical significance does not always imply practical importance.
  - Consider the effect size and its real-world implications.

9. Common Applications of T-Tests

- **Real-world Examples:**
    - Comparing the effectiveness of two different medications in clinical trials.
    - Analyzing the impact of a new teaching method on student performance.
- **Case Studies:**
    - **Health Sciences:** Testing if there is a difference in blood pressure levels between patients receiving two different treatments.
    - **Education:** Examining if there is a significant difference in reading comprehension scores before and after implementing a new curriculum.

Recommended Readings:

- "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne
- "Introduction to the Practice of Statistics" by David S. Moore, George P. McCabe, and Bruce A. Craig
- "Fundamentals of Biostatistics" by Bernard Rosner

TOPIC 4: CHI-SQUARE HYPOTHESIS TESTING

Course Description:

This course note focuses on Chi-Square hypothesis testing, a statistical method used to test relationships between categorical variables. It covers the types of Chi-Square tests, the conditions under which they are appropriate, and the interpretation of results.

Objectives:

- Understand the concept and purpose of Chi-Square tests in hypothesis testing.
- Learn the different types of Chi-Square tests and their applications.
- Conduct Chi-Square tests and interpret their results.
- Apply Chi-Square tests to real-world data scenarios.

Course Outline:

1. **Introduction to Chi-Square Tests**
    - Definition and Purpose
    - Types of Chi-Square Tests
2. **Conditions for Using Chi-Square Tests**
    - Categorical Data
    - Independence of Observations
    - Expected Frequency Conditions
3. **Steps in Conducting a Chi-Square Test**
    - Formulating Hypotheses
    - Choosing the Significance Level (α)
    - Calculating the Chi-Square Statistic
    - Decision Rule and P-value
    - Drawing Conclusions
4. **Types of Chi-Square Tests**
    - Chi-Square Test for Goodness of Fit
    - Chi-Square Test for Independence
    - Chi-Square Test for Homogeneity
5. **Chi-Square Test for Goodness of Fit**

    - Hypothesis Testing for Distributional Assumptions
    - Worked Examples
6. **Chi-Square Test for Independence**
    - Hypothesis Testing for Association Between Two Categorical Variables
    - Worked Examples
7. **Chi-Square Test for Homogeneity**
    - Hypothesis Testing for Comparing Proportions Across Groups
    - Worked Examples
8. **Interpreting Chi-Square Test Results**
    - P-values and Degrees of Freedom
    - Practical vs. Statistical Significance
9. **Common Applications of Chi-Square Tests**
    - Real-world Examples
    - Case Studies

Detailed Course Notes:

1. Introduction to Chi-Square Tests

- **Definition and Purpose:**
    - **Chi-Square Test:** A statistical test used to determine if there is a significant association between categorical variables or if a sample data matches a population distribution.
    - Purpose: To test hypotheses about categorical data.
- **Types of Chi-Square Tests:**
    - **Chi-Square Test for Goodness of Fit:** Tests if observed frequencies match expected frequencies based on a specific distribution.
    - **Chi-Square Test for Independence:** Tests if there is an association between two categorical variables.
    - **Chi-Square Test for Homogeneity:** Tests if different populations have the same proportions of a characteristic.

2. Conditions for Using Chi-Square Tests

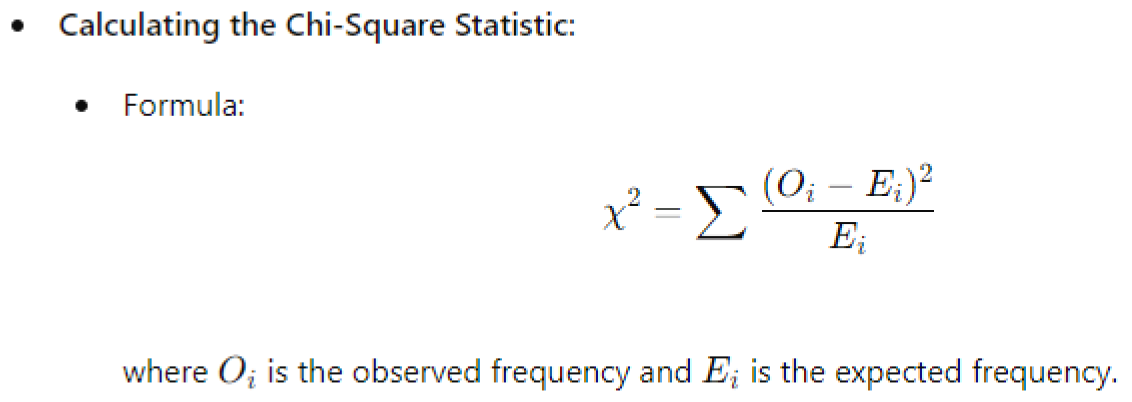
- **Categorical Data:**

- Data should be in categories, not numerical.

- **Independence of Observations:**
  - Each observation should be independent of others.
- **Expected Frequency Conditions:**
  - Generally, expected frequency in each cell should be at least 5 for the Chi-Square test to be valid.

3. Steps in Conducting a Chi-Square Test

- **Formulating Hypotheses:**
  - Define the null hypothesis (H0) and the alternative hypothesis (H1 or Ha).
  - Example for Goodness of Fit: H0: The data follows a specified distribution, Ha: The data does not follow the specified distribution.
  - Example for Independence: H0: There is no association between the variables, Ha: There is an association between the variables.
- **Choosing the Significance Level (α):**
  - Commonly used levels: 0.05, 0.01.
  - Represents the probability of committing a Type I error.



