**Module 2 Assignment-Chi-Square Testing and ANOVA**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

[**ALY6015.71591.202515**](https://northeastern.instructure.com/courses/196378) **: Intermediate Analytics**

**Professor: Roy Wada**

**November 9th, 2024**

**Overview and Rationale**

This assignment demonstrates the application of chi-square and ANOVA testing techniques to solve various types of statistical problems. By analyzing different datasets using chi-square tests and ANOVA, this report provides insights into categorical data distribution, independence between categorical variables, and differences in means across multiple groups. The purpose is to enhance understanding of data analysis techniques and support hypothesis testing with statistical rigor.

**Module Outcomes**

This assignment links directly to the following learning outcomes from the course syllabus:

1. Use “R” effectively to process, analyze, and depict data.
2. Develop more advanced models to interpret data.
3. Prepare complex datasets for analysis.

**Assignment Summary**

The assignment includes various problems solved using R, with step-by-step hypothesis testing and analysis for each case. The problems cover:

* **Chi-Square Goodness-of-Fit Tests**: Assessing whether observed distributions match expected probabilities.
* **Chi-Square Test of Independence**: Testing whether two categorical variables are independent.
* **One-Way ANOVA**: Identifying significant differences among multiple group means.
* **Two-Way ANOVA**: Analyzing interaction effects between two independent variables on a dependent variable.

The report includes all necessary hypothesis tests, critical values, computed test values, and resulting decisions. Each section is accompanied by R code (included in the appendix) and visualizations where appropriate.

**Analysis**

**Section 1: Chi-Square Goodness of Fit**

**Problem 6: Blood Types**

* **Objective**: Determine if the distribution of blood types among hospital patients matches the distribution in the general population.
* **Hypothesis**:
  + 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.
* **Assumptions**: The observed frequencies for each blood type are compared against the expected frequencies based on the known general population distribution (α = 0.10).
* **Test Results**:
  + Chi-Squared value: 5.4714
  + Degrees of Freedom: 3
  + p-value: 0.1404
* **Decision**: Since p > 0.10, we fail to reject the null hypothesis. This suggests there is no significant difference between the blood type distribution among hospital patients and the general population.
* **Summary Statistics**:

| **Blood Type** | **Observed** | **Expected** |
| --- | --- | --- |
| A | 12 | 10 |
| B | 8 | 14 |
| O | 24 | 18 |
| AB | 6 | 8 |

* **Visualization**:

A graph of a bar graph

Description automatically generated

**Problem 8: On-Time Performance by Airlines**

* **Objective**: Evaluate if the observed on-time performance distribution for a major airline company matches the government’s statistics.
* **Hypothesis**:
  + Null Hypothesis (H0): The airline’s on-time performance distribution matches the government statistics.
  + Alternative Hypothesis (H1): The airline’s on-time performance distribution differs from the government statistics.
* **Assumptions**: Observed frequencies are compared to expected frequencies based on government statistics (α = 0.05).
* **Test Results**:
  + Chi-Squared value: 39.504
  + Degrees of Freedom: 3
  + p-value: 1.357e-08
* **Decision**: Since p < 0.05, we reject the null hypothesis. This indicates that there is a significant difference between the airline’s on-time performance and the government statistics.
* **Summary Statistics**:

| **Action** | **Observed** | **Expected** |
| --- | --- | --- |
| On time | 125 | 141.6 |
| Weather Delay | 40 | 16.4 |
| NAS Delay | 10 | 18.0 |
| Late Arrival | 25 | 24.0 |

* **Visualization**:

A graph with different colored squares

Description automatically generated

**Section 2: Chi-Square Test of Independence**

**Problem 8: Ethnicity and Movie Admissions**

* **Objective**: Test if movie attendance is independent of ethnicity.
* **Hypotheses**:
  + **Null Hypothesis (H0)**: Movie attendance by year is independent of ethnicity.
  + **Alternative Hypothesis (H1)**: Movie attendance by year depends on ethnicity.
* **Critical Value**: At α = 0.05 and with 3 degrees of freedom, the critical value is **7.815**.
* **Test Results**:
  + Chi-Squared value: **32.677**
  + Degrees of Freedom: **3**
  + p-value: **3.767e-07**
* **Decision**: Since **p < 0.05**, we reject the null hypothesis. Movie attendance appears dependent on ethnicity.

