**CSC 435 LAB 03 Points: 20**

**T-test, Chi-Square, & ANOVA**

Your Team Member names:

1. Stryder R. Coleman

2. Padraic Darby

This document consists of data on how much and how frequently students spend using their zip cards. The data has many demographic variables such as gender, level of academic career, student type (full time, half time etc.), enrolled hours, college they belong to. The data has four continuous variables such as age, hours enrolled, how much they spend using their zip card and how often they use their zip card for purchases.

**Purpose:**

This lab introduces hypothesis testing concepts using R. After completing the tasks in this lab, you should be able to:

* State Null and Alternate Hypothesis based on the research question.
* Select an appropriate test and run the analysis – (independent sample t-test, One-Way ANOVA, Two-Way ANOVA, & Chi-Square)
* Test hypotheses based on the results from the analysis
* Answer the research/business question.

**Lab Instructions:**

1. **Compare** the amount spent (TRANSACTION\_AMOUNT) based on gender (Only men and women. This data set has an additional category called “Unknown”. You need to exclude it. Test if average amount spent by men is greater than the average amount spent by women.
2. Type of Test you would perform for this analysis (Ex: **T-Test**, Chi-Square, Oneway-Anova, Twoway-Anova). **Copy the code and the result below.**

t.test(TRANSACTION\_AMOUNT ~ GENDER, data = df, var.equal = TRUE, alternative = "greater")

Two Sample t-test

data: TRANSACTION\_AMOUNT by GENDER

t = -3.4369, df = 24044, p-value = 0.9997

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-38.47698 Inf

sample estimates:

mean in group Female mean in group Male

269.2269 295.2495

1. State Null and Alternate hypotheses

Null hypothesis: true difference in means is less than or equal to 0

alternative hypothesis: true difference in means is greater than 0

1. What p-value would you use to test the hypotheses & what decision about the hypotheses would you make based on the p-value?

a = .05

**p-value = 0.9997**

Reject the Null hypothesis.

1. What is your conclusion?

There is a significant difference between the average spent by males and females with men spending more than females on average.

1. Is ACAD\_ORG (colleges the students belong to) dependent on Gender? Test the hypothesis for the following five Colleges.

SUNY at Brockport (4628)

College of Business Administration (1959)

College of Education (2091)

College of Engineering (1858)

College of Health Professions (1758)

1. Type of Test you would perform for this analysis (Ex: T-Test, **Chi-Square**, Oneway-Anova, Twoway-Anova). **Copy the code and the result below.**

chisq.test(df$GENDER, df$ACAD\_ORG)

Pearson's Chi-squared test

data: df$GENDER and df$ACAD\_ORG

X-squared = 2133.7, df = 14, p-value < 2.2e-16

1. State Null and Alternate hypotheses

Null: Gender and ACAD\_ORG are independent of each other

Alternative: Gender and ACAD\_ORG are dependent

1. What p-value would you use to test the hypotheses? What decision about the hypotheses would you make based on the p-value?

a = .05

p-value < 2.2e-16

Reject Null Hypothesis

1. What is your conclusion?

Gender and Acad\_Org are dependent of eachother.

1. Test if there is a difference in the average Transaction Amount (TRANSACTION\_AMOUNT) based on the academic level (ACAD\_LEVEL). Use **Freshmen, Sophomore, Junior, and Senior from the ACAD\_LEVEL**. If the model is significant (or reject null hypothesis), perform pairwise comparison using Tukey test and identify the pairs with statistically significant differences.
2. Type of Test you would perform for this analysis (Ex: T-Test, Chi-Square, **Oneway-Anova**, Two way-Anova). **Copy the code and the result below.**

af <- df %>% filter(ACAD\_LEVEL %in% c("Freshman","Sophomore","Junior","Senior"))

> res.aov <- aov(af$TRANSACTION\_AMOUNT ~ af$ACAD\_LEVEL)

> res.aov

Call:

aov(formula = af$TRANSACTION\_AMOUNT ~ af$ACAD\_LEVEL)

Terms:

af$ACAD\_LEVEL Residuals

Sum of Squares 92082185 6732971974

Deg. of Freedom 3 17285

Residual standard error: 624.1209

Estimated effects may be unbalanced

> TukeyHSD(res.aov)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = af$TRANSACTION\_AMOUNT ~ af$ACAD\_LEVEL)

$`af$ACAD\_LEVEL`

diff lwr upr p adj

Junior-Freshman 33.39949 -1.419354 68.21833 0.0656114

Senior-Freshman -145.58027 -178.450934 -112.70961 0.0000000

Sophomore-Freshman 44.78296 12.124763 77.44115 0.0024141

Senior-Junior -178.97976 -216.587464 -141.37206 0.0000000

Sophomore-Junior 11.38347 -26.038673 48.80561 0.8628642

Sophomore-Senior 190.36323 154.746585 225.97987 0.0000000

1. State Null and Alternate hypotheses

Null: The means of the different groups are equal

Alternative: At least one of the means are not equal to the others

1. What p-value would you use to test the hypotheses? What decision about the hypotheses would you make based on the p-value?

a = .05

Junior-Freshman 0.0656114

Senior-Freshman 0.0000000

Sophomore-Freshman 0.0024141

Senior-Junior 0.0000000

Sophomore-Junior 0.8628642

Sophomore-Senior 0.0000000

Reject the null hypothesis

1. Which pairs of students have statistically significant differences in their spending?

Senior-Freshman, Senior-Freshman, Senior-Junior, and Sophomore-Senior differ significantly in average spending from each other.

1. Test if there is a difference in average Age of the student based on Gender (Male/Female) and Academic load (ACAD\_LOAD) (use students who are enrolled full-time, half-time, & three quarter time). Also, check whether the interaction is significant.
2. Type of Test you would perform for this analysis (Ex: T-Test, Chi-Square, One-way Anova, **Two-way Anova**). **Copy the code and the result below.**

**res.aov2 <- aov(d4$AGE ~ d4$GENDER + d4$ACAD\_LOAD)**

**summary(res.aov2)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**d4$GENDER 1 1865 1865 47.21 6.54e-12**

**d4$ACAD\_LOAD 2 88461 44231 1119.78 < 2e-16**

**Residuals 21878 864170 39**

**---**

**Signif. codes: 0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**TukeyHSD(res.aov2)**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = d4$AGE ~ d4$GENDER + d4$ACAD\_LOAD)**

**$d4$GENDER**

**diff lwr upr p adj**

**Male-Female -0.5839522 -0.7505325 -0.4173718 0**

**$d4$ACAD\_LOAD**

**diff lwr upr p adj**

**Enrolled Half-Time-Enrolled Full-Time 5.716742 5.430108 6.003377 0**

**Three Quarter Time-Enrolled Full-Time 1.920278 1.585782 2.254774 0**

**Three Quarter Time-Enrolled Half-Time -3.796464 -4.206112 -3.386817 0**

1. What are p-values and what decision would you make based on the p-values?

a = 0.5

**Male-Female 0**

**Enrolled Half-Time-Enrolled Full-Time 0**

**Three Quarter Time-Enrolled Full-Time 0**

**Three Quarter Time-Enrolled Half-Time 0**

Reject the Null hypothesis.

1. What is your overall conclusion (not statistical term)?

Each variable differs significantly from the other variables.