**ALY6015 80472 Intermediate Analytics SEC 04 Spring 2023 CPS**

**Module 2 Assignment — Chi Square and ANOVA REPORT**

Logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Zihan Ma

ma.zihan1@northeastern.edu

**Instructor**

 Prof. Valeriy Shevchenko

**Date**

06/05/2023

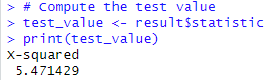
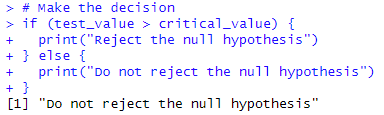
**Chi Square and ANOVA**

**Assignment Summary:**

This course assignment primarily focused on two statistical analysis techniques - Chi Square test and Analysis of Variance (ANOVA). These techniques were applied across various data sets and tasks, including AmesHousing dataset, data related to blood type distribution, airline performance, movie admissions, sodium amounts in various foods, sales across industries, and others.

**Tasks:**

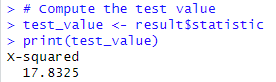
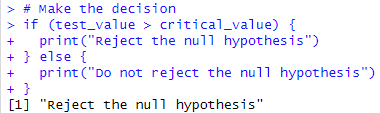
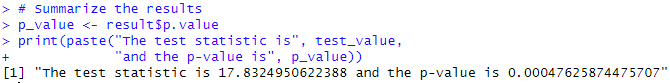
**Task 1: Chi-Square Goodness-of-Fit Test for Blood Type Distribution Comparison**

1. The null hypothesis (H0) stated that the distribution of blood types in the hospital is the same as that in the general population.   
   The alternative hypothesis (H1) proposed that the distribution of blood types in the hospital is not the same as that in the general population.
2.   
   The critical value for this test was found to be approximately 6.251.
3.   
   The computed Chi-Square test statistic from our data was approximately 5.471, which is less than the critical value.
4.   
   Because the computed test value did not exceed the critical value, we did not reject the null hypothesis. This means that, based on our sample, we do not have enough evidence to conclude that the distribution of blood types in the hospital differs from that in the general population.
5. A picture containing text, font, screenshot, line

   Description automatically generated  
   The p-value of the test was approximately 0.140, which is greater than the significance level. Based on this Chi-Square goodness-of-fit test, we concluded that there is not enough statistical evidence at the 0.10 significance level to suggest that the distribution of blood types in the hospital differs from that in the general population.

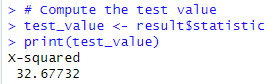
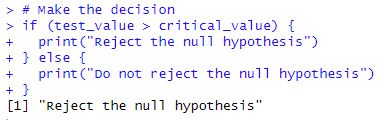
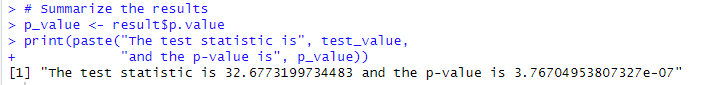
**Task 2: Chi-Square Goodness-of-Fit Test for On-Time Performance by Airlines**

1. Null hypothesis (H0): The distribution of on-time performance by the major airline company is the same as that described by the Bureau of Transportation Statistics.  
     
   Alternative hypothesis (H1): The distribution of on-time performance by the major airline company is not the same as that described by the Bureau of Transportation Statistics.
2. A picture containing text, font, screenshot, line

   Description automatically generated  
   We conducted the Chi-Square test at a significance level of α = 0.05. The critical value for this test, with 3 degrees of freedom, was found to be approximately 7.815.
3.   
   The computed Chi-Square test statistic from our data was approximately 17.83, which exceeds the critical value.
4.   
   Since the computed test value exceeded the critical value, we reject the null hypothesis. This means that, based on our sample, we have strong evidence to conclude that the distribution of on-time performance by the major airline differs from that in the government's statistics.
5.   
   The p-value of the test was approximately 0.000476, which is much lower than the significance level. This further supports our decision to reject the null hypothesis, as a lower p-value indicates strong evidence against the null hypothesis.  
   Based on this Chi-Square goodness-of-fit test, we concluded that there is enough statistical evidence at the 0.05 significance level to suggest that the on-time performance distribution of the major airline differs significantly from that of the government's statistics.

