BIA 654 B: Experimental Design II

**Ephraim Schoenbrun**

CWID: 104-20-702

Date of Submission: **February 21, 2017**

Assignment:

**Assignment #4**

# **Ethical Conduct**

|  |
| --- |
| The following statement is printed in the Stevens Graduate Catalog and applies to all students taking Stevens courses, on and off campus.  “Cheating during in-class tests or take-home examinations or homework is, of course, illegal and immoral. A Graduate Academic Evaluation Board exists to investigate academic improprieties, conduct hearings, and determine any necessary actions. The term ‘academic impropriety’ is meant to include, but is not limited to, cheating on homework, during in-class or take home examinations and plagiarism.”  Consequences of academic impropriety are severe, ranging from receiving an “F” in a course, to a warning from the Dean of the Graduate School, which becomes a part of the permanent student record, to expulsion.  *Reference: The Graduate Student Handbook, Academic Year 2003-2004 Stevens*  *Institute of Technology, page 10.*  Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.  Signature: Date: February 21, 2017  Please note that assignments in this class may be submitted to [www.turnitin.com](http://www.turnitin.com/), a web-based anti-plagiarism system, for an evaluation of their originality. |

BIA 654 Homework 4

1. You are conducting a study to see if students do better when they study all at once or in intervals. One group of 12 participants took a test after studying for one hour continuously. The other group of 12 participants took a test after studying for three twenty minute sessions. The first group had a mean score of 75 and a variance of 120. The second group had a mean score of 86 and a variance of 100. Assuming the normal populations, independent samples, and equal population variances conditions hold, are the mean test scores of these two groups significantly different at the 0.05 level?

**Test: Two Sample, pooled-variance t Test**

**Goal: Form a confidence interval for the difference between two population means**

**Point Estimate: or = -11**

**Pooled SP: ((n1-1) \* v1 + (n2-1) \* v2)/df or ((12-1)\*120 + (12-1)\*100)/22 = 10.488**

**Standard Error: Pooled SP \* sqrt((1/n1 + 1/n2)) = 4.282**

**TSTAT: (mean1-mean2-mean:null)/se = -2.569**

**95% Confidence interval (t-computed): 2.074**

**The two values are significantly different since the t-statistic exceeds the t computed for the degrees of freedom the distribution would account for.**

R Code

m0 <- 0

m1 <- 75

m2 <- 86

v1 <- 120

v2 <- 100

n1 <- 12

n2 <- 12

df <- n1+n2-2

pvar <- ((n1-1)\*v1 + (n2-1)\*v2)/df

psd <- sqrt(pvar)

se <- psd\*sqrt((1/n1 + 1/n2))

tstat <- (m1-m2-m0)/se

tcrit <- pt(t, .975, df = 22)

dist <- se \* tcrit

pest <- m1-m2

ci <- c(pest + dist, pest - dist)

1. Find a file NHANES.dat. It contains a large data set related to patients profile and their health status. Suppose we are interested in whether BMI (Body Mass Index) is different between Male and Female groups. Is there an evidence of equal variance at significance level 0.05? Perform an appropriate two-sample t-test.

**Test: F test (too sensitive to non-normal distribution) and Levene test (more robust)**

**Assumptions: Not normally distributed (Shapiro Wilks: Male - W = 0.92699, p-value < 2.2e-16; Female: W = 0.94261, p-value < 2.2e-16; and qqplot shows not normal distribution)**

