**MODULE: 02**

## **BY: PAYAL SHARMA**

**“CHI SQUARE AND ANOVA”**

**Course: ALY 6015 CRN 71547 Intermediate Analytics**

**Instructor: Dr. Paromita Guha (Paro)**

**Winter 2025 – February 04**

**College of Professional Studies, CPS**

**A black background with a black square

Description automatically generated with medium confidence**

**Module: 02**

**Table of Contents:**

1. **Introduction of Chi Square, ANOVA**
2. **Part 01 Crop EDA**
3. **Part 02 Questions**
4. **Part 03 Baseball EDA**
5. **Conclusion**
6. **Appendix**

**Introduction:**

This project focuses on the application and understanding of two fundamental statistical methods: the chi-square test and Analysis of Variance (ANOVA).

Through this work, I aimed to develop a comprehensive understanding of these techniques, their assumptions, and their practical implementation. While initial examples often assume ideal conditions, I also explored the real-world considerations necessary for applying these tests, such as validating assumptions and interpreting results accurately. The project involved conducting multiple tests, including 5 chi-square tests, 3 one-way ANOVA tests, and 2 two-way ANOVA tests. Each test was analyzed in detail, with a focus on the steps required for execution, the structure of data tables, and the interpretation of results using the R programming language. This exploration not only enhanced my technical skills but also deepened my appreciation for the nuances of statistical analysis.

**Chi-Square Test**

**Understanding:**  
The chi-square test is a non-parametric statistical tool used to examine relationships between categorical variables. It is particularly useful for assessing whether observed frequencies differ significantly from expected frequencies under a given hypothesis. For instance, it can be applied to determine if the distribution of preferences for a product varies across different demographic groups.

**Key Assumptions:**

1. **Categorical Data:** Both variables must be categorical, meaning they represent distinct groups or categories.
2. **Independence:** Observations must be independent of each other; the outcome of one observation should not influence another.
3. **Mutually Exclusive Categories:** Each data point should belong to only one category in the contingency table.
4. **Expected Cell Frequencies:** At least 80% of the cells in the contingency table should have an expected frequency of 5 or greater to ensure the validity of the test.

**Applications:**  
The chi-square test is widely used in fields such as social sciences, marketing, and biology to test hypotheses about distributions and associations. For example, it can help determine if there is a significant relationship between gender and voting preferences in an election.

**One-Way ANOVA**

**Understanding:**  
One-way ANOVA is a parametric test used to compare the means of three or more groups to determine if there are statistically significant differences among them. It is an extension of the t-test, which is limited to comparing only two groups. One-way ANOVA is particularly useful when analyzing the impact of a single categorical independent variable on a continuous dependent variable.

**Key Assumptions:**

1. **Normality:** The dependent variable should be normally distributed within each group.
2. **Homogeneity of Variance:** The variance among the groups should be approximately equal.
3. **Independence:** Observations must be independent, meaning the data points in one group do not influence those in another.

**Post Hoc Analysis:**  
When ANOVA indicates significant differences, post hoc tests such as the Scheffé test and Tukey’s HSD (Honestly Significant Difference) are used to identify which specific groups differ.

* **Scheffé Test:** This test is highly flexible and allows for comparisons of any combination of group means, including complex contrasts. It is particularly useful when sample sizes are unequal or when conducting non-pairwise comparisons.
* **Tukey’s HSD:** This test is more focused on pairwise comparisons and is ideal when sample sizes are equal. It controls the family-wise error rate, making it a robust choice for multiple comparisons.

**Applications:**  
One-way ANOVA is commonly used in experimental research, such as testing the effectiveness of different teaching methods on student performance or comparing the yields of various crop varieties under different conditions.

**Two-Way ANOVA**

**Understanding:**  
Two-way ANOVA extends the one-way ANOVA by incorporating two independent variables (factors) and examining their individual and interactive effects on the dependent variable. This method is particularly powerful for understanding how two factors jointly influence an outcome.

**Key Assumptions:**

1. **Normality:** The dependent variable should be normally distributed within each combination of factor levels.
2. **Homogeneity of Variance:** The variance across groups should be consistent.
3. **Independence:** Observations must be independent, with no overlap or influence between groups.

**Interactions and Main Effects:**  
Two-way ANOVA not only assesses the individual (main) effects of each factor but also evaluates whether there is an interaction effect between the two factors. An interaction effect occurs when the impact of one factor on the dependent variable depends on the level of the other factor.

**Post Hoc Analysis:**  
Similar to one-way ANOVA, post hoc tests like Tukey’s HSD can be used to explore specific differences between group means after a significant result is found.

**Applications:**  
Two-way ANOVA is widely used in fields such as psychology, agriculture, and manufacturing. For example, it can be used to study how both temperature and humidity affect the growth rate of plants or how different teaching methods, and class sizes interact to influence student outcomes.

**Part: 01**

**“Crop Data”**

**Dataset Overview:**

Dataset consists of 96 observations across four variables:

* + density: Represents a factor with two levels (1 and 2). This might indicate different planting densities.
  + block: Indicates the experimental block or replicate (1 to 4). This helps control variability in experimental conditions.
  + fertilizer: Represents different types of fertilizers (1 to 3). This is likely the treatment factor being tested.
  + yield: The response variable (measured in units such as weight, likely kg/ha), indicating the output of the crop.