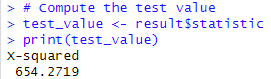
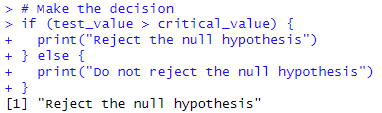
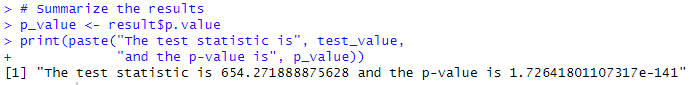
**Task 3: Chi-Square Test of Independence: Ethnicity and Movie Admissions**

1. Null hypothesis (H0): Movie attendance by year is independent of ethnicity.  
     
   Alternative hypothesis (H1): Movie attendance by year is not independent of ethnicity, implying that it depends on ethnicity.
2. A picture containing text, font, screenshot, line

   Description automatically generated  
   The critical value from the Chi-Square distribution for this test with 3 degrees of freedom at α=0.05 is approximately 7.815.
3.   
   The computed Chi-Square test statistic is approximately 32.68.
4.   
   Since the computed test statistic (32.68) is larger than the critical value (7.81), we reject the null hypothesis.
5.   
   The test statistic is approximately 32.68, and the p-value is approximately 3.77e-07. Given that the p-value is less than the significance level (α=0.05), there is very strong evidence to reject the null hypothesis. This suggests that movie attendance by year is dependent on ethnicity. Thus, the results support the alternative hypothesis and the claim that ethnicity does have an effect on movie attendance.

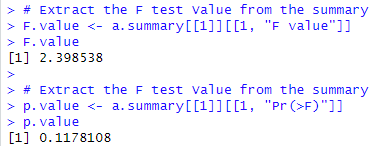
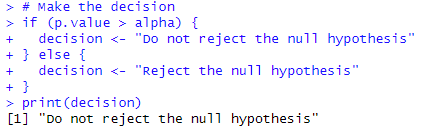
**Task 4: Chi-Square Test of Independence: Ethnicity and Movie Admissions**

1. Null Hypothesis (H₀): There is no relationship between rank and branch of the Armed Forces.  
     
   Alternative Hypothesis (H₁): There is a relationship between rank and branch of the Armed Forces.
2. A picture containing text, font, screenshot, line

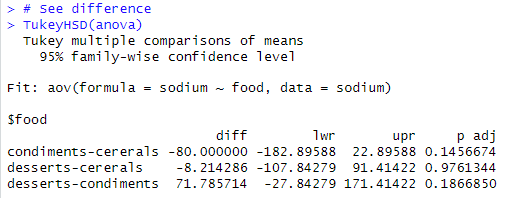
   Description automatically generated  
   The critical value from the Chi-Square distribution for this test with 3 degrees of freedom at α=0.05 is approximately 7.815.
3.   
   To assess the relationship, a Chi-Square test statistic was computed. The test statistic value was found to be 654.2719.
4.   
   Since the test value (654.2719) exceeds the critical value (7.814728), we reject the null hypothesis.
5.   
   The p-value associated with the test statistic is approximately 1.72641801107317e-141, which is significantly smaller than the significance level of 0.05. This indicates strong evidence against the null hypothesis. Therefore, we conclude that there is sufficient evidence to support the claim that a relationship exists between rank and branch of the Armed Forces.

**Task 5: Difference in Mean Sodium Amounts Among Condiments, Cereals, and Desserts**

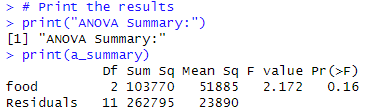
1. Null Hypothesis (H₀): There is no difference in mean sodium amounts among the three categories: condiments, cereals, and desserts.  
     
   Alternative Hypothesis (H₁): There is a difference in mean sodium amounts among the three categories.
2. A picture containing text, screenshot, font, line

   Description automatically generated  
   Find 2 degrees of freedom for the numerator (k - 1) and 19 degrees of freedom for the denominator (N - k).
3.   
   The ANOVA summary table shows that the F test value is 2.399, indicating the ratio of the between-group variance to the within-group variance.
4.   
   The p-value associated with the F test is 0.118, which is greater than the significance level of 0.05.

Based on the decision rule, we fail to reject the null hypothesis.

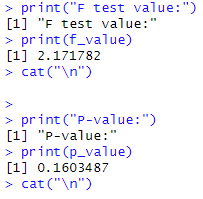
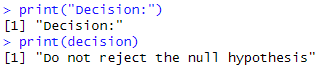
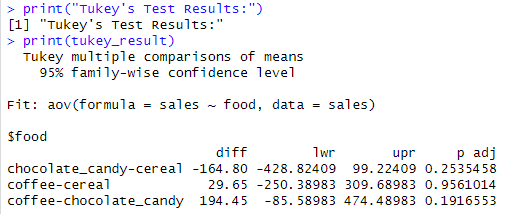
1.   
   Based on the analysis at a significance level of α = 0.05, the one-way ANOVA does not provide sufficient evidence to reject the null hypothesis. Therefore, we do not have enough evidence to conclude that there is a significant difference in the mean sodium amounts among the three categories: condiments, cereals, and desserts.

**Task 6: Difference in Means of Sales Among Cereal, Chocolate Candy, and Coffee Industries**

1. Null Hypothesis (H₀): There is no significant difference in the means of sales among the Cereal, Chocolate Candy, and Coffee industries.  
     