**Problem 10: Women in the Military**

* **Objective**: Test if there is a relationship between rank and branch of service in the military.
* **Hypotheses**:
  + **Null Hypothesis (H0)**: Rank and branch of service are independent.
  + **Alternative Hypothesis (H1)**: Rank and branch of service are dependent.
* **Critical Value**: At α = 0.05 and with 3 degrees of freedom, the critical value is **7.815**.
* **Test Results**:
  + Chi-Squared value: **654.27**
  + Degrees of Freedom: **3**
  + p-value < **2.2e-16**
* **Decision**: Since **p < 0.05**, we reject the null hypothesis. There is a significant relationship between rank and branch of service.

**Section 3: One-Way ANOVA**

**Problem 8: Sodium Contents of Foods**

* **Objective**: Test if there is a significant difference in mean sodium content among condiments, cereals, and desserts.
* **Hypothesis**:
  + Null Hypothesis (H0): There is no difference in mean sodium content among condiments, cereals, and desserts.
  + Alternative Hypothesis (H1): There is a difference in mean sodium content among the food types.
* **Critical Value:** The critical F-value at α = 0.05 is approximately 3.55
* **Test Results**:
  + F value: 3.194
  + p-value: 0.0637
* **Decision**: Since p > 0.05, we fail to reject the null hypothesis. There is no significant difference in sodium content among the food types.
* **Summary**: Although there are variations in mean sodium content across the food types, these differences are not statistically significant and could be due to random sampling variability.
* **Summary Statistics**:

| **Food Type** | **Mean Sodium** | **SD** |
| --- | --- | --- |
| Cereals | 246 | 62.7 |
| Condiments | 165 | 69.9 |
| Desserts | 249 | 86.1 |

* **Visualization**:

A graph of a bar chart

Description automatically generated with medium confidence

**Section 4: One-Way ANOVA with Post-Hoc Test**

**Problem 10: Sales for Leading Companies**

* **Objective**: Test if there is a significant difference in mean sales among different company categories.
* **Hypothesis**:
  + Null Hypothesis (H0): There is no difference in mean sales among the company categories.
  + Alternative Hypothesis (H1): There is a difference in mean sales among the company categories.
* **Critical Value**: Since α = 0.01, we are using this as the threshold for significance.
* **Test Results**:
  + F value: 1.413
  + p-value: 0.296
* **Decision**: Since p > 0.01, we fail to reject the null hypothesis. There is no significant difference in mean sales among categories at the 0.01 significance level.
* **Summary Statistics**:

| **Category** | **Mean Sales** | **SD** |
| --- | --- | --- |
| Candy | 186 | 69.0 |
| Cereal | 330 | 142 |
| Chocolate | 175 | 118 |
| Coffee | 392 | 264 |

* **Visualization**:

A graph of sales

Description automatically generated

**Problem 12: Per-Pupil Expenditures**

* **Objective**: Test if there is a significant difference in mean per-pupil expenditures among the three sections of the country (Eastern, Middle, and Western).
* **Hypotheses**:
  + **Null Hypothesis (H0)**: There is no difference in mean per-pupil expenditures among the three sections.
  + **Alternative Hypothesis (H1)**: There is a difference in mean per-pupil expenditures among the three sections.
* **Significance Level**: α = 0.05
* **Test Results**:
  + **F value**: 0.649
  + **p-value**: 0.543
  + **Degrees of Freedom**: 2 for the section variable, and 10 for the residuals
* **Decision**: Since **p > 0.05**, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in mean per-pupil expenditures among the three sections of the country at the 0.05 significance level.
* **Summary Statistics**:

| **Section** | **Mean Expenditures** | **SD** | **Sample Size** |
| --- | --- | --- | --- |
| Eastern | 6091 | 817 | 5 |
| Middle | 6520 | 652 | 4 |
| Western | 6832 | 1372 | 4 |

* **Post-hoc Test**: Since we did not reject the null hypothesis, there is no need to perform a post-hoc test (such as Tukey's or Scheffé's) because we only conduct post-hoc comparisons when there is a significant result in ANOVA.
* **Conclusion:** There is no significant difference in mean per-pupil expenditures among the Eastern, Middle, and Western sections of the country. Any observed differences in the sample means are likely due to random sampling variability rather than true differences among the sections.
* **Visualization**:

A graph with different colored squares

Description automatically generated

**Section 5: Two-Way ANOVA**

**Problem 10: Increasing Plant Growth**

* **Objective**: Test the effects of light strength and plant food type on plant growth and determine if there is an interaction effect between these factors.