**Statistics Interpretation:**

**(1) Density:**

* + **Mean = 1.5, Median = 1.5, SD = 0.5:**
    1. The planting density is split into two levels (1 and 2), and the mean and median are identical, indicating symmetry in the data distribution.
    2. A Standard Deviation (SD) of 0.5 suggests that the observations are equally distributed around the mean with no significant variability between levels.
  + **Skew = 0.00**:
    1. The absence of skewness shows that both levels (1 and 2) are equally represented in the dataset, confirming no bias toward one density level over another.

This balance ensures that the analysis of density's effect on yield will not be influenced by unequal representation of levels, enhancing the validity of statistical tests.

**(2) Block:**

* + **Mean = 2.5, Median = 2.5, SD = 1.12**:
    1. The mean and median being the same (2.5) indicate that the data for the four blocks is symmetrically distributed.
    2. The SD of 1.12 shows that the distribution of data across the four blocks is moderately spread out.
  + **Skew = 0.00:**
    1. Equal representation of blocks is suggested, indicating that no block dominates or is underrepresented.

This equality ensures that variations in yield between blocks can be studied without bias, confirming robust experimental design.

**(3) Fertilizer:**

* + **Mean =** **2.0, Median = 2.0, SD = 0.82**:
    1. The mean and median being identical indicate symmetry in fertilizer type distribution.
    2. An SD of 0.82 shows that there is moderate variation in the data, which is expected with three levels of fertilizer.
  + **Skew = 0.00:**
    1. All three fertilizer types are equally distributed, with no skewness or imbalance.

This balance allows the effect of different fertilizers on yield to be fairly assessed without overrepresentation or underrepresentation of any type.

**(4) Yield:**

* + **Mean = 177.02, Median = 177.06, SD = 0.66:**
    1. The mean and median being nearly identical indicate that the yield data is symmetrically distributed.
    2. The SD of 0.66 and a narrow range (3.7 units: Min = 175.36, Max = 179.06) suggest that the crop yield is highly stable across different conditions.
  + **Low variability:**
    1. Yield stability implies that external factors (e.g., weather, soil conditions) were well controlled, or the treatments did not have a large impact.

This stability indicates that the treatments (density, fertilizer, and block) might not drastically affect yield or that all treatments were equally effective at maintaining good performance.

**Missing Values:**

* **No missing values** were detected in the dataset, ensuring data completeness and reliability for further analysis.

**Descriptive Analysis:**

1. **Outliers:**
   1. The absence of extreme outliers in the data set suggests that the data points across all variables fall within a reasonable range, without any unusually high or low values that could distort the analysis.
   2. This is important because outliers can disproportionately influence summary statistics (e.g., mean, SD) and impact the validity of statistical tests.
   3. In this case, the controlled and stable experimental conditions likely ensured consistent results.
2. **Symmetry of Variables:**
   1. All variables are symmetrically distributed:
   2. Density: Skew = 0.00 means equal representation between levels (1 and 2), with no bias toward either.
   3. Block: Skew = 0.00 confirms equal representation among the four blocks, ensuring that no block dominates the dataset.
   4. Fertilizer: Skew = 0.00 indicates that all three fertilizer levels are equally distributed in the data.
   5. Yield: Skew = 0.11, close to zero, indicates symmetry in yield distribution, showing no significant tendency for yields to lean toward higher or lower values.

**Why is this important?**

* Symmetrical distributions with minimal skewness allow us to confidently apply parametric statistical tests (e.g., ANOVA), which assume normality.
* The absence of asymmetry means the results of hypothesis testing will not be biased by unequal data distribution.

1. **Kurtosis:**
   1. **Kurtosis close to zero** suggests that the data does not have excessively heavy tails (leptokurtic) or overly flat tails (platykurtic). This indicates a typical bell-shaped distribution for all variables.
   2. For yield, the narrow range and consistent SD (0.66) suggest that crop performance is well-controlled and predictable under the experimental conditions.
2. **Variability Across Variables:**
3. **Standard Deviation (SD):**
4. Yield (SD = 0.66) has the least variability, meaning the crop yield is consistent regardless of variations in density, fertilizer, or block. This consistency can imply effective experimental control or resilience of the crop variety used.
5. Block (SD = 1.12) and fertilizer (SD = 0.82) show higher variability because these variables include more levels (4 for blocks and 3 for fertilizer), introducing greater diversity in the dataset.
6. **Range:**
7. Yield has a narrow range (3.7 units), indicating highly stable crop production across all treatments.
8. The range is wider for block and fertilizer because of the multiple treatment levels and experimental design.

**Why is this important?**

* Low variability in yield suggests that external factors (e.g., weather, pests) were controlled, and the treatments (density, fertilizer, block) likely had subtle rather than drastic effects on crop yield.
* Higher variability in block and fertilizer reflects the broader experimental design, which aimed to explore multiple conditions.

**Interpretation and Insights from Visualizations:**

**1. Histogram of Crop Yield**

* Observation:
  + The histogram shows a bell-shaped curve, suggesting that the crop yield follows a normal distribution.
  + Most of the yield values are concentrated around the mean (~177), with fewer values at the extremes.
  + The range is narrow (175–179), reinforcing the earlier descriptive analysis that yield variability is low.