   Alternative Hypothesis (H₁): There is a significant difference in the means of sales among the industries.
2.   
   The ANOVA summary table shows an F-value of 2.172 with a corresponding p-value of 0.160. The p-value is above the significance level of 0.01.
3. A picture containing text, font, screenshot, line

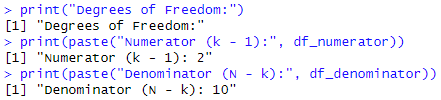
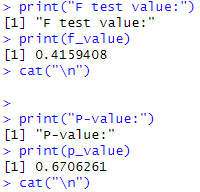
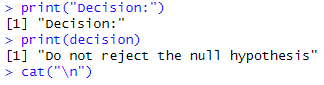
   Description automatically generated  
   Numerator (k - 1): 2

Denominator (N - k): 11

1.   
   The F test value is 2.172.  
   The p-value is 0.160.
2.   
   Based on the p-value exceeding the significance level, we do not have sufficient evidence to reject the null hypothesis. Therefore, we do not conclude a significant difference in the means of sales among the Cereal, Chocolate Candy, and Coffee industries.
3.   
   The post hoc Tukey's test was performed to examine pairwise differences in means. The results indicate that there are no significant differences between Chocolate Candy and Cereal, Coffee and Cereal, and Coffee and Chocolate Candy, based on the 95% family-wise confidence level.

**Task 7: Difference in Means of Per-Pupil Expenditures Among Eastern, Middle, and Western Thirds**

1. Null Hypothesis (H₀): There is no significant difference in the means of per-pupil expenditures among the Eastern third, Middle third, and Western third.  
     
   Alternative Hypothesis (H₁): There is a significant difference in the means of per-pupil expenditures among the thirds.
2. A picture containing text, font, screenshot, line

   Description automatically generated  
   The ANOVA summary table shows an F-value of 0.416 with a corresponding p-value of 0.671. The p-value is above the significance level of 0.05.
3.   
   Numerator (k - 1): 2  
   Denominator (N - k): 10
4.   
   The F test value is 0.   
   The p-value is 0.671.
5.   
   Based on the p-value exceeding the significance level, we do not have sufficient evidence to reject the null hypothesis. Therefore, we do not conclude a significant difference in the means of per-pupil expenditures among the Eastern third, Middle third, and Western third.

**Task 8: Difference in Means of Two Factors Among Increasing Plant Growth**

1. **Hypotheses  
     
   For Grow-lights:**Null Hypothesis (H0): There is no difference in mean growth for different Grow-lights.

Alternative Hypothesis (H1): There is a significant difference in mean growth for different Grow-lights.

**For Plant Food:**

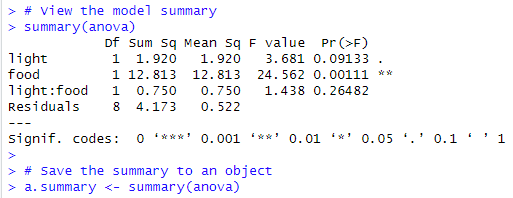
Null Hypothesis (H0): There is no difference in mean growth for different types of Plant Food.

Alternative Hypothesis (H1): There is a significant difference in mean growth for different types of Plant Food.

**For interaction of Grow-light and Plant Food:**

Null Hypothesis (H0): There is no interaction between Grow-light and Plant Food affecting plant growth.

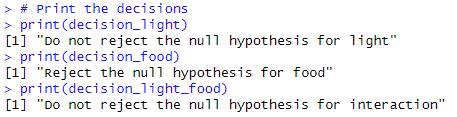
Alternative Hypothesis (H1): There is an interaction between Grow-light and Plant Food affecting plant growth.

1. **Test Results  
     
   For Grow-lights**, the F-value was 3.681 with a p-value of 0.09133. Since the p-value is greater than the significance level (α = 0.05), we do not reject the null hypothesis. Hence, we conclude that there is no significant difference in plant growth for different Grow-lights.

**For Plant Food**, the F-value was 24.562 with a p-value of 0.00111. Since the p-value is less than the significance level (α = 0.05), we reject the null hypothesis. This suggests that there is a significant difference in plant growth for different types of Plant Food.

