Northeastern University
College of Professional Studies

ALY 6015 Intermediate Analytics Module 2 Assignment

Class ALY6015 – Intermediate Analytics

Module 2 Assignment: Chi Square and ANOVA

Overview and Rationale

In this assignment, you will use your knowledge of chi-square and ANOVA testing to solve various types of problems.

Module Outcomes

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

- Use "R" effectively to process, analyze and depict data
- Develop more advanced models to interpret data
- Prepare complex dataset for analysis

Submission Requirements

1. Complete paperwork in MS Word format (.docx) must include:

- Title Page
    - Your name (as registered in Canvas)
    - Assignment name
    - Class number, name and CRN Number
    - Your contact information (NEU email)
- Assignment summary section.(Explain assignment summary, plans, goals, dataset).
- Each step of the research with supporting screenshots, charts, results generated by R code.
- Explain each screen shot from the data standpoint.
- Each output generated by R code must be present and explained in the paperwork.
- Each output, chart, table, screenshot shown in the paperwork must have corresponding R code that generates it.
- Final conclusions section. (Explain if goals were achieved as expected or not, summary of you findings about analyzed data).
- References (optional)

2. Complete R code file meet the following criteria:

- Submitted in R script (.r file format). Only .r file format will be accepted.
- Each line of code must be commented. (Explain why do you execute this line of code, not what the command does).
- Code must be runnable on any computer. Any errors in executing R code will results in significant points deduction. (*Follow the guidelines provided in the class and user R code examples provided in Canvas*)

3. Submit dataset(s) used in the research.

Assignment Summary

Complete the following problems using R. Be sure to show your work and include the hypothesis tests, the critical values, the computed test values, and the resulting decisions where applicable.

Instructions

Perform the following steps.

1. State the hypotheses and identify the claim.
2. Find the critical value.
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

Use the traditional method of hypothesis testing unless otherwise specified. Assume all assumptions are met.

Task 1: Blood Types

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 $\alpha = 0.10$, can it be concluded that the distribution is the same as that of the general population?

Task 2: On-Time Performance by Airlines

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

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

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

Source: Transtats: OST_R|BTS

Perform the following steps.

1. State the hypotheses and identify the claim.
2. Find the critical value.
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

Use the traditional method of hypothesis testing unless otherwise specified. Assume all assumptions are valid.

Task 3: Ethnicity and Movie Admissions

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

Task 4: Women in the Military

This table lists the numbers of officers and enlisted personnel for women in the military. At $\alpha = 0.05$, 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

Source: New York Times Almanac

Assume that all variables are normally distributed, that the samples are independent, that the population variances are equal, and that the samples are simple random samples, one from each of the populations. Also, for each exercise, perform the following steps.

1. State the hypotheses and identify the claim.
2. Find the critical value.
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

Use the traditional method of hypothesis testing unless otherwise specified.

Task 5: Sodium Contents of Foods

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

Source: The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter

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.

Task 6: Sales for Leading Companies

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

Cereal	Chocolate Candy	Coffee
578	311	261
320	106	185
264	109	302
249	125	689

Cereal	Chocolate Candy	Coffee
237	173	

Source: Information Resources, Inc.

Task 7: Per-Pupil Expenditures

The expenditures (in dollars) per pupil for states in three sections of the country are listed. Using $\alpha = 0.05$, 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		

Source: New York Times Almanac

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.)*

Task 8: Increasing Plant Growth

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 $\alpha = 0.05$.

Task 9: Baseball Dataset Analysis

Use R to complete the following steps. Assume the expected frequencies are equal and $\alpha = 0.05$.

1. Download the file base baseball.csv from the course resources and import the file into R.
2. Perform EDA on the imported data set. Write a paragraph or two to describe the data set using descriptive statistics and plots. Are there any trends or anything of interest to discuss?
3. Assuming the expected frequencies are equal, perform a Chi-Square Goodness-of-Fit test to determine if there is a difference in the number of wins by decade. Be sure to include the following:
    a. State the hypotheses and identify the claim.
    b. Find the critical value ($\alpha = 0.05$) (From table in the book).
    c. Compute the test value.
    d. Make the decision. Clearly state if the null hypothesis should or should not be rejected and why.
    e. Does comparing the critical value with the test value provide the same result as comparing the p-value from R with the significance level?

