**Assignment – 9.1**

NOTE: I have cut and paste the r code in this word file, and when it was showing in blue in colour, to bring everything to one font size and colour, it got standardised. But it is the true code with the relevant output

**Question No. 1:**

1. Find the probability of z being greater than 2.64

> pnorm(2.64,0,1,lower.tail = FALSE)

[1] 0.004145301

1. Find the probability of absolute value of z being greater than 1.39

> pnorm(1.39,0,1,lower.tail = FALSE)+ pnorm(-1.39,0,1)

[1] 0.1645289

**Question No. 2:**

Since the data that is given is based on admissions, the acceptance rate is not readily available in the database of UCBAdmissions in r. Therefore there is a need to calculate acceptance rate as Admissions Frequency / Applied Frequency.

Additionally the given database is in the form of a 3 dimensional table, with gender wise admission, and rejection details in each table, and in this manner department wise tables are created. There are 6 departments.

To test whether the proclaimed 40% proportion is true, there are ways in which we can do it. We can test it by aggregating the data at department level, irrespective of the gender, or we can aggregate at department and gender level also. The difference is at department level, the sample size would be 6, at both levels it would be 12. I have attempted to do it at both levels to find out whether the announcement is correct.

> UCBAdmissions

> univ <- apply(UCBAdmissions,c(1,3),sum)

> univ

Dept

Admit A B C D E F

Admitted 601 370 322 269 147 46

Rejected 332 215 596 523 437 668

> univ <- as.data.frame(univ)

# to convert it into a data frame for further analysis

> data <- t(univ)

# to convert the orientation of the data frame into long from wide for convinient columnar manipulation

> data <- as.data.frame(data)

> head(data)

Admitted Rejected

A 601 332

B 370 215

C 322 596

D 269 523

E 147 437

F 46 668

> library(dplyr)

> data <- mutate(data, "Applied" = Admitted + Rejected)

# creating a new column Applied

> data <- mutate(data, "AccRate" = Admitted / Applied)

# creating a new column AccRate

> data

Admitted Rejected Applied AccRate

1 601 332 933 0.64415863

2 370 215 585 0.63247863

3 322 596 918 0.35076253

4 269 523 792 0.33964646

5 147 437 584 0.25171233

6 46 668 714 0.06442577

> prop.test(mean(data$Admitted), mean(data$Applied),0.4,"less", 0.99, TRUE)

1-sample proportions test with continuity correction

data: mean(data$Admitted) out of mean(data$Applied), null probability 0.4

X-squared = 0.42129, df = 1, p-value = 0.2581

alternative hypothesis: true p is less than 0.4

98 percent confidence interval: 0.0000000 0.4253922

sample estimates: p 0.3877596

So the above test shows that the p value is not strong and significant enough at 98% confidence intervals or at 2% significance level (for left of the mean, since the hypothesis is it is significantly less than 40%) to reject the null hypothesis that the acceptance rate is 40% and accept the alternate that it is less than 40%.

Both Department and Gender wise aggregation gives the following result

> dfUCB <- as.data.frame(UCBAdmissions)

> head(dfUCB)

Admit Gender Dept Freq

1 Admitted Male A 512

2 Rejected Male A 313

3 Admitted Female A 89

4 Rejected Female A 19

5 Admitted Male B 353

6 Rejected Male B 207

> library(sqldf)

> sqldf('select \* from dfUCB where Admit like "Admitted"')

Admit Gender Dept Freq

1 Admitted Male A 512

2 Admitted Female A 89

3 Admitted Male B 353

4 Admitted Female B 17

5 Admitted Male C 120

6 Admitted Female C 202

> mfdata <- sqldf('select \* from dfUCB where Admit like "Admitted"')

> mfdata <- cbind(mfdata,sqldf('select Admit,Freq from dfUCB where Admit like "Rejected"'))

> mfdata

Admit Gender Dept Freq Admit Freq

1 Admitted Male A 512 Rejected 313

2 Admitted Female A 89 Rejected 19

3 Admitted Male B 353 Rejected 207

4 Admitted Female B 17 Rejected 8

5 Admitted Male C 120 Rejected 205

6 Admitted Female C 202 Rejected 391

> library(dplyr)

> colnames(mfdata)[c(1,5)] <- c("Offered","Denied")

> colnames(mfdata)[c(4,6)] <- c("FreqOff","FreqDeni")

> mfdata <- mutate(mfdata, "Applied" = FreqOff + FreqDeni)

> mfdata <- mutate(mfdata, "AccRate" = FreqOff / Applied)

> prop.test(mean(mfdata$FreqOff), mean(mfdata$Applied), 0.4,"less", 0.98, TRUE)

1-sample proportions test with continuity correction

data: mean(mfdata$FreqOff) out of mean(mfdata$Applied), null probability 0.4

X-squared = 0.18722, df = 1, p-value = 0.3326

alternative hypothesis: true p is less than 0.4

98 percent confidence interval: 0.0000000 0.4415984

sample estimates: p 0.3877596

The p value in this test also cannot reject the null hypothesis that the acceptance rate is 40%.

This procedure may be wrong, since it is not testing proportions, but testing of means. Yet for cross verification.

Suppose we assume 0.4 as a mean proportion what the University is stating, then for the given sample of department and gender based aggregation, we can get the mean and standard deviation of the acceptance rate, and assuming that this variable is normally distributed we can also test for the hypothesis like a z score to find the probability of the acceptance rate being less than 40%.

> pnorm(0.4,mean(mfdata$AccRate),sd(mfdata$AccRate))

[1] 0.5012134

> mean(mfdata$AccRate)

[1] 0.3992677

> sd(mfdata$AccRate)

[1] 0.2407534

This shows that it is 50.12%, that means that we are not confident 99% that it is less than 40%, so again we reject our alternate view that the acceptance rate may be less than 40% and have to accept that it is 40% as stated by the university. Further with the mean and sd we can create the confidence interval approximately as follows.

0.3992677 + (2.576 \* 0.2407534) = [1] 1.019448

and

0.3992677 – (2.576 \* 0.2407534) = [1] -0.2209131

The range shows very clearly that it is too wide, with a positive value greater than 0.4, for confidently concluding that the acceptance rate is 40%.