- Calculating the Chi-Square Statistic:
  - Formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency and E_i is the expected frequency.

- **Decision Rule and P-value:**
  - Compare the calculated Chi-Square value to critical values from the Chi-Square distribution table (based on degrees of freedom and significance level).

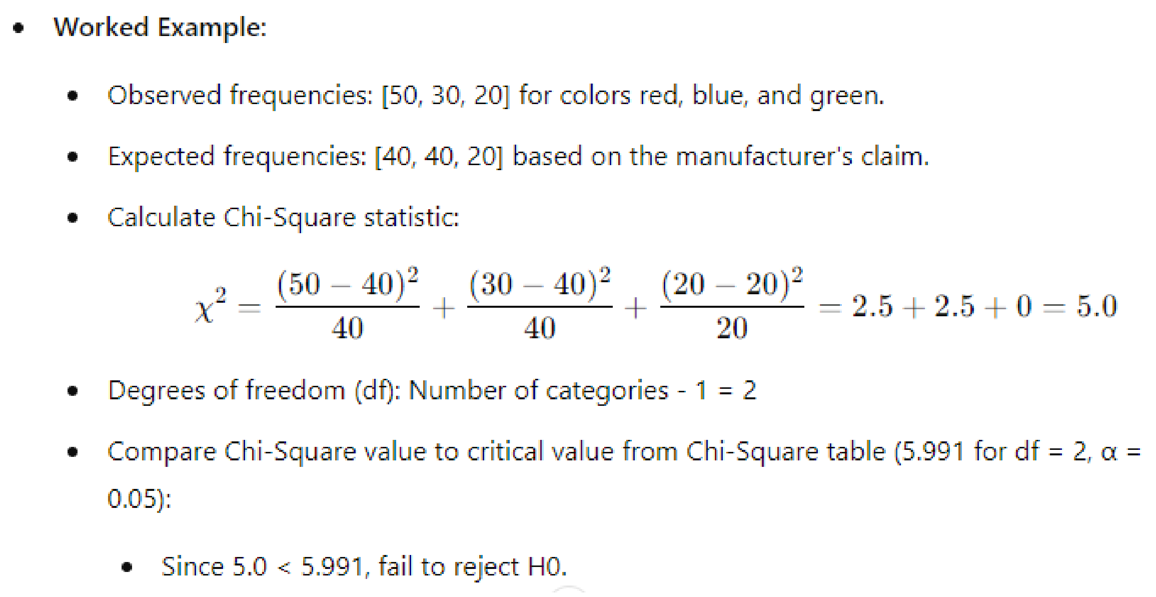
  - Alternatively, compare the p-value to the significance level α: if p-value ≤ α, reject H0; otherwise, fail to reject H0.

- **Drawing Conclusions:**
  - Based on the comparison of the p-value and the significance level.
  - State the conclusion in the context of the research question.

4. Types of Chi-Square Tests

Chi-Square Test for Goodness of Fit

- **Hypothesis Testing for Distributional Assumptions:**
  - Example Scenario: Testing if the distribution of colors in a bag of candies matches the manufacturer's claimed distribution.



- Worked Example:
  - Observed frequencies: [50, 30, 20] for colors red, blue, and green.
  - Expected frequencies: [40, 40, 20] based on the manufacturer's claim.
  - Calculate Chi-Square statistic:

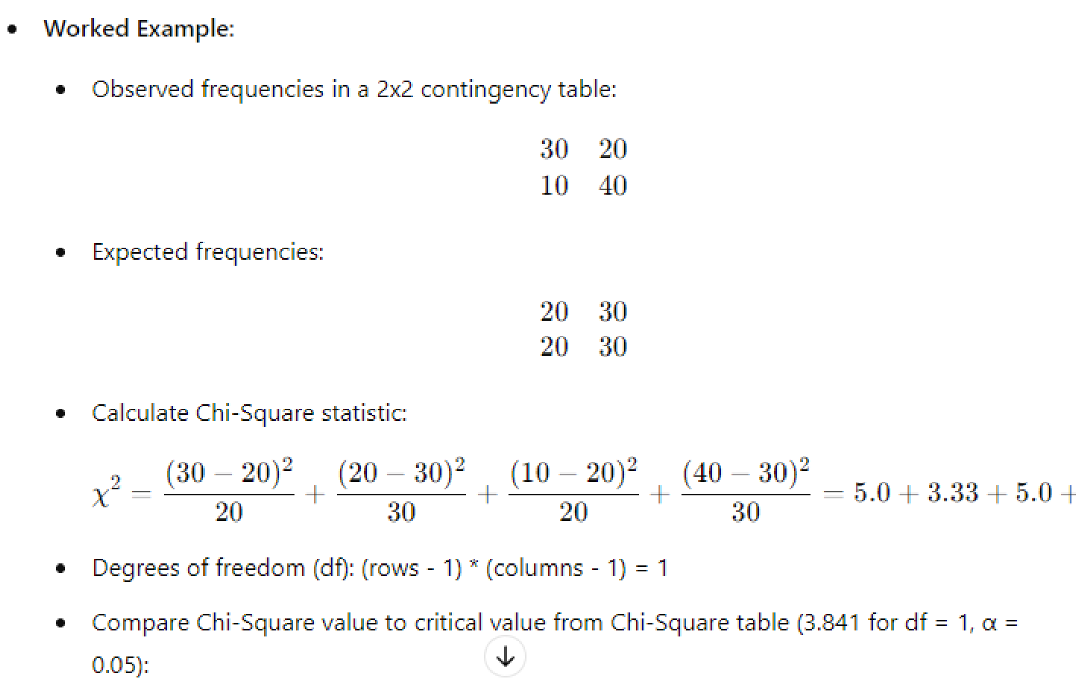
$$\chi^2 = \frac{(50-40)^2}{40} + \frac{(30-40)^2}{40} + \frac{(20-20)^2}{20} = 2.5 + 2.5 + 0 = 5.0$$