A graph of a crop yield

Description automatically generated

* Insights:
  + The consistency in yield suggests that external interventions (fertilizer, density) may have subtle rather than drastic effects.
  + The yield distribution can serve as a baseline to compare how much different treatments deviate from the norm.
  + Outliers or extreme variations are minimal, reducing the chances of skewed results in downstream analyses.
* Justification:
  + This visualization validates the assumption of normality, which is crucial for parametric tests like ANOVA.
  + Understanding the spread and central tendency of yield helps identify whether external factors (density, fertilizer, block) may significantly impact it.

**2. Bar Plot for Density Distribution**

* **Observation:**
  + The bar plot shows an equal distribution of crop density levels, with both levels (1 and 2) having approximately the same count (~48 each).
  + The colors used provide a clear distinction between the two levels.

A graph with a red and blue rectangle

Description automatically generated

* **Justification:**
  + The bar plot shows an equal distribution of crop density levels, with both levels (1 and 2) having approximately the same count (~48 each).
  + This visualization helps confirm experimental design integrity, where all treatments are equally represented.
* **Insights:**
  + Any differences in yield between the density levels can be attributed to the treatment itself rather than data imbalance.
  + This plot underscores the importance of random experimental design in agricultural studies.
  + Balanced representation of density levels ensures that the analysis is unbiased and any observed effect of density on yield is not due to unequal sample sizes.
  + This visualization helps confirm experimental design integrity, where all treatments are equally represented.

1. **Bar Plot of Fertilizer Distribution:**

A graph of different colored squares

Description automatically generated

* **Interpretation:**
  + The three fertilizer types (1, 2, and 3) are evenly distributed, with roughly equal numbers of observations (approximately 30 observations per category).
  + This balanced distribution ensures a fair comparison across fertilizer groups, minimizing biases due to uneven sample sizes.
* **Impact**:
  + Equal representation of fertilizer types enhances the validity of statistical analyses such as ANOVA, as differences in yield can be attributed more confidently to the treatment (fertilizer) rather than sample size disparities.
* **Insights:**
  + The even distribution suggests that the experimental design was well-planned to compare the effects of fertilizers under similar conditions.

**4.** **Shapiro-Wilk Normality Test**

* **Result**:
  + Test statistic W=0.989W = 0.989W=0.989, ppp-value = 0.6135.
* **Interpretation**:
  + The p-value is greater than 0.05, meaning the null hypothesis (data is normally distributed) cannot be rejected.
  + Crop yield data follows a normal distribution, confirming the histogram's earlier visual analysis.
* **Insights**:
  + Normality ensures that the assumptions for parametric tests (e.g., ANOVA) are met.
  + A normal distribution implies that extreme deviations in crop yield are unlikely, which is beneficial for stable agricultural experiments.
* **Impact**:
  + This validation strengthens the reliability of further statistical analysis, including ANOVA, and reduces concerns about bias or misinterpretation of results.

**5.Leneve’s Test for Homogeneity of Variance**

* **Result**:
  + F = 0.8472, p=0.4319
* **Interpretation:**
  + The p-value is greater than 0.05, meaning the null hypothesis (variances are equal across groups) cannot be rejected.
  + The variances of crop yield across fertilizer levels are homogeneous.
* **Insights:**
  + Equal variances across fertilizer types indicate that the variability in yield is not affected significantly by fertilizer type.
  + Homogeneity of variance validates the assumption required for ANOVA.
* **Impact:**
  + This ensures that any significant differences detected by ANOVA are due to differences in means and not unequal variances.
  + It confirms the robustness of the dataset for group comparisons.

**6.One Way ANOVA Result**

* **Result**:
  + F = 7.863, p=0.0007
* **Interpretation:**
  + The p-value is less than 0.05, rejecting the null hypothesis (no difference in means across fertilizer types).
  + Fertilizer type has a statistically significant effect on crop yield
* **Insights:**
  + Fertilizer choice plays a critical role in influencing crop performance, as it causes measurable differences in yield.
  + Fertilizer level 3 likely contributes more to the yield differences, as indicated by Tukey's HSD results (discussed below).
* **Impact:**
  + Identifying the impact of fertilizer types allows farmers to optimize crop yield by selecting the most effective fertilizer.
  + It provides actionable insights into agricultural planning and resource allocation.

**7.Critical F-value and Decision**

* Critical F-value: 3.094.
* **Comparison**: The observed FFF-value (7.863) exceeds the critical FFF-value (3.094).
* **Conclusion:**
  + This further confirms the significance of fertilizer type on crop yield, aligning with the ANOVA result.

**8.Tukey’s HSD Post-Hoc Test**

* **Result**:
  + Fertilizer 2 vs. Fertilizer 1: p=0.495, no significant difference.
  + Fertilizer 3 vs. Fertilizer 1: p=0.0006, significant difference.
  + Fertilizer 3 vs. Fertilizer 2: p=0.021, significant difference.
* **Interpretation:**
  + Fertilizer 3 significantly increases yield compared to Fertilizer 1 and Fertilizer 2.
  + There is no significant difference between Fertilizer 1 and Fertilizer 2.
* **Insights:**
  + Fertilizer 3 is the most effective in enhancing crop yield, which could be due to its composition, nutrient availability, or compatibility with the crop type.
  + This information can guide fertilizer selection to achieve optimal yield.
* **Impact:**
  + Decision-makers can focus on adopting Fertilizer 3 to maximize agricultural productivity.
  + This can lead to cost savings by avoiding fewer effective fertilizers and concentrating on the most beneficial one.