**F test**

**data: male$BMI and female$BMI**

**F = 0.67806, num df = 2662, denom df = 2724, p-value < 2.2e-16**

**Alternative hypothesis: true ratio of variances is not equal to 1**

**95% Confidence Interval: Between 0.629 and 0.731**

**sample estimates: ratio of variances: 0.678**

**Result: Reject null hypothesis that they are equal variances**

**Levene test - absolute deviations from the mean**

**Test Statistic = 102.66, p-value < 2.2e-16**

**Result: Reject null hypothesis that they are equal variances**

R Code

library(tidyverse)

library(plyr)

library(foreign)

library(lattice)

library(graphics)

library(lawstat)

data <- read.csv("C:/Users/Stevens/Desktop/R/NHANES\_for\_PopHealthPortal.dat")

data <- rename(data, replace = c("Body.Mass.Index..kg.m..2."="BMI"))

data2 <- select(data, BMI, Gender)

data <- subset(data2, !is.na(data$BMI))

male <- subset(data, Gender == "Male", "BMI")

female <- subset(data, Gender == "Female", "BMI")

mmean <- mean(male$BMI)

fmean <- mean(female$BMI)

msd <- sd(male$BMI)

fsd <- sd(female$BMI)

require(graphics)

qqnorm(male$BMI)

qqline(male$BMI, datax = FALSE, distribution = qnorm, probs = c(0.25, 0.75), qtype = 7)

shapiro.test(male$BMI)

qqnorm(female$BMI)

qqline(female$BMI, datax = FALSE, distribution = qnorm, probs = c(0.25, 0.75), qtype = 7)

shapiro.test(female$BMI)

ftest <- var.test(male$BMI, female$BMI, ratio = 1, conf.level = .95)

levene.test(data$BMI, data$Gender, location="mean")

1. The female cuckoo lays her eggs into the nests of foster parents. The foster parents are usually deceived, probably because of the similarity in the sizes of the eggs. Lengths of cuckoo eggs (in millimeters) found in the nests of hedge sparrows, robins, and wrens are shown below:
   * Hedge sparrow: 22.0, 23.9, 20.9, 23.8, 25.0, 24.0, 21.7, 23.8, 22.8, 23.1, 23.1, 23.5, 23.0, 23.0
   * Robin: 21.8, 23.0, 23.3, 22.4, 23.0, 23.0, 23.0, 22.4, 23.9, 22.3, 22.0, 22.6, 22.0, 22.1, 21.1, 23.0
   * Wren: 19.8, 22.1, 21.5, 20.9, 22.0, 21.0, 22.3, 21.0, 20.3, 20.9, 22.0, 20.0, 20.8, 21.2, 21.0

It is believed that the size of the egg influences the female cuckoo in her selection of the foster parents. Do the data support this hypothesis? Test whether or not the mean lengths of cuckoo eggs found in the nests of the three foster-parent species are the same. (Here, don’t forget to check each of the underlying ‘assumptions.’)

**Test: Completely Random Design**

**Equal variances: Levene test - Test Statistic = Test Statistic = 0.70873, p-value = 0.4981 -> Equal**

**Normal distribution: Shapiro Wilks – Sparrow: W = 0.9556, p-value = 0.6505; Robin: W = 0.96526, p-value = 0.7571; Wren: W = 0.94185, p-value = 0.4062**

**All Normally Distributed and have Equal Variance**

**Null Hypothesis: µSparrow = µRobin = µWren**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Df** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr(>F)** |
| **group** | **2** | **31.112** | **15.556** | **22.329** | **2.48E-07** |
| **Residuals** | **42** | **29.261** | **0.6967** |  |  |

**Result: Reject null hypothesis**

R Code

install.packages(c("tidyr", "devtools"))

install.packages("lawstat")

sparrow <- c(22.0, 23.9, 20.9, 23.8, 25.0, 24.0, 21.7, 23.8, 22.8, 23.1, 23.1, 23.5, 23.0, 23.0)

Robin <- c(21.8, 23.0, 23.3, 22.4, 23.0, 23.0, 23.0, 22.4, 23.9, 22.3, 22.0, 22.6, 22.0, 22.1, 21.1, 23.0)

Wren <- c(19.8, 22.1, 21.5, 20.9, 22.0, 21.0, 22.3, 21.0, 20.3, 20.9, 22.0, 20.0, 20.8, 21.2, 21.0)

sparrowdf <- data.frame(group = "sparrow", mm = sparrow)

robindf <- data.frame(group = "robin", mm = Robin)

wrendf <- data.frame(group = "wren", mm = Wren)

data <- rbind(sparrowdf, robindf, wrendf)

levene.test(data$mm, data$group, location="mean")

shapiro.test(sparrow)

shapiro.test(Robin)

shapiro.test(Wren)

anova <- lm(mm ~ group, data = data)

anova(anova)

aov <- aov(mm ~ group, data= data)

summary(aov)

1. Recall the class note example on three promotions and sales volume difference percentage data (find it on Feb. 9 lecture slide, p.7). Perform one-way ANOVA procedure to test

H0 : µ1 = µ2 = µ3, vs. H1 : Not H0 .