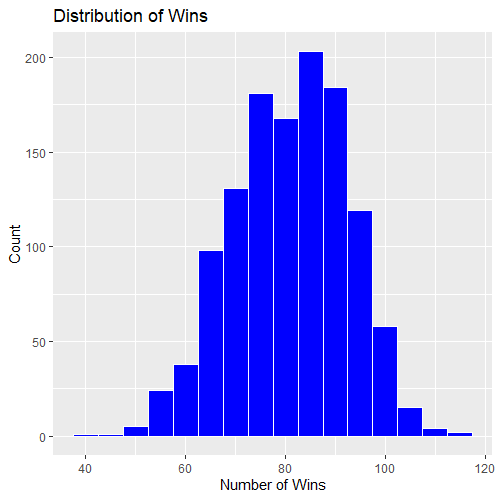
**For the interaction between Grow-light and Plant Food**, the F-value was 1.438 with a p-value of 0.26482. Since the p-value is greater than the significance level (α = 0.05), we do not reject the null hypothesis. This indicates that there is no significant interaction effect of Grow-light and Plant Food on plant growth.

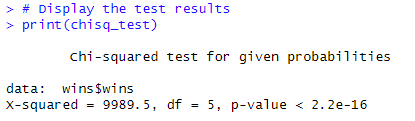
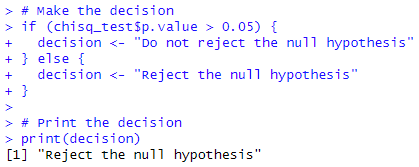
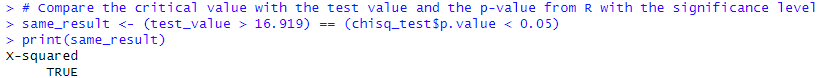
1. **Hoc Analysis  
   A screenshot of a computer code

   Description automatically generated with medium confidence**For the significant factor, Plant Food, we conducted a post-hoc analysis using Tukey's HSD test. The test showed that the difference in mean plant growth between Plant Food A and Plant Food B was significant, with a mean difference of -2.07 and a p-value of 0.0011124.
2. **Conclusion  
   **Based on the results of our analysis, we can conclude that the type of Plant Food has a significant effect on plant growth, while the strength of the Grow-light does not. Moreover, there is no significant interaction effect between Grow-light and Plant Food on plant growth.

**Task 9: Baseball Dataset Analysis**

1. **Data Exploration:**A picture containing text, font, number, screenshot

   Description automatically generated  
   The dataset consists of 1232 rows and 15 columns. It contains information about different baseball teams' performance across different years, including details such as the number of Runs Scored (RS), Runs Allowed (RA), Wins (W), On-Base Percentage (OBP), Slugging Percentage (SLG), Batting Average (BA), whether they made the playoffs, and their ranks if they did. The exploration showed that the data varies across different columns.
2. **Exploratory Data Analysis:**  
   The histogram plotted showed the distribution of the number of wins across different teams and years.
3. **Extracting Decade and Summarizing Wins:**A screenshot of a computer code

   Description automatically generated with medium confidence  
   A new column "Decade" was created by subtracting the remainder of the year divided by 10 from the year. Then, the total number of wins was summarized for each decade.
4. **Chi-Square Goodness-of-Fit Test:**  
   The null hypothesis (H0) is that the number of wins is equally distributed across decades. The alternative hypothesis (H1) is that the number of wins is not equally distributed across decades. The Chi-Square test resulted in a test statistic of approximately 9989.5, which is significantly higher than the critical value (approximately 16.919 for df=5, α=0.05).
5. **Conclusion**:  
     
   Based on the Chi-Square test, the null hypothesis was rejected, indicating that there is a significant difference in the number of wins by decade.
6. **Consistency Check:**  
     
   Comparing the critical value with the test value and the p-value from R with the significance level (α=0.05), both comparisons gave the same result. This provides additional assurance that the Chi-Square test was conducted correctly.

**Task 10: Crop Data Analysis**

1. **Data Preparation:**A black text on a white background

   Description automatically generated with low confidence  
   The data file crop\_data.csv was loaded into R and the variables density, fertilizer, and block were converted into factors.
2. **Hypotheses:  
     
   Null Hypothesis (H0):** Fertilizer and density have no impact on yield.

**Alternative Hypothesis (H1):** Fertilizer and/or density have an impact on yield.  
The histogram plotted showed the distribution of the number of wins across different teams and years.

1. **Two-way ANOVA Test Results:**A screenshot of a computer

   Description automatically generated with low confidence  
   **fertilizer:** The p-value is 0.000273, which is less than the significance level (0.05). This suggests that different levels of fertilizer have a significant effect on crop yield.

**density:** The p-value is 0.000186, which is less than the significance level (0.05). This suggests that different densities have a significant effect on crop yield.

**fertilizer:density:** The p-value is 0.532500, which is greater than the significance level (0.05). This suggests that the interaction of different densities and levels of fertilizer does not have a significant effect on crop yield.

1. **Conclusions:**  
   Based on the results of the Two-way ANOVA test, we reject the null hypothesis for both fertilizer and density but fail to reject it for their interaction term fertilizer:density. This indicates that while both fertilizer and density individually have a significant impact on yield, their interaction does not appear to have a significant impact within this dataset.