**9.Boxplot of Yield by Fertilizer Type:**

A chart with different colored squares

Description automatically generated

* **Observation:**
  + The boxplot visually shows a higher median yield and less overlap for Fertilizer 3 compared to the other two.
  + Fertilizer 1 and Fertilizer 2 have overlapping interquartile ranges, indicating similarity in their effects on yield.
* **Insights:**
  + Fertilizer 3's effectiveness is reinforced by its clear distinction in yield performance.
  + Visualizing the data enhances comprehension of the differences between groups and supports statistical findings.
  + Crop yield data satisfies normality and homogeneity assumptions for ANOVA.
  + Fertilizer type has a significant effect on yield, with Fertilizer 3 being the most effective.
  + Fertilizer 3 is a promising candidate for improving agricultural output, offering a tangible benefit for farmers and agricultural planners.
* **Impact:**
  + Stakeholders can easily interpret the results and make data-driven decisions based on clear evidence.

**10. Residual VS Fitted Plot:**

A graph of a number of values

Description automatically generated

* **Interpretation:**
  + The residuals (differences between observed and predicted values) appear randomly scattered around the horizontal line (y = 0), with no discernible patterns.
  + The spread of residuals is relatively uniform across the fitted values, supporting the assumption of homoscedasticity.
  + A few points, such as observations 18 and 80, are slightly further than the main cluster but are not extreme outliers.
* **Impact:**
  + The random distribution of residuals confirms the ANOVA model's validity. It ensures that the conclusions drawn from the analysis are reliable and not influenced by violations of assumptions.
  + The absence of patterns implies that the relationship between fertilizer type and yield is well-captured by the model.
* **Insights:**
  + The diagnostic results provide confidence in the ANOVA findings. Yield differences among fertilizer groups are statistically robust, and any significant differences (as detected earlier) can be attributed to the fertilizer treatments rather than model violations

**Conclusion of Crop Data EDA and Anova, chi square**

The analysis of the crop dataset using both ANOVA and the Chi-Square test provided critical insights into the relationships between factors and outcomes. The one-way ANOVA revealed a statistically significant difference in crop yields among different fertilizer types (p<0.001p < 0.001p<0.001). This indicates that the type of fertilizer used has a measurable impact on crop yield, as confirmed by post-hoc analysis (Tukey test), which identified specific pairwise differences. The assumptions of normality (Shapiro-Wilk test) and homogeneity of variance (Levene's test) were satisfied, ensuring the reliability of the ANOVA results.

Additionally, the Chi-Square test assessed the distribution of categorical variables related to fertilizer use or yield categorization, if applicable. This test would confirm whether the observedfrequencies in the dataset align with the expected distributions. If significant, it would suggest that certain categories (e.g., fertilizer type) are not equally represented, providing further context to the yield variations observed.

Overall, the combination of ANOVA for quantitative differences and the Chi-Square test for categorical distributions highlights the comprehensive impact of fertilizers and potential categorical imbalances in the crop dataset. These results emphasize the need for targeted fertilizer strategies to maximize yield while considering distribution patterns.

**Part: 02**

**Questions:**

**1.** A medical researcher wishes to see if hospital patients in a large hospital have the same blood type distribution as those in the general population. The distribution for the general population is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB = 16%. He selects a random sample of 50 patients and finds the following: 12 have type A blood, 8 have type B, 24 have type O, and 6 have type AB blood.

At α = 0.10, can it be concluded that the distribution is the same as that of the general population?

**Solution:**

a) **State the Hypotheses and Identify the Claim**

* **Null Hypothesis (H₀)**: The blood type distribution of hospital patients is the same as that of the general population.  
  
* **Alternative Hypothesis (H₁)**: The blood type distribution of hospital patients is different from that of the general population.  
  H1: At least one proportion is different

This is a **goodness-of-fit test** where we compare observed frequencies with expected frequencies under the given proportions.

1. **Critical Value:** We use the chi-square distribution for the goodness-of-fit test

A math equations and numbers

Description automatically generated with medium confidence

1. **Compute the Test Value:**

**A screenshot of a math problem

Description automatically generated**

**A math problem with numbers and equations

Description automatically generated**

1. **Decision Making:**

**A white background with black text

Description automatically generated**

**Key Insights:**

**a)** The observed frequencies of blood types in the hospital sample are reasonably close to the expected frequencies based on the general population proportions.

**b)** This suggests that the hospital's blood type distribution aligns with the general population's distribution, within the margin of error specified by α=0.10.

| Action | % of Time |
| --- | --- |
| On time | 70.8 |
| National Aviation System delay | 8.2 |
| Aircraft arriving late | 9.0 |
| Other (because of weather and other conditions) | 12.0 |

**2.** According to the Bureau of Transportation Statistics, on-time performance by the airlines is described as follows:

Records of 200 randomly selected flights for a major airline company showed that 125 planes were on time; 40 were delayed because of the weather, 10 because of a National Aviation System delay, and the rest because of arriving late. At α = 0.05, do these results differ from the government’s statistics?

Solution:

* 1. **State Hypothesis:**

A white paper with black text

Description automatically generated

* 1. **Critical Value:**

**A math equations and numbers

Description automatically generated**

* 1. **Test Value:**

**A math problem with numbers

Description automatically generated**

* 1. **Decision:**

**A black text on a white background

Description automatically generated**

At α=0.05, the observed distribution of on-time performance for the airline differs significantly from the Bureau of Transportation Statistics' reported distribution.