Here is some code to get you started. Be sure to import the dplyr and tidyverse packages.

```
# Extract decade from year bb$Decade <- bb$Year - (bb$Year %% 10)

# Create a wins table by summing the wins by decade wins <- bb %>% group_by(Decade) %>%
summarize(wins = sum(W)) %>% as.tibble()
```

Task 10: Crop Data Analysis

Use R to complete the following steps. Assume the expected frequencies are equal and $\alpha = 0.05$.

1. Download the file crop_data.csv from the course resources and import the file into R.
2. Be sure to convert the variables density, fertilizer and block to R factors.
3. Include a null and alternate hypothesis for both factors and the interaction
4. Perform a Two-way ANOVA test using yield as the dependent variable and fertilizer and density as the independent variables. Explain the results of the test. Is there reason to believe that fertilizer and density have an impact on yield?

Chi Square and ANOVA Assignment Rubric

Chi Square and ANOVA Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Analysis	**25 to >23.25 pts** **Above Standards** Incorporates R code and the outputs. Uses the correct statistical test for the problem and obtains the correct results. Provides detailed analysis of the output focusing on significance results. Uses visualizations to make major points.	**23.25 to >17.5 pts** **Meets Standards** Provides all R code and the outputs. Uses the correct statistical method for the problem, performs the steps correctly. Includes interpretation of the output, graphs, figures, charts, and tables and the significance of the results in the analysis.	**17.5 to >15.0 pts** **Approaching Standards** Provides R codes and outputs, but the R code does not match the outputs or is missing some code or outputs. Uses the correct statistical test for the problem, but does not perform steps correctly or obtains incorrect results. Includes limited interpretations, charts, and tables and the significance of the results in the analysis.	**15 to >0 pts** **Below Standard** Does not use the correct statistical test for the problem. The conclusion does not summarize or attempt to make sense of the results. Conclusions do not reflect an understanding or reflect a misunderstanding of the material.	25 pts

Chi Square and ANOVA Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Interpretation	**25 to >23.25 pts** **Above Standards** Wraps up the findings in a conclusion that provides an answer to the question(s) posed in the introduction. Makes specific recommendations based on the data presented.	**23.25 to >17.5 pts** **Meets Standards** The conclusion summarizes and makes sense of the results, making good points that reflect clear understanding of the assignment material.	**17.5 to >15.0 pts** **Approaching Standards** The conclusion summarizes and makes sense of the results, making good points that reflect a basic understanding of the assignment material.	**15 to >0 pts** **Below Standard** Does not provide R code or its outputs or minimal R code is provided. Includes few interpretations, charts, or tables. Does not identify the significance of the results in the analysis.	25 pts

Chi Square and ANOVA Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Data Visualizations	**25 to >23.25 pts** **Above Standards** Data visualizations are appropriate for the level and type of analysis. Graphs, figures and tables communicate insights and significance to the reader. Graphs are well formatted and labeled	**23.25 to >17.5 pts** **Meets Standards** Data visualizations are useful for the level and type of analysis. Graphs, figures and tables communicate the significance of the results to the reader.	**17.5 to >15.0 pts** **Approaching Standards** Data visualizations are useful for the level and type of analysis, but graphs, figures and tables do not clearly communicate the significance of the results to the reader.	**15 to >0 pts** **Below Standard** Data visualizations are used minimally or not at all. If graphs, figures and tables are used, it is unclear what they are intended to communicate or why.	25 pts

Chi Square and ANOVA Assignment Rubric

Criteria	Ratings				Pts
This criterion is linked to a Learning Outcome Writing Mechanics, Title Page, & References	**25 to >23.25 pts** **Above Standards** There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references.	**23.25 to >17.5 pts** **Meets Standards** There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references.	**17.5 to >15.0 pts** **Approaching Standards** There are very few errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references.	**15 to >0 pts** **Below Standard** There are more than five errors in grammar, spelling, and punctuation; or the usage of title page, citations, and references are incomplete.	25 pts
					Total Points: 100