**Solution Steps**

**a. Hypotheses**

1. **Main Effect of Light**:
   * **Null Hypothesis (H0)**: There is no difference in mean plant growth between the two light strengths.
   * **Alternative Hypothesis (H1)**: There is a difference in mean plant growth between the two light strengths.
2. **Main Effect of Plant Food**:
   * **Null Hypothesis (H0)**: There is no difference in mean plant growth between the two types of plant food.
   * **Alternative Hypothesis (H1)**: There is a difference in mean plant growth between the two types of plant food.
3. **Interaction Effect**:
   * **Null Hypothesis (H0)**: There is no interaction effect between light strength and plant food type on plant growth.
   * **Alternative Hypothesis (H1)**: There is an interaction effect between light strength and plant food type on plant growth.

**b. Critical Value**

* **Significance Level**: α = 0.05.
* Using an F-distribution table or software, we can determine the critical values based on degrees of freedom for each effect.

**c. Summary Table and Test Values**

The ANOVA results you provided include the F values and p-values for each factor and the interaction:

| **Effect** | **F Value** | **p-value** |
| --- | --- | --- |
| **Light** | **15.531** | **0.00429** |
| **Food** | **0.259** | **0.62482** |
| **Interaction** | **0.036** | **0.85352** |

**d. Decision**

* **Effect of Light**: Since **p = 0.00429 < 0.05**, we reject the null hypothesis for light. This indicates that there is a significant difference in mean plant growth with respect to light strength.
* **Effect of Food**: Since **p = 0.62482 > 0.05**, we fail to reject the null hypothesis for food. This suggests no significant difference in mean plant growth with respect to plant food type.
* **Interaction**: Since **p = 0.85352 > 0.05**, we fail to reject the null hypothesis for the interaction effect. This suggests no significant interaction between light strength and plant food type.

**Summary Statistics**:

| **Light** | **Food** | **Mean Growth** | **SD** |
| --- | --- | --- | --- |
| Grow-light 1 | A | 9.1 | 0.173 |
| Grow-light 1 | B | 8.93 | 0.451 |
| Grow-light 2 | A | 7.13 | 1.35 |
| Grow-light 2 | B | 6.77 | 1.12 |

* **Visualization**:

A graph with a bar and a bar chart

Description automatically generated with medium confidence

**Conclusion**

* **Main Effect of Light**: Significant—plant growth differs significantly between light strengths.
* **Main Effect of Food**: Not significant—no difference in growth between plant food types.
* **Interaction**: Not significant—no interaction effect between light strength and plant food type on plant growth.

The results suggest that only light strength impacts plant growth significantly, while plant food type and the interaction between light and food type do not significantly influence growth.

**Section 6: Baseball Dataset Analysis (EDA & Chi-Square Goodness-of-Fit)**

**Exploratory Data Analysis (EDA)**

The **baseball dataset** includes data on wins by various teams across decades. For this analysis, we examine the distribution of wins over different decades. A brief exploratory data analysis (EDA) reveals the following summary statistics for the number of wins by decade:

* **Descriptive Statistics**:
  + The dataset spans multiple decades, with each row representing the sum of wins for a particular decade.
  + Calculated mean wins, standard deviation, and other key statistics are presented in the summary table below.

**Summary Statistics Table for Wins by Decade**

| **Statistic** | **Value** |
| --- | --- |
| Mean Wins | 16,612 |
| SD Wins | 5,761 |
| Count (n) | 6 |

**Hypothesis Testing: Chi-Square Goodness-of-Fit**

* **Objective**: Test whether the observed distribution of wins by decade is consistent with an expected distribution.
* **Hypothesis**:
  + Null Hypothesis (H0): The distribution of wins by decade is consistent.
  + Alternative Hypothesis (H1): The distribution of wins by decade is not consistent.
* **Assumptions**: The expected frequencies are equal (α = 0.05).

**Test Details**

1. **Critical Value**: For α = 0.05 with 5 degrees of freedom, the critical chi-square value from the chi-square table is approximately 11.07.
2. **Test Statistic Calculation**: The test statistic (X-squared) was calculated using the chi-square goodness-of-fit test in R. The computed chi-square value was **9989.5**.
3. **Decision**:
   * Since the computed test value (9989.5) is much greater than the critical value (11.07), we reject the null hypothesis.
   * **Conclusion**: There is strong evidence that the distribution of wins by decade is not consistent.

* **Visualization**:

A graph of a bar chart

Description automatically generated with medium confidence

**Conclusion**

This assignment demonstrated the use of chi-square and ANOVA tests to understand the distribution and relationship of categorical variables, as well as to compare group means. The baseball dataset analysis specifically showed a significant difference in the distribution of wins by decade, indicating variability across the years. Through hypothesis testing and visualization, this assignment provided insights into data patterns and group comparisons across various problems.