  - Degrees of freedom (df): Number of categories - 1 = 2
  - Compare Chi-Square value to critical value from Chi-Square table (5.991 for df = 2, α = 0.05):
    - Since 5.0 < 5.991, fail to reject H0.

  - Conclusion: There is no significant difference between the observed and expected distributions.

Chi-Square Test for Independence

- **Hypothesis Testing for Association Between Two Categorical Variables:**
  - Example Scenario: Testing if gender is independent of voting preference in an election.



- Worked Example:
  - Observed frequencies in a 2x2 contingency table:

$$\begin{matrix} 30 & 20 \\ 10 & 40 \end{matrix}$$

  - Expected frequencies:

$$\begin{matrix} 20 & 30 \\ 20 & 30 \end{matrix}$$

  - Calculate Chi-Square statistic:

$$\chi^2 = \frac{(30-20)^2}{20} + \frac{(20-30)^2}{30} + \frac{(10-20)^2}{20} + \frac{(40-30)^2}{30} = 5.0 + 3.33 + 5.0 +$$

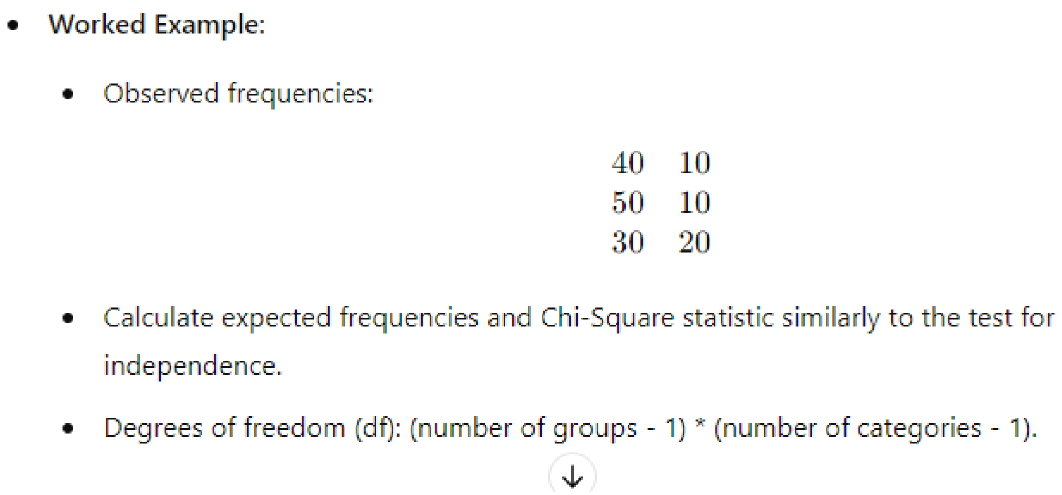
  - Degrees of freedom (df): (rows - 1) * (columns - 1) = 1
  - Compare Chi-Square value to critical value from Chi-Square table (3.841 for df = 1, α = 0.05):

    - Since 16.66 > 3.841, reject H0.

  - Conclusion: There is a significant association between gender and voting preference.

Chi-Square Test for Homogeneity

- **Hypothesis Testing for Comparing Proportions Across Groups:**
  - Example Scenario: Testing if different schools have the same proportion of students passing a standardized test.



- Worked Example:
  - Observed frequencies:

$$\begin{matrix} 40 & 10 \\ 50 & 10 \\ 30 & 20 \end{matrix}$$

  - Calculate expected frequencies and Chi-Square statistic similarly to the test for independence.
  - Degrees of freedom (df): (number of groups - 1) * (number of categories - 1).

8. Interpreting Chi-Square Test Results

- **P-values and Degrees of Freedom:**
  - **P-value:** The probability of observing the test statistic as extreme as, or more extreme than, the value observed under H0.
  - **Degrees of Freedom:** Determines the critical value from the Chi-Square distribution table.
- **Practical vs. Statistical Significance:**
  - Statistical significance does not always imply practical importance.
  - Consider the effect size and its real-world implications.

9. Common Applications of Chi-Square Tests

- **Real-world Examples:**
  - Testing if customer satisfaction is independent of store location.
  - Analyzing if the distribution of a genetic trait matches expected Mendelian ratios.
- **Case Studies:**
  - **Market Research:** Testing if purchasing preferences are independent of demographic factors.
  - **Healthcare:** Examining if the occurrence of a disease is independent of geographic region.

Recommended Readings:

- "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne
- "Introduction to the Practice of Statistics" by David S. Moore, George P. McCabe, and Bruce A. Craig
- "Fundamentals of Biostatistics" by Bernard Rosner

TOPIC 5: CORRELATION ANALYSIS

Course Description:

This course note provides a comprehensive understanding of correlation analysis, a statistical method used to measure the strength and direction of the relationship between two quantitative variables. It covers the types of correlation coefficients, methods for calculating them, and interpretation of the results.