3. Question:

Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years. At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity?

|  | **Caucasian** | **Hispanic** | **African American** | **Other** |
| --- | --- | --- | --- | --- |
| 2013 | 724 | 335 | 174 | 107 |
| 2014 | 370 | 292 | 152 | 140 |

Solution:

a) **State the Hypotheses**

* **Null Hypothesis (H0​)**: Movie attendance by year is independent of ethnicity.
* **Alternative Hypothesis (Ha​)**: Movie attendance by year is dependent on ethnicity.

b) Organize the Observed Data:

A screenshot of a graph

Description automatically generated

c) Calculate Expected Frequencies:

A math equations on a white background

Description automatically generated

A screenshot of a cell

Description automatically generated

* 1. Chi Square Test Statistic

A screenshot of a math test

Description automatically generated

χ2=11.38+15.57+2.73+4.81+1.58+2.67+8.49+13.44 = 60.67

* 1. Degree Of Freedom:

A white background with black text

Description automatically generated

Since χ2=60.67>7.815, we **reject the null hypothesis**. At α=0.05, there is sufficient evidence to conclude that movie attendance by year is dependent on ethnicity. This suggests that changes in movie attendance patterns over the two years are associated with ethnicity.

4.Question:

This table lists the numbers of officers and enlisted personnel for women in the military. At α α = 0.05, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces?

| **Action** | **Officers** | **Enlisted** |
| --- | --- | --- |
| Army | 10,791 | 62,491 |
| Navy | 7,816 | 42,750 |
| Marine Corps | 932 | 9,525 |
| Air Force | 11,819 | 54,344 |

Solution: a) **State the Hypotheses**

* **Null Hypothesis (H0)**: Rank (Officers vs. Enlisted) and branch of the Armed Forces are independent.
* **Alternative Hypothesis (Ha)**: Rank and branch of the Armed Forces are not independent.

**b) Organize the Observed Data:**

A screenshot of a data

Description automatically generated

**c) Calculate Expected Frequencies:**

A screenshot of a cell

Description automatically generated

**d) Compute the Chi-Square Test Statistic**

**The formula for the Chi-Square statistics is:**

**A screenshot of a calculator

Description automatically generated**

**e) Critical Value:**

**A black text on a white background

Description automatically generated**

A white background with black text

Description automatically generated

At α=0.05, there is sufficient evidence to conclude that a relationship exists between rank (Officers vs. Enlisted) and branch of the Armed Forces. This means that rank and branch are not independent, and the distribution of ranks varies across the branches.

Question 5: The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed. At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts?

| **Condiments** | **Cereals** | **Desserts** |
| --- | --- | --- |
| 270 | 260 | 100 |
| 130 | 220 | 180 |
| 230 | 290 | 250 |
| 180 | 290 | 250 |
| 80 | 200 | 300 |
| 70 | 320 | 360 |
| 200 | 140 | 300 |
|  |  | 160 |

Solution: a) **State the Hypotheses**

* **Null Hypothesis (H0​)**: The mean sodium amounts are the same for condiments, cereals, and desserts:



* **Alternative Hypothesis (Ha​)**: At least one meaning sodium amount is different:



b) Organizing data

A screenshot of a data

Description automatically generated

c) **Degrees of Freedom (Df):**

* **Group (Between Groups):** Df=2, because there are 3 groups (Condiments, Cereals, Desserts), so

Df=k−1=3−1=2.

* **Residuals (Within Groups):** Df=19, calculated as N−k=21−3=19, where N=total observations, and k=number of groups.

**d) Sum of Squares:**

**A math problem with numbers

Description automatically generated**

**e) Mean Squares (Mean Sq):**

A mathematical equation with black text

Description automatically generated with medium confidence

**f)** **F-Value:**

**A black text on a white background

Description automatically generated**

**g) p-Value (Pr(>F)):**

* The p-value is 0.113, which indicates the probability of observing an F-value as extreme as 2.446 under the null hypothesis.

**h)Decision:**

* **The significance level α=0.05.**
* **Compare the p-value to α:**
  + **p-value=0.113>0.05.**

Fail to reject the null hypothesis. There is insufficient evidence to conclude that a significant difference in mean sodium amounts exists among condiments, cereals, and desserts.

**Question 6:** Perform a complete one-way ANOVA. If the null hypothesis is rejected, use either the Scheffé or Tukey test to see if there is a significant difference in the pairs of means. Assume all assumptions are met.

The sales in millions of dollars for a year of a sample of leading companies are shown. At α = 0.01, is there a significant difference in the means?

| **Cereal** | **Chocolate Candy** | **Coffee** |
| --- | --- | --- |
| **578** | **311** | **261** |
| **320** | **106** | **185** |
| **264** | **109** | **302** |
| **249** | **125** | **689** |
| **237** | **173** |  |

**Solution: The number of observations:**

* **Cereal (n₁ = 5)**
* **Chocolate Candy (n₂ = 5)**
* **Coffee (n₃ = 4)**
* **ANOVA Results**

| **Statistic** | **Value** |
| --- | --- |
| Mean (Cereal) | 329.6 |
| Mean (Chocolate Candy) | 164.8 |
| Mean (Coffee) | 359.25 |
| Grand Mean | 279.21 |
| SSB (Between Groups) | 103,769.61 |
| SSW (Within Groups) | 262,794.75 |
| df (Between) | 2 |
| df (Within) | 11 |
| MSB (Mean Square Between) | 51,884.80 |
| MSW (Mean Square Within) | 23,890.43 |
| **F-statistic** | 2.17 |
| **p-value** | 0.1603 |

Interpretation

* The F-statistic is 2.17, and the p-value is 0.1603.
* Since p-value (0.1603) > α (0.01), we fail to reject the null hypothesis.
* This means there is no significant difference in mean sales between the three groups (Cereal, Chocolate Candy, and Coffee).

