DATS 6101 HW 4

Chris Montgomery

2/11/2019

```
## [1] "/Users/chrismontgomery/Documents/active_projects/DATS6101_DataScience"
```

Q1.

```
#Construct t-intervals for the gre and gpa data for all applicants at 0.80 level #and 0.99 level.
conf.level <- c(.8, .99)
variable <- c("gre", "gpa")
label <- c()
interval <- c()
a <- 1
for (i in variable){
  for (j in conf.level){
    label[a] <- paste( i, toString(j), sep = "-")
    t.interval <- t.test(df[,i], conf.level = j)$conf.int
    print(paste( label [a],  t.interval))
 a <- a + 1
  }
}
```

```
## [1] "gre-0.8 580.285704407918" "gre-0.8 595.114295592082"
## [1] "gre-0.99 572.750963649866" "gre-0.99 602.649036350134"
## [1] "gpa-0.8 3.36547376090035" "gpa-0.8 3.41432623909965"
## [1] "gpa-0.99 3.34065071776371" "gpa-0.99 3.43914928223629"
```

Q2

```
#Repeat the same calculation in Question 1 but for admitted (1) and rejected (0) #separately
admit <- c("admitted", "rejected")
conf.level <- c(.8, .99)
variable <- c("gre", "gpa")
label <- c()
interval <- c()
a <- 1

for (i in variable){
  for (j in conf.level){
    for(k in 0:1){

    label[a] <- paste( i, toString(j),k, sep = "-")
    t.interval <- t.test(df[df$admit ==k ,i], conf.level = j)$conf.int
    print(paste( label [a],  t.interval))
 a <- a + 1
  }
```

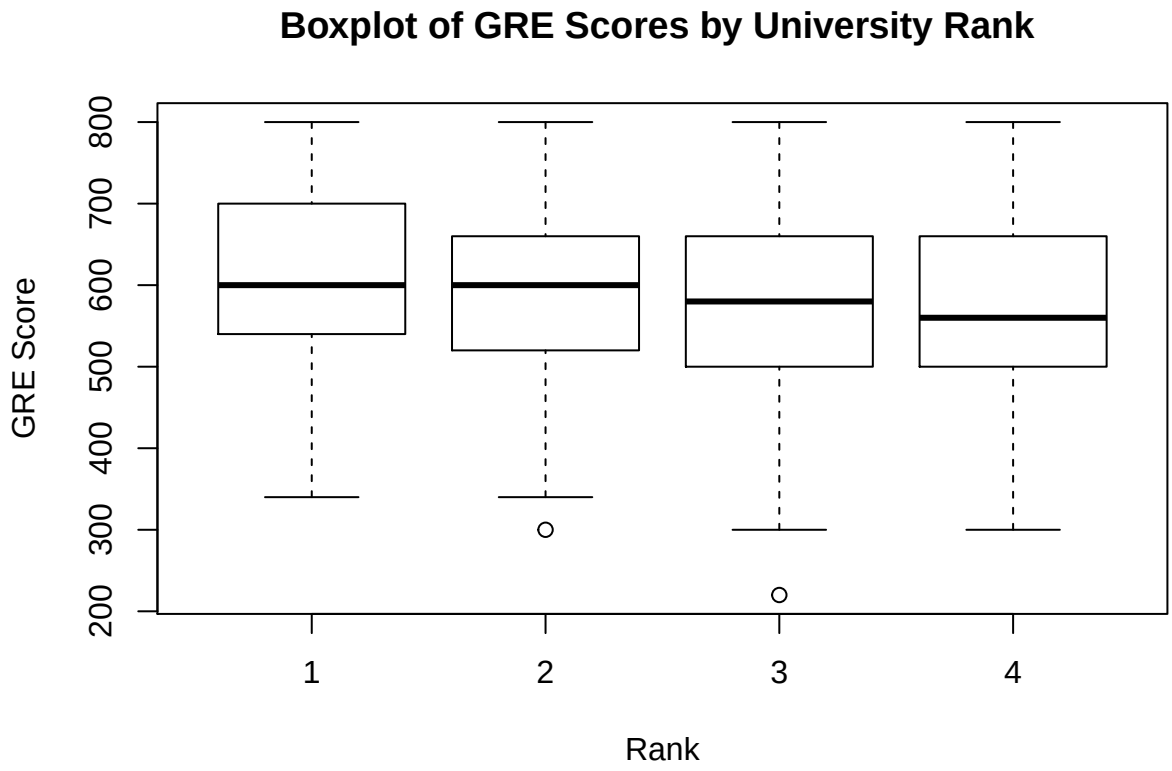
```
}
}
```

```
## [1] "gre-0.8-0 564.180795137426" "gre-0.8-0 582.192831236201"
## [1] "gre-0.8-1 606.450056868856" "gre-0.8-1 631.345218721695"
## [1] "gre-0.99-0 555.001751539152" "gre-0.99-0 591.371874834474"
## [1] "gre-0.99-1 593.627601465313" "gre-0.99-1 644.167674125239"
## [1] "gpa-0.8-0 3.31437683978476" "gpa-0.8-0 3.3730224276145"
## [1] "gpa-0.8-1 3.44689441951894" "gpa-0.8-1 3.53153077733145"
## [1] "gpa-0.99-0 3.2844906900689"  "gpa-0.99-0 3.40290857733037"
## [1] "gpa-0.99-1 3.40330177569584" "gpa-0.99-1 3.57512342115455"
```