Objectives:

- Understand the concept and purpose of correlation analysis.
- Learn the different types of correlation coefficients and their applications.
- Calculate correlation coefficients using different methods.
- Interpret the results of correlation analysis.
- Apply correlation analysis to real-world data scenarios.

Course Outline:

1. **Introduction to Correlation Analysis**
    - Definition and Purpose
    - Types of Correlation
2. **Types of Correlation Coefficients**
    - Pearson’s Correlation Coefficient
    - Spearman’s Rank Correlation Coefficient
    - Kendall’s Tau Correlation Coefficient
3. **Conditions for Using Correlation Analysis**
    - Linearity
    - Level of Measurement
    - Homoscedasticity
4. **Steps in Conducting Correlation Analysis**
    - Data Collection
    - Checking Assumptions
    - Calculating the Correlation Coefficient
    - Hypothesis Testing

    - Interpreting Results
5. **Pearson's Correlation Coefficient**
    - Definition and Formula
    - Calculation and Interpretation
    - Worked Examples
6. **Spearman's Rank Correlation Coefficient**
    - Definition and Formula
    - Calculation and Interpretation
    - Worked Examples
7. **Kendall's Tau Correlation Coefficient**
    - Definition and Formula
    - Calculation and Interpretation
    - Worked Examples
8. **Interpreting Correlation Coefficients**
    - Strength and Direction
    - Significance Testing
    - Practical vs. Statistical Significance
9. **Common Applications of Correlation Analysis**
    - Real-world Examples
    - Case Studies

Detailed Course Notes:

1. Introduction to Correlation Analysis

- **Definition and Purpose:**
    - **Correlation Analysis:** A statistical technique used to quantify the relationship between two variables.
    - Purpose: To determine the strength and direction of the relationship between variables and to identify patterns.
- **Types of Correlation:**
    - **Positive Correlation:** Both variables move in the same direction.

  - **Negative Correlation:** Variables move in opposite directions.
  - **No Correlation:** No discernible relationship between variables.

2. Types of Correlation Coefficients

- **Pearson’s Correlation Coefficient (r):**
  - Measures linear relationships between two continuous variables.
  - Values range from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.
- **Spearman’s Rank Correlation Coefficient (ρ or rs):**
  - Measures the strength and direction of the relationship between two ranked variables.
  - Suitable for ordinal data or non-linear relationships.
- **Kendall’s Tau Correlation Coefficient (τ):**
  - Measures the association between two variables based on the ranks of data.
  - Less sensitive to errors in data than Spearman's coefficient.

3. Conditions for Using Correlation Analysis

- **Linearity:**
  - Pearson’s correlation requires a linear relationship between variables.
  - Spearman’s and Kendall’s correlations do not require linearity.
- **Level of Measurement:**
  - Pearson’s: Interval or ratio data.
  - Spearman’s and Kendall’s: Ordinal, interval, or ratio data.
- **Homoscedasticity:**
  - The spread of one variable is consistent across the range of another variable.

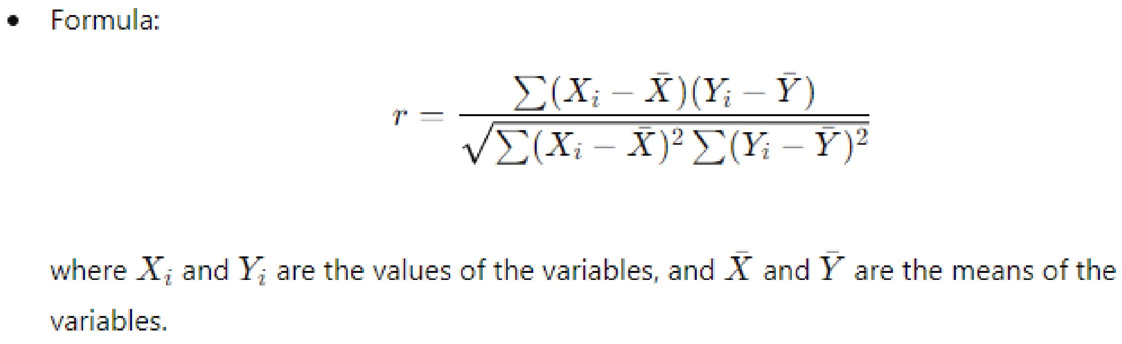
4. Steps in Conducting Correlation Analysis

- **Data Collection:**
  - Gather data for the variables of interest.
- **Checking Assumptions:**
  - Ensure the data meets the assumptions for the chosen correlation coefficient.

- **Calculating the Correlation Coefficient:**
  - Use appropriate formulas or statistical software.
- **Hypothesis Testing:**
  - Null hypothesis (H0): There is no correlation between the variables.
  - Alternative hypothesis (H1): There is a correlation between the variables.
- **Interpreting Results:**
  - Determine the strength and direction of the relationship.
  - Assess statistical significance.