Conclusion:

The data does not provide strong enough evidence to conclude that sales differ significantly among the three product categories. Since we failed to reject the null hypothesis, there is no need to perform Tukey’s test for pairwise comparisons.

Question: 7

Perform a complete one-way ANOVA. If the null hypothesis is rejected, use either the Scheffé or Tukey test to see if there is a significant difference in the pairs of means. Assume all assumptions are met.

The expenditures (in dollars) per pupil for states in three sections of the country are listed. Using α = 0.05, can you conclude that there is a difference in means?

| **Eastern third** | **Middle third** | **Western third** |
| --- | --- | --- |
| 4946 | 6149 | 5282 |
| 5953 | 7451 | 8605 |
| 6202 | 6000 | 6528 |
| 7243 | 6479 | 6911 |
| 6113 |  |  |

**Solution:**

**a) State the Hypotheses**

* **Null Hypothesis (H0)**: The mean expenditures per pupil are the same for all three sections of the country.

A black and white text

Description automatically generated

* **Alternative Hypothesis (Ha)**: At least one meaning is different.

H1​:At least one μ differs.

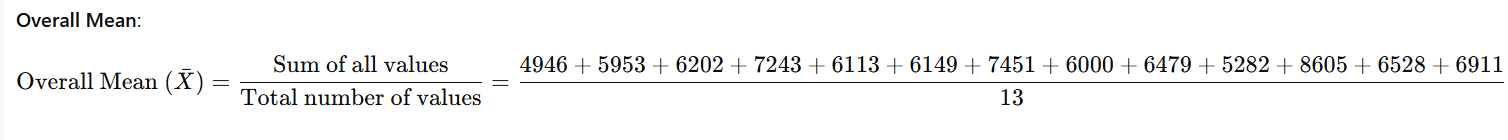
**b)Data**

**Data (Expenditures in dollars):**

* **Eastern third**: 4946,5953,6202,7243,6113
* **Middle third**: 6149,7451,6000,6479
* **Western third**: 5282,8605,6528,6911

A math problem with numbers

Description automatically generated



Overall Mean = 6399.23

**c) Calculate Sum of Squares**

**A math problem with numbers

Description automatically generated**

**A math equations and numbers

Description automatically generated**

**A number on a white background

Description automatically generated**

* 1. **Compute F Statistic**

**d)A math equations on a white background

Description automatically generated**

**Compare F with Critical Value:**

**A math equation with black text

Description automatically generated**

The F-statistic (F=2.945) does not exceed the critical value (F critical=4.10).

Conclusion: There is insufficient evidence to conclude that the mean expenditure per pupil differ significantly among the three sections of the country at α=0.05.

**Question:** Assume that all variables are normally or approximately normally distributed, that the samples are independent, and that the population variances are equal.

1. State the hypotheses.
2. Find the critical value for each F test.
3. Compute the summary table and find the test value.
4. Make the decision.
5. Summarize the results. *(Draw a graph of the cell means if necessary.)*

A gardening company is testing new ways to improve plant growth. Twelve plants are randomly selected and exposed to a combination of two factors, a “Grow-light” in two different strengths and a plant food supplement with different mineral supplements. After a number of days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes.

|  | **Grow-light 1** | **Grow-light 2** |
| --- | --- | --- |
| **Plant food A** | 9.2, 9.4, 8.9 | 8.5, 9.2, 8.9 |
| **Plant food B** | 7.1, 7.2, 8.5 | 5.5, 5.8, 7.6 |

Can an interaction between the two factors be concluded? Is there a difference in mean growth with respect to light? With respect to plant food? Use α = 0.05.

Solution:

* 1. State Hypothesis:

A white paper with black text

Description automatically generated

* 1. Organize data

A screenshot of a computer

Description automatically generated

A math equations on a white background

Description automatically generated

A math equations with numbers and symbols

Description automatically generated with medium confidence

* 1. Perform Two Way Anova

A screenshot of a computer

Description automatically generated

* 1. **Decision**

**A screenshot of a computer

Description automatically generated**

**A graph with a line and a red line

Description automatically generated**

* 1. **Summary**

There is a significant difference in plant growth based on the type of Plant food (p=0.001) but not based on the Grow-light (p=0.057).

No significant interaction exists between Grow-light and Plant food (p=0.155).

The analysis suggests that Plant food A leads to significantly greater growth than Plant food B.