Q3

```
#Make (box-) plots showing the gre distribution among applicants from different #school rankings.

boxplot(df$gre ~ df$rank, main = "Boxplot of GRE Scores by University Rank ", xlab = "Rank", ylab = "GR
```




```
?boxplot
```

Q4.

```
t.test(df[df$admit == 0, "gre"], df[df$admit == 1, "gre"])
```

```
##
##  Welch Two Sample t-test
```

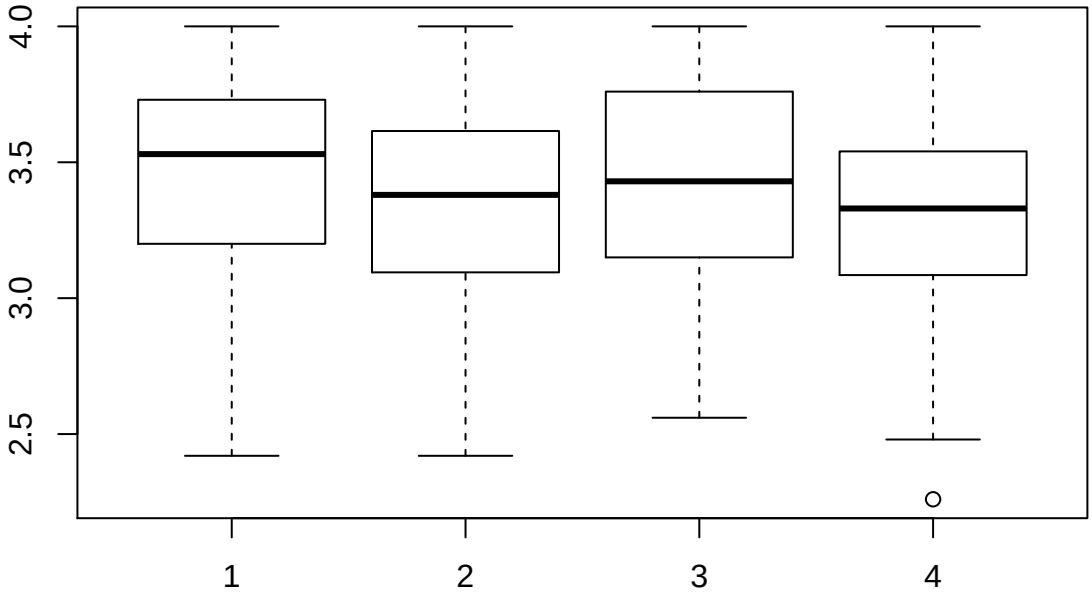
```
##
## data:  df[df$admit == 0, "gre"] and df[df$admit == 1, "gre"]
## t = -3.8292, df = 260.18, p-value = 0.0001611
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -69.21683 -22.20482
## sample estimates:
## mean of x mean of y
##  573.1868  618.8976
```

The above two sample t-test asseses whether the average GRE scores for admitted and rejected students significantly differ. Formally, we are testing the null hypothesis H_0: $\mu_{rejected} = \mu_{accepted}$ with μ representing mean GRE scores for each group. As such, we determine the alternative hypothesis to be $H_1 : \mu_{accepted} \neq \mu_{rejected}$. According to the results of the above t-test, we can formally reject the null hypothesis at the $> 99\%$ confidence level. Given a t-statistic of -3.83 and p-value of .0002, we say there is a less than 1% chance the difference between sample means arose from random sampling. Given the sign of the t-stat and confidence interval, we can say that the average score for accepted students is between 22 and 69 points higher than the rejected average.

Q5

```
#Repeat questions 3 and 4 for the gpa data. Explain whether the result is #reasonable to you or not.

boxplot(df$gpa~df$rank)
```




```
# Compute the analysis of variance
anova <- aov(gpa ~ rank, data = df)
# Summary of the analysis
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## rank          1   0.19  0.1908   1.318  0.252
## Residuals   398  57.60  0.1447
```

The above boxplot appears to show limited variation within the GPA distributions for students grouped

by undergraduate intsitutional ranking. This seems fairly intuitive, as GPA, unlike GRE scores, tend to be determined by within university characteristics. Put simply, the low variation in GPA between groups can be explained by the fact taht low ranking institutions can give students 4.0s just as top tier institutions can fail their students. Likewise, the above finding may show that students of varying GPAs apply to grad school at roughly the same proportions, regardless of undergraduate institution ranking. The above anova test confirms that variation in GPA scores between groups is not statistically significant.

```
t.test(df[df$admit == 0, "gpa"], df[df$admit == 1, "gpa"])
```

```
##
##  Welch Two Sample t-test
##
## data:  df[df$admit == 0, "gpa"] and df[df$admit == 1, "gpa"]
## t = -3.6379, df = 250.05, p-value = 0.0003339
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2242921 -0.0667338
## sample estimates:
## mean of x mean of y
##  3.343700  3.489213
```

The above t-test demonstrates that the average gpa of admitted students significantly differs from rejected students. Based on the direction of the t-statistics and reported 95% confidence interval, we can say that we are 95% confident the mean gpd for admitted students was $[.067, .22]$ points higher than the rejected student mean. This finding should be fairly intuitive, if we believe that gpa is a considering factor in the college admission process.

Q6

```
#Construct a contingency table between admit and rank.
table1 <- table(df$admit,df$rank, dnn = c("Rejected/Admitted", "Rank")); table1
```

```
##                  Rank
## Rejected/Admitted  1  2  3  4
##                 0 28 97 93 55
##                 1 33 54 28 12
```

Q7

```
chi.test <- chisq.test(table1)
chi.test$residuals
```

```
##                  Rank
## Rejected/Admitted          1          2          