5. Pearson's Correlation Coefficient

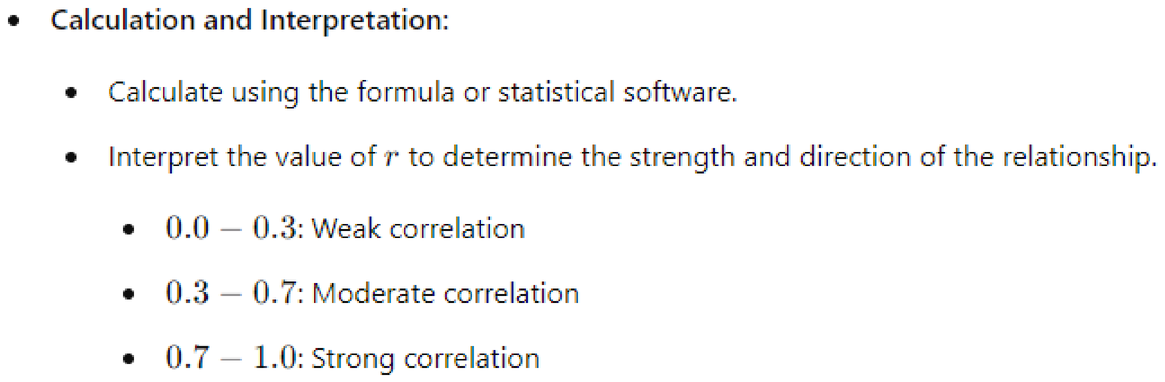
- **Definition and Formula:**
  - Measures the linear relationship between two continuous variables.



- Formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where X_i and Y_i are the values of the variables, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.

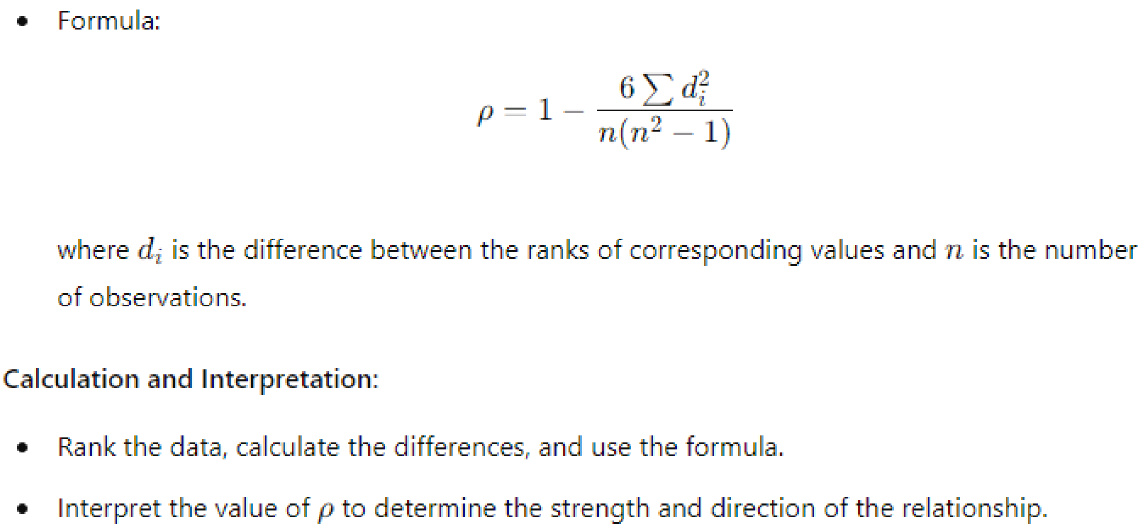


- Calculation and Interpretation:
  - Calculate using the formula or statistical software.
  - Interpret the value of r to determine the strength and direction of the relationship.
    - $0.0 - 0.3$: Weak correlation
    - $0.3 - 0.7$: Moderate correlation
    - $0.7 - 1.0$: Strong correlation

- **Worked Examples:**
  - Example Scenario: Analyzing the correlation between study hours and exam scores.
  - Given data for study hours and scores, calculate r and interpret the result.

6. Spearman's Rank Correlation Coefficient

- **Definition and Formula:**
  - Measures the strength and direction of the relationship between two ranked variables.



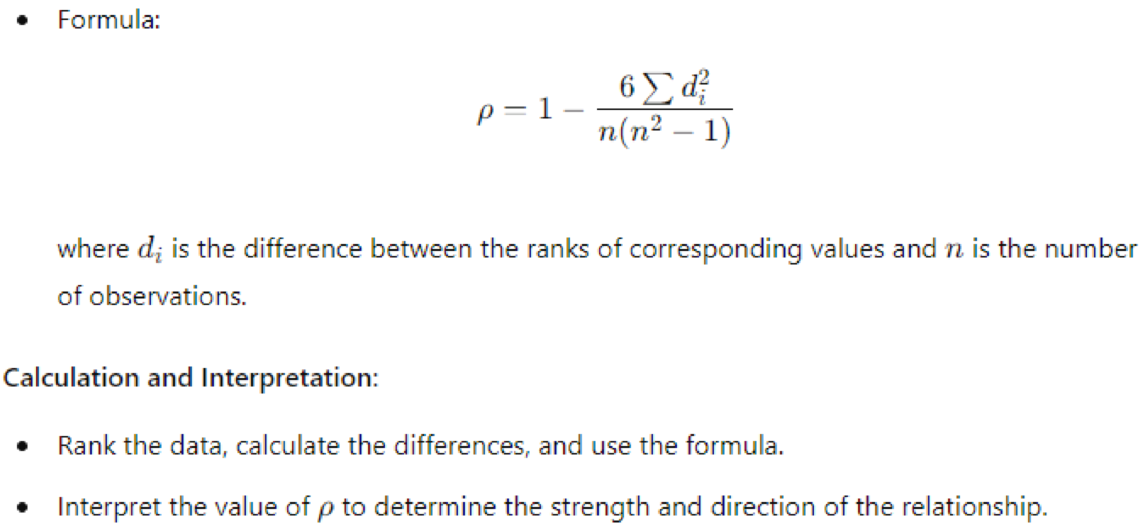
- Formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of corresponding values and n is the number of observations.