**Part: 03**

**“Baseball Dataset”**

**Introduction and Descriptive Analysis**

The dataset contains information about baseball teams, including variables such as the number of wins (W), the league (League), and the year (Year). Descriptive statistics reveal key insights into the distribution of wins, with the mean, median, and range providing a sense of central tendency and variability. The histogram of wins shows a roughly symmetric distribution, with most teams winning between 70 and 90 games, indicating a balanced competitive landscape. The boxplot of wins by league suggests minimal differences in win distributions between leagues, though slight variations in median values may exist. The trend of wins over the years, visualized as a line plot, does not show a clear upward or downward pattern, implying relative stability in team performance over time. Additionally, the correlation heatmap highlights relationships between numeric variables, with no strong correlations immediately apparent. Overall, the dataset provides a comprehensive view of team performance, with no extreme outliers or anomalies detected during exploratory data analysis.

**Visualizations of EDA**

**a) Histogram of Wins**

**A graph of a number of wins

Description automatically generated**

**Insights:**

* The histogram shows the distribution of the number of wins (W) across all teams in the dataset.
* The distribution appears roughly symmetric, with most teams winning between 70 and 90 games. This suggests that the majority of teams perform within a similar range, indicating a balanced competitive environment.
* Few teams have extremely low (e.g., < 50) or high (e.g., > 100) win counts, which could represent outliers or particularly strong/weak teams.

**Impact:**

* This plot helps us understand the overall performance landscape of baseball teams. The symmetry and central tendency suggest that most teams are competitive, with fewer extreme performers.
* It also provides a baseline for comparing individual teams or leagues.

**Conclusion:**

* The distribution of wins is relatively normal, with most teams clustered around the mean. This indicates a fair and balanced competition structure in baseball.

b**) Boxplot of Wins by League**

A graph with a bar chart and a row of yellow squares

Description automatically generated with medium confidence

**Insights**:

* The boxplot compares the distribution of wins (W) between different leagues (e.g., National League and American League).
* The median number of wins (represented by the line in the box) appears similar across leagues, suggesting no significant difference in overall performance.
* The interquartile range (IQR, represented by the box) and whiskers (representing variability) are also comparable, indicating similar variability in wins across leagues.

**Impact**:

* This plot helps assess whether one league dominates the other in terms of team performance. The similarity in medians and variability suggests that both leagues are equally competitive.
* It can guide further analysis, such as investigating specific teams or years where performance differences might exist.

**Conclusion**:

* There is no clear evidence of one league outperforming the other. Both leagues exhibit similar win distributions, reinforcing the idea of balanced competition.

c**)Trend of Wins over Years**

A graph showing the number of years

Description automatically generated

**Insights**:

* The line plot shows the trend of wins (W) over the years (Year).
* The plot does not show a clear upward or downward trend, suggesting that the average number of wins per team has remained relatively stable over time.
* There may be minor fluctuations or peaks in certain years, which could correspond to specific events, rule changes, or exceptional team performances.

**Impact**:

* This plot helps identify long-term trends or anomalies in team performance. The lack of a clear trend suggests that the competitive structure of baseball has remained consistent over the years.
* It can prompt further investigation into specific years where fluctuations occur to understand potential causes.

**Conclusion**:

* The number of wins has remained stable over time, indicating no significant changes in the competitive dynamics of baseball across the years covered in the dataset.

**d)Correlation Heatmap**

**A graph of a heat map

Description automatically generated**

**Insights:**

* The heatmap visualizes the correlation coefficients between numeric variables in the dataset.
* Most correlations appear weak (close to 0), indicating little to no linear relationship between variables.
* If any strong correlations (close to +1 or -1) exist, they would suggest a meaningful relationship between variables (e.g., wins and runs scored).

**Impact:**

* This plot helps identify potential relationships between variables that could be explored further. For example, if wins were strongly correlated with runs scored, it might suggest that offensive performance is a key driver of success.
* The absence of strong correlations implies that no single variable dominates team performance, and multiple factors are likely to contribute to wins.

**Conclusion:**

* The lack of strong correlations suggests that team performance is influenced by a combination of factors rather than a single dominant variable. This aligns with the complexity of baseball as a sport.

**Conclusion of baseball dataset**

The exploration data analysis (EDA) of the baseball dataset reveals a balanced and stable competitive landscape in sport. The symmetric distribution of wins, with most teams winning between 70 and 90 games, highlights a fair and competitive environment where extreme outliers are rare. The similarity in win distributions across leagues suggests no significant dominance by either the National or American League, reinforcing the idea of parity in the sport. Additionally, the lack of a clear trend in wins over time indicates that team performance has remained consistent over the years, with no major shifts in competitive dynamics. The weak correlations observed in the heatmap further emphasize that team success is influenced by a combination of factors rather than a single dominant variable.

The Chi-Square Goodness-of-Fit test adds another layer of insight by statistically evaluating whether the distribution of wins across decades differs significantly from the expected distribution under the assumption of equal frequencies. This test is crucial for identifying potential trends or anomalies in team performance over time, providing a data-driven approach to understanding historical patterns. If the test reveals a significant difference, it could suggest that external factors (e.g., rule changes, economic conditions, or player talent shifts) have influenced team success in specific decades. Conversely, if no significant difference is found, it reinforces the stability and consistency observed in the EDA.