**Appendix**

The R code used in this report is provided here for reproducibility:

|  |
| --- |
| # Load necessary libraries  library(dplyr)  library(tidyverse)  library(ggplot2)  library(plotly)  # Set alpha level for tests  alpha <- 0.10  # -----------------------------------------  # Section 11-1: Chi-Square Goodness of Fit  # -----------------------------------------  # Problem 6: Blood Types  observed <- c(12, 8, 24, 6)  expected\_prob <- c(0.20, 0.28, 0.36, 0.16) \* sum(observed)  blood\_test <- chisq.test(observed, p = expected\_prob / sum(expected\_prob))  print(blood\_test)  # Summary statistics  blood\_summary <- data.frame(Blood\_Type = c("A", "B", "O", "AB"),  Observed = observed, Expected = expected\_prob)  print(blood\_summary)  # 3D Bar Plot without unsupported attributes  suppressWarnings({  fig <- plot\_ly(blood\_summary, x = ~Blood\_Type, y = ~Observed, type = "bar", color = ~Blood\_Type) %>%  layout(title = "3D Bar Plot of Blood Type Distribution",  scene = list(xaxis = list(title = "Blood Type"),  yaxis = list(title = "Observed")))  fig  })  # Problem 8: On-Time Performance by Airlines  observed <- c(125, 40, 10, 25)  expected\_prob <- c(0.708, 0.082, 0.09, 0.12) \* sum(observed)  airline\_test <- chisq.test(observed, p = expected\_prob / sum(expected\_prob))  print(airline\_test)  # Summary statistics  airline\_summary <- data.frame(Action = c("On time", "Weather Delay", "NAS Delay", "Late Arrival"),  Observed = observed, Expected = expected\_prob)  print(airline\_summary)  # 3D Bar Plot without unsupported attributes  suppressWarnings({  fig <- plot\_ly(airline\_summary, x = ~Action, y = ~Observed, type = "bar", color = ~Action) %>%  layout(title = "3D Bar Plot of Airline On-Time Performance",  scene = list(xaxis = list(title = "Action"),  yaxis = list(title = "Observed")))  fig  })  # -----------------------------------------  # Section 11-2: Chi-Square Test of Independence  # -----------------------------------------  # Problem 8: Ethnicity and Movie Admissions  movie\_data <- matrix(c(724, 370, 335, 292, 174, 152, 107, 140),  nrow = 2, byrow = TRUE)  colnames(movie\_data) <- c("Caucasian", "Hispanic", "African American", "Other")  rownames(movie\_data) <- c("2013", "2014")  movie\_test <- chisq.test(movie\_data)  print(movie\_test)  # Problem 10: Women in the Military  military\_data <- matrix(c(10791, 7816, 932, 11819, 62491, 42750, 9525, 54344),  nrow = 2, byrow = TRUE)  colnames(military\_data) <- c("Army", "Navy", "Marine Corps", "Air Force")  rownames(military\_data) <- c("Officers", "Enlisted")  military\_test <- chisq.test(military\_data)  print(military\_test)  # -----------------------------------------  # Section 12-1: One-Way ANOVA  # -----------------------------------------  # Problem 8: Sodium Contents of Foods  sodium\_data <- data.frame(  sodium = c(270, 130, 230, 180, 80, 70, 200, 160, 260, 220, 290, 290, 200, 320, 140, 100, 180, 250, 250, 300, 360, 300),  food\_type = factor(rep(c("Condiments", "Cereals", "Desserts"), c(8, 7, 7)))  )  sodium\_anova <- aov(sodium ~ food\_type, data = sodium\_data)  summary(sodium\_anova)  # Summary table  sodium\_summary <- sodium\_data %>%  group\_by(food\_type) %>%  summarize(Mean = mean(sodium), SD = sd(sodium), n = n())  print(sodium\_summary)  # 3D Bar Plot without unsupported attributes  suppressWarnings({  fig <- plot\_ly(sodium\_summary, x = ~food\_type, y = ~Mean, type = "bar", color = ~food\_type) %>%  layout(title = "3D Bar Plot of Sodium Contents by Food Type",  scene = list(xaxis = list(title = "Food Type"),  yaxis = list(title = "Mean Sodium")))  fig  })  # -----------------------------------------  # Section 12-2: One-Way ANOVA with Post-Hoc Test  # -----------------------------------------  # Problem 10: Sales for Leading Companies  sales\_data <- data.frame(  sales = c(578, 320, 264, 249, 237, 311, 106, 109, 125, 173, 261, 185, 302, 689),  category = factor(rep(c("Cereal", "Chocolate", "Candy", "Coffee"), c(5, 3, 3, 3)))  )  sales\_anova <- aov(sales ~ category, data = sales\_data)  summary(sales\_anova)  # Post-hoc Tukey test if ANOVA is significant  if (summary(sales\_anova)[[1]][["Pr(>F)"]][1] < alpha) {  tukey\_test <- TukeyHSD(sales\_anova)  print(tukey\_test)  }  # Summary table for Sales  sales\_summary <- sales\_data %>%  group\_by(category) %>%  summarize(Mean = mean(sales), SD = sd(sales), n = n())  print(sales\_summary)  # 3D Bar Plot without unsupported attributes  suppressWarnings({  fig <- plot\_ly(sales\_summary, x = ~category, y = ~Mean, type = "bar", color = ~category) %>%  layout(title = "3D Bar Plot of Sales by Company Category",  scene = list(xaxis = list(title = "Category"),  yaxis = list(title = "Mean Sales")))  fig  })  # -----------------------------------------  # Section 12-3: Two-Way ANOVA  # -----------------------------------------  # Problem 10: Plant Growth  growth\_data <- data.frame(  growth = c(9.2, 9.4, 8.9, 8.5, 9.2, 8.9, 7.1, 7.2, 8.5, 5.5, 5.8, 7.6),  light = factor(rep(c("Grow-light 1", "Grow-light 2"), each = 6)),  food = factor(rep(c("A", "B"), times = 6))  )  growth\_anova <- aov(growth ~ light \* food, data = growth\_data)  summary(growth\_anova)  # Summary table for Plant Growth  growth\_summary <- growth\_data %>%  group\_by(light, food) %>%  summarize(Mean = mean(growth), SD = sd(growth), n = n())  print(growth\_summary)  # 3D Bar Plot without unsupported attributes  suppressWarnings({  fig <- plot\_ly(growth\_summary, x = ~light, y = ~Mean, type = "bar", color = ~food) %>%  layout(title = "3D Bar Plot of Plant Growth by Light and Food Type",  scene = list(xaxis = list(title = "Light"),  yaxis = list(title = "Mean Growth")))  fig  })  # -----------------------------------------  # Baseball Dataset Analysis (EDA & Chi-Square Goodness-of-Fit)  # -----------------------------------------  # Load baseball.csv data  baseball\_data <- read.csv('baseball.csv')  # Add Decade column  baseball\_data$Decade <- baseball\_data$Year - (baseball\_data$Year %% 10)  # Summing wins by decade  wins\_by\_decade <- baseball\_data %>%  group\_by(Decade) %>%  summarize(total\_wins = sum(W))  # Summary statistics table  summary\_table <- wins\_by\_decade %>%  summarize(Mean\_Wins = mean(total\_wins), SD\_Wins = sd(total\_wins), n = n())  print(summary\_table)  # Chi-Square Goodness-of-Fit for Wins by Decade  observed\_wins <- wins\_by\_decade$total\_wins  expected\_wins <- rep(mean(observed\_wins), length(observed\_wins))  wins\_test <- chisq.test(observed\_wins, p = rep(1/length(observed\_wins), length(observed\_wins)))  print(wins\_test)  # 3D Bar Plot without unsupported attributes  suppressWarnings({  fig <- plot\_ly(wins\_by\_decade, x = ~Decade, y = ~total\_wins, type = "bar", color = ~Decade) %>%  layout(title = "3D Bar Plot of Wins by Decade",  scene = list(xaxis = list(title = "Decade"),  yaxis = list(title = "Total Wins")))  fig  }) |

**References**

**Chi-Square Goodness-of-Fit Test Explanation**:

* Laerd Statistics. (n.d.). *Chi-Square Goodness-of-Fit Test in SPSS Statistics*. Retrieved November 7, 2024, from <https://statistics.laerd.com/spss-tutorials/chi-square-goodness-of-fit-test-in-spss-statistics.php>

**ANOVA (Analysis of Variance) Overview:**

* Statistics Solutions. (n.d.). *ANOVA (Analysis of Variance)*. Retrieved November 8, 2024, from <https://www.statisticssolutions.com/anova/>

**R Programming for Statistical Analysis:**

* R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

**Data Visualization with Plotly in R:**

* Plotly Technologies Inc. (n.d.). *Plotly for R*. Retrieved November 8, 2024, from <https://plotly.com/r/>