Calculation and Interpretation:

- Rank the data, calculate the differences, and use the formula.
- Interpret the value of ρ to determine the strength and direction of the relationship.



- Formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of corresponding values and n is the number of observations.

Calculation and Interpretation:

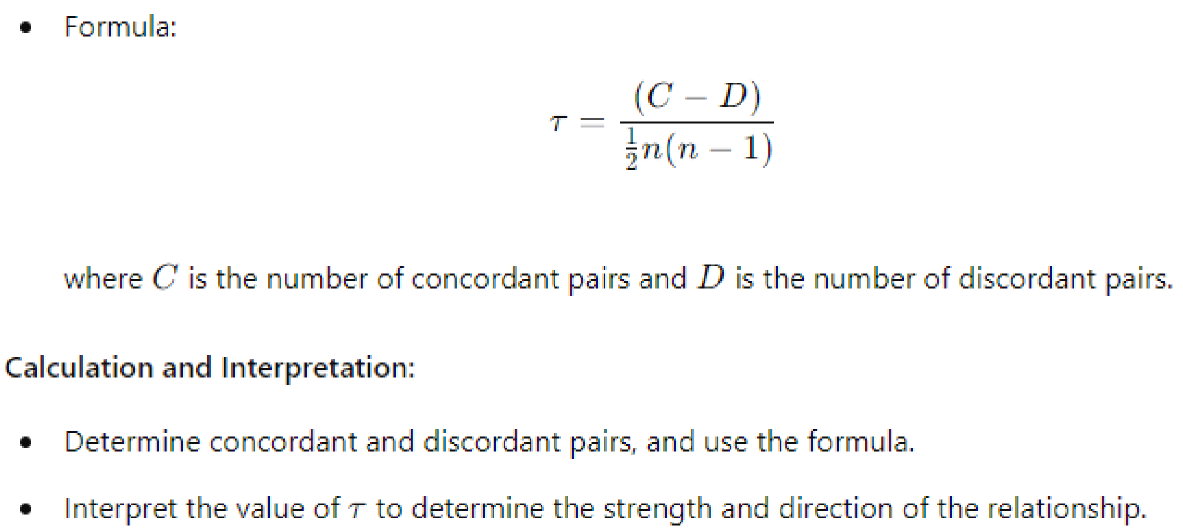
- Rank the data, calculate the differences, and use the formula.
- Interpret the value of ρ to determine the strength and direction of the relationship.

- **Worked Examples:**
  - Example Scenario: Analyzing the correlation between employee ranks and performance scores.
  - Given ranked data, calculate ρ and interpret the result.

7. Kendall's Tau Correlation Coefficient

- **Definition and Formula:**

- Measures the association between two variables based on the ranks of data.



- Formula:

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n-1)}$$

where C is the number of concordant pairs and D is the number of discordant pairs.

Calculation and Interpretation:

- Determine concordant and discordant pairs, and use the formula.
- Interpret the value of τ to determine the strength and direction of the relationship.

- **Worked Examples:**
  - Example Scenario: Analyzing the correlation between customer satisfaction ranks and loyalty ranks.
  - Given ranked data, calculate τ and interpret the result.

8. Interpreting Correlation Coefficients

- **Strength and Direction:**
  - Positive values indicate a positive relationship; negative values indicate a negative relationship.
  - The closer the value is to ±1, the stronger the relationship.
- **Significance Testing:**
  - Use p-values to determine if the correlation is statistically significant.
  - Null hypothesis: There is no correlation between the variables.
  - If p-value ≤ α, reject the null hypothesis.
- **Practical vs. Statistical Significance:**
  - Statistical significance does not always imply practical importance.
  - Consider the effect size and its real-world implications.

9. Common Applications of Correlation Analysis

- **Real-world Examples:**

    - Examining the relationship between marketing spend and sales revenue.
    - Analyzing the correlation between temperature and ice cream sales.
- **Case Studies:**
    - **Finance:** Studying the correlation between stock prices of two companies.
    - **Healthcare:** Investigating the correlation between BMI and blood pressure levels.

Recommended Readings:

- "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne
- "Introduction to the Practice of Statistics" by David S. Moore, George P. McCabe, and Bruce A. Craig
- "Fundamentals of Biostatistics" by Bernard Rosner

TOPIC 6: REGRESSION ANALYSIS

Course Description:

This course note provides a comprehensive understanding of regression analysis, a statistical method used to examine the relationship between a dependent variable and one or more independent variables. It covers simple linear regression, multiple linear regression, assumptions, interpretation, and applications.

Objectives:

- Understand the concept and purpose of regression analysis.
- Learn the different types of regression models and their applications.
- Conduct regression analysis using statistical software.
- Interpret the results of regression analysis.
- Apply regression analysis to real-world data scenarios.