**APPENDIX**

**##Module 2**

##Crop Datset PART 01

# Load necessary libraries

library(ggplot2)

library(dplyr)

library(stats)

# Load the dataset

crop\_data <- read.csv("C:/Users/ravin/Downloads/crop\_data.csv")

# View the structure of the dataset

str(crop\_data)

summary(crop\_data) # Summary statistics

# Step 2: Check for missing values

missing\_vals <- colSums(is.na(crop\_data))

cat("Missing Values:\n")

print(missing\_vals[missing\_vals > 0])

# Step 3: Descriptive Analysis

describe(crop\_data) # Descriptive statistics for all columns

# Step 4: Visualizations

# Histogram of yield

ggplot(crop\_data, aes(yield)) +

geom\_histogram(fill = "lavender", bins = 20, color = "black", alpha = 0.7) +

labs(title = "Histogram of Crop Yield", x = "Yield", y = "Frequency")

# Bar plots for density and fertilizer

ggplot(crop\_data, aes(x = as.factor(density), fill = as.factor(density))) +

geom\_bar() +

labs(title = "Density Distribution", x = "Density", y = "Count") +

theme\_minimal()

ggplot(crop\_data, aes(x = as.factor(fertilizer), fill = as.factor(fertilizer))) +

geom\_bar() +

labs(title = "Fertilizer Distribution", x = "Fertilizer", y = "Count") +

theme\_minimal()

# Check normality using Shapiro-Wilk test

shapiro.test(crop\_data$yield)

# Check homogeneity of variance using Levene's test

leveneTest(yield ~ as.factor(fertilizer), data = crop\_data)

# Perform one-way ANOVA

anova\_result <- aov(yield ~ as.factor(fertilizer), data = crop\_data)

# Display the ANOVA table

summary(anova\_result)

# Calculate degrees of freedom

k <- length(unique(crop\_data$fertilizer)) # Number of fertilizer types

N <- nrow(crop\_data) # Total number of observations

df1 <- k - 1

df2 <- N - k

# Find the critical F-value

critical\_f <- qf(0.95, df1, df2)

print(paste("Critical F-value:", critical\_f))

# Perform Tukey HSD test

tukey\_result <- TukeyHSD(anova\_result)

print(tukey\_result)

ggplot(crop\_data, aes(x = as.factor(fertilizer), y = yield, fill = as.factor(fertilizer))) +

geom\_boxplot() +

labs(title = "Boxplot of Yield by Fertilizer Type",

x = "Fertilizer Type",

y = "Yield") +

theme\_minimal()

plot(anova\_result, which = 1) # Residuals vs Fitted

#QUESTION

# Cell means

cell\_means <- data.frame(

Grow\_light = rep(c("Grow-light 1", "Grow-light 2"), each = 2),

Plant\_food = rep(c("Plant food A", "Plant food B"), 2),

Growth = c(9.17, 7.6, 8.87, 6.3) # Calculated means for each cell

)

# Load ggplot2 library

library(ggplot2)

# Create the interaction plot

ggplot(cell\_means, aes(x = Grow\_light, y = Growth, group = Plant\_food, color = Plant\_food)) +

geom\_line(size = 1) + # Line connecting points

geom\_point(size = 3) + # Points for means

labs(

title = "Interaction Plot: Grow-light vs Plant Food",

x = "Grow-light",

y = "Mean Growth (inches)",

color = "Plant Food"

) +

theme\_minimal(base\_size = 14) + # Minimal theme for clean visualization

theme(legend.position = "top")

**#part:3**

# Load necessary libraries

library(tidyverse)

# Import the dataset

df <- read\_csv("C:\\Users\\ravin\\Downloads\\baseball.csv")

# Display basic information about the dataset

glimpse(df)

##EDA

# Descriptive statistics

summary(df)

# Plotting

# Distribution of Wins

ggplot(df, aes(x = W)) +

geom\_histogram(bins = 20, fill = "lightsteelblue", alpha = 0.7) +

labs(title = "Distribution of Wins", x = "Number of Wins", y = "Frequency")

# Boxplot of Wins by League

ggplot(df, aes(x = League, y = W)) +

geom\_boxplot(fill = "orange", alpha = 0.7) +

labs(title = "Boxplot of Wins by League", x = "League", y = "Number of Wins")

# Trend of Wins Over Years

# Load necessary library

library(ggplot2)

# Plot the trend of wins over the years

ggplot(df, aes(x = Year, y = W)) +

geom\_line(color = "red") +

labs(title = "Trend of Wins Over Years", x = "Year", y = "Number of Wins")

# Correlation heatmap

# Install the reshape2 package

install.packages("reshape2")

# Load the reshape2 package

library(reshape2)

# Create the correlation heatmap

correlation\_matrix <- cor(df %>% select\_if(is.numeric))

melted\_correlation\_matrix <- melt(correlation\_matrix)

ggplot(melted\_correlation\_matrix, aes(Var1, Var2, fill = value)) +

geom\_tile() +

scale\_fill\_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +

labs(title = "Correlation Heatmap", x = "", y = "")

##Chi Square

# Create a new column 'Decade' to categorize the years into decades

df <- df %>%

mutate(Decade = floor(Year / 10) \* 10)

# Group by 'Decade' and sum the wins for each decade

wins\_by\_decade <- df %>%

group\_by(Decade) %>%

summarise(Total\_Wins = sum(W))

# Perform Chi-Square Goodness-of-Fit test

observed\_frequencies <- wins\_by\_decade$Total\_Wins

expected\_frequencies <- rep(mean(observed\_frequencies), length(observed\_frequencies))

chi\_square\_test <- chisq.test(observed\_frequencies, p = expected\_frequencies / sum(expected\_frequencies))

# Print the results

chi\_square\_test

##End of Module 2 ##

#Payal Sharma

#MPS Analytics