1. Carry this out via hand calculations (it is worth doing this practice at least once!).
2. Now, do it again via statistical software package and compare with the result in (a).

|  |  |  |  |
| --- | --- | --- | --- |
| Promotions | 1 | 2 | 3 |
| Value 1 | 9.5 | 8.5 | 7.7 |
| Value 2 | 3.2 | 9 | 11.3 |
| Value 3 | 4.7 | 7.9 | 9.7 |
| Value 4 | 7.5 | 5 | 11.5 |
| Value 5 | 8.3 | 3.2 | 12.4 |
| size | 5 | 5 | 5 |
|  | 6.64 | 6.72 | 10.52 |
| variance | 6.82 | 6.28 | 3.43 |

**= 7.96**

**Sum of Squares Between Groups (SSB):**

**=**

**+ + =**

**8.712 + 7.688 + 32.768 = 49.168**

**Mean Square Between (MSB):**

**= = 24.584**

**Sum of Squares Within Groups (SSW)**

**=**

**(9.5 - 6.64)^2) + (3.2 - 6.64)^2) + (4.7 - 6.64)^2) + (7.5 - 6.64)^2) + (8.3 - 6.64)^2) = 27.272**

**(8.5 - 6.72)^2) + (9 - 6.72)^2) + (7.9 - 6.72)^2) + (5 - 6.72)^2) + (3.2 - 6.72)^2) = 25.108**

**(7.7 - 10.52)^2) + (11.3 - 10.52)^2) + (9.7 - 10.52)^2) + (11.5 - 10.52)^2) + (12.4 - 10.52)^2) = 13.728**

**27.272 + 25.108 + 13.728 = 66.108**

**Mean Square Within (MSW) =**

**= = 5.509**

**Total Sum of Squares**

**66.108 + 49.168 = 115.276**

**Mean Square Total**

**= = 6.685**

**F Statistic**

**= = 4.463**

**F critical value at α = .05 with df1 = 2, df2 = 12 = 3.8853**

**Since FSTAT of 4.463 > α of 3.8853, we reject the null hypothesis.**

**With R**

**Equal Variance: Levene Test Result = equal**

**Test Statistic = 0.86375, p-value = 0.4462**

**Normal Distribution: Shapiro Wilks Result = normal**

**Promo 1: W = 0.93739, p-value = 0.6475**

**Promo 2:** **W = 0.87967, p-value = 0.3078**

**Promo 3: W = 0.92229, p-value = 0.5448**

**ANOVA test**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Df** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr(>F)** |
| **promo** | **2** | **49.168** | **24.584** | **4.4625** | **0.03557** |
| **Residuals** | **12** | **66.108** | **5.509** |  |  |

R Code

datamaster <- read.csv("C:/Users/Stevens/Desktop/R/promotions.csv", header=TRUE)

promo1df <- data.frame(promo = "1", volume = data$X1)

promo2df <- data.frame(promo = "2", volume = data$X2)

promo3df <- data.frame(promo = "3", volume = data$X3)

data <- rbind(promo1df, promo2df, promo3df)

levene.test(data$volume, data$promo, location="mean")

promo1 <- datamaster[, "X1"]

promo2 <- datamaster[, "X2"]

promo3 <- datamaster[, "X3"]

shapiro.test(promo1)

shapiro.test(promo2)

shapiro.test(promo3)

anova <- lm(volume ~ promo, data = data)

anova(anova)