Course Outline:

1. **Introduction to Regression Analysis**
    - Definition and Purpose
    - Types of Regression Models
2. **Simple Linear Regression**
    - Model Specification
    - Estimation of Parameters
    - Assumptions of Simple Linear Regression
3. **Multiple Linear Regression**
    - Model Specification
    - Estimation of Parameters
    - Assumptions of Multiple Linear Regression
4. **Assumptions of Regression Analysis**
    - Linearity
    - Independence
    - Homoscedasticity
    - Normality

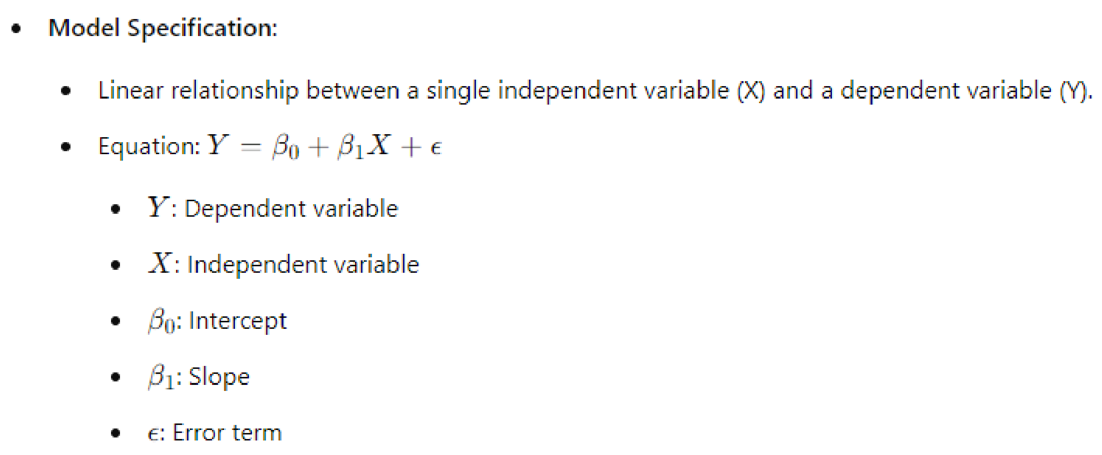
5. **Steps in Conducting Regression Analysis**
    - Data Collection
    - Model Specification
    - Estimation of Parameters
    - Hypothesis Testing
    - Model Diagnostics and Validation
6. **Interpreting Regression Results**
    - Coefficients and Significance
    - R-squared and Adjusted R-squared
    - Residual Analysis
    - Practical vs. Statistical Significance
7. **Common Applications of Regression Analysis**
    - Real-world Examples
    - Case Studies

Detailed Course Notes:

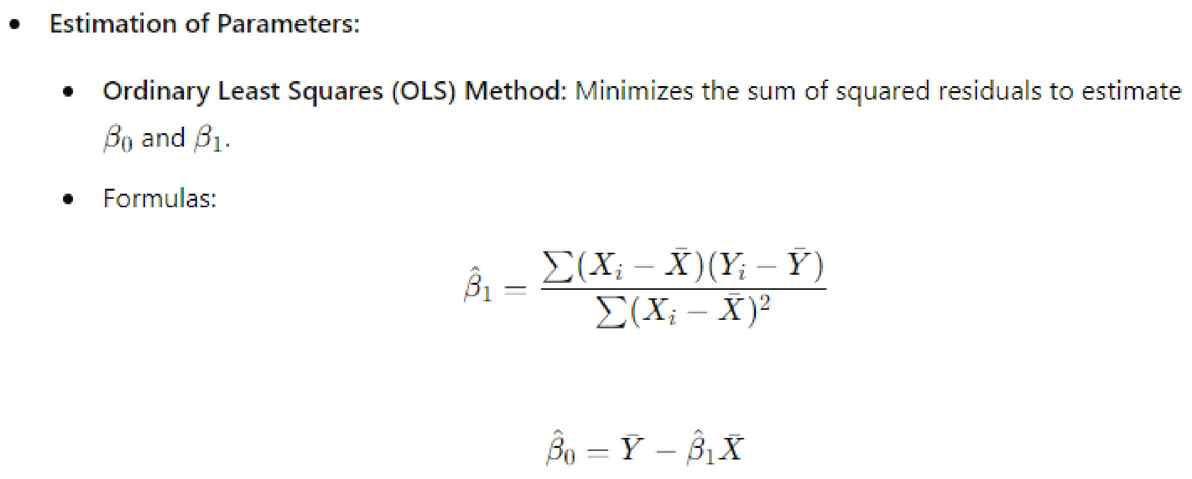
1. Introduction to Regression Analysis

- **Definition and Purpose:**
    - **Regression Analysis:** A statistical technique used to examine the relationship between a dependent variable and one or more independent variables.
    - Purpose: To predict the value of the dependent variable based on the values of the independent variables and to understand the relationships between variables.
- **Types of Regression Models:**
    - **Simple Linear Regression:** One independent variable.
    - **Multiple Linear Regression:** More than one independent variable.
    - **Logistic Regression:** Dependent variable is categorical.
    - **Polynomial Regression:** Non-linear relationship between variables.

2. Simple Linear Regression



- **Model Specification:**
  - Linear relationship between a single independent variable (X) and a dependent variable (Y).
  - Equation: $Y = \beta_0 + \beta_1 X + \epsilon$
    - Y: Dependent variable
    - X: Independent variable
    - β_0: Intercept
    - β_1: Slope
    - ϵ: Error term



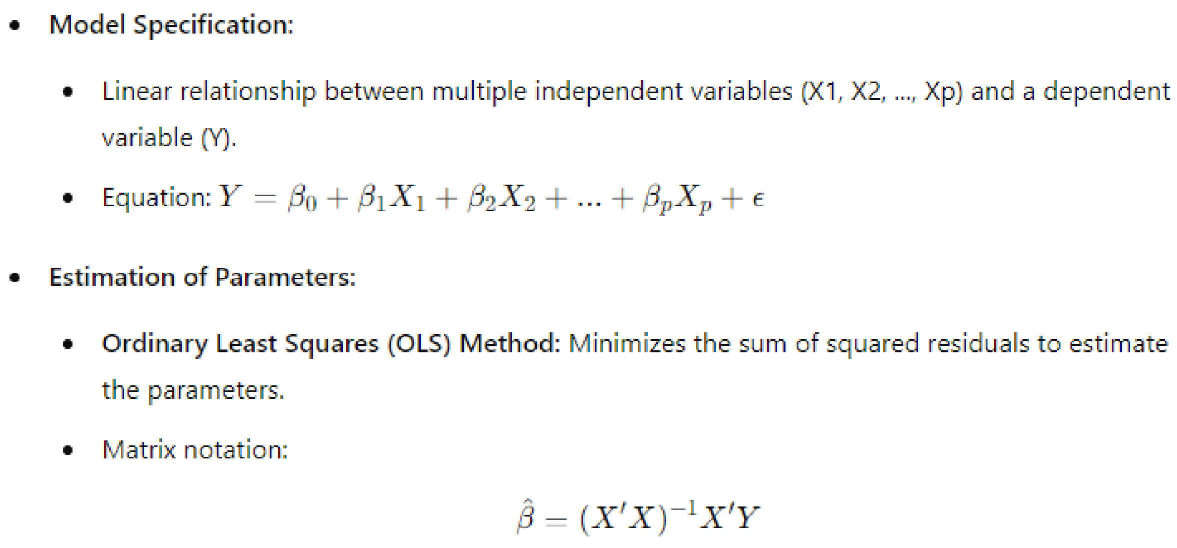
- **Estimation of Parameters:**
  - **Ordinary Least Squares (OLS) Method:** Minimizes the sum of squared residuals to estimate β_0 and β_1.
  - Formulas:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- **Assumptions of Simple Linear Regression:**
  - Linearity: The relationship between X and Y is linear.
  - Independence: Observations are independent.
  - Homoscedasticity: Constant variance of errors.
  - Normality: Errors are normally distributed.

3. Multiple Linear Regression



- Model Specification:
  - Linear relationship between multiple independent variables (X1, X2, ..., Xp) and a dependent variable (Y).
  - Equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$
- Estimation of Parameters:
  - **Ordinary Least Squares (OLS) Method:** Minimizes the sum of squared residuals to estimate the parameters.
  - Matrix notation:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- **Assumptions of Multiple Linear Regression:**
  - Linearity: The relationship between each X and Y is linear.
  - Independence: Observations are independent.
  - Homoscedasticity: Constant variance of errors.
  - Normality: Errors are normally distributed.
  - No multicollinearity: Independent variables are not highly correlated.

4. Assumptions of Regression Analysis

- **Linearity:** The relationship between the independent and dependent variables is linear.
- **Independence:** Observations of the dependent variable are independent of each other.
- **Homoscedasticity:** The variance of the error terms is constant across all levels of the independent variables.
- **Normality:** The error terms are normally distributed.

5. Steps in Conducting Regression Analysis

- **Data Collection:**
  - Gather data for the dependent and independent variables.
- **Model Specification:**
  - Define the regression model to be used based on the data and research question.
- **Estimation of Parameters:**
  - Use statistical software to estimate the parameters of the regression model.

- **Hypothesis Testing:**
    - Null hypothesis (H0): No relationship between the independent and dependent variables.
    - Alternative hypothesis (H1): There is a relationship between the independent and dependent variables.
    - Use t-tests for individual coefficients and F-tests for the overall model.
- **Model Diagnostics and Validation:**
    - Check for violations of regression assumptions.
    - Use diagnostic plots (e.g., residual plots) and statistical tests (e.g., Breusch-Pagan test for homoscedasticity).

6. Interpreting Regression Results

- **Coefficients and Significance:**
    - Coefficients (β\betaβ): Indicate the direction and strength of the relationship between each independent variable and the dependent variable.
    - p-values: Assess the statistical significance of each coefficient.
- **R-squared and Adjusted R-squared:**
    - **R-squared (R^2):** Proportion of variance in the dependent variable explained by the independent variables.
    - **Adjusted R-squared:** Adjusts R^2 for the number of predictors in the model.
- **Residual Analysis:**
    - Analyze residuals to check for patterns that might indicate violations of regression assumptions.
- **Practical vs. Statistical Significance:**
    - Statistical significance does not always imply practical importance.
    - Consider the effect size and its real-world implications.

7. Common Applications of Regression Analysis

- **Real-world Examples:**
    - Predicting house prices based on features like size, location, and number of bedrooms.
    - Analyzing the impact of advertising spend on sales revenue.
- **Case Studies:**

- **Economics:** Studying the relationship between GDP growth and unemployment rates.
- **Healthcare:** Investigating the impact of lifestyle factors on health outcomes.

Recommended Readings:

- "Applied Regression Analysis" by Norman R. Draper and Harry Smith
- "Regression Analysis by Example" by Samprit Chatterjee and Ali S. Hadi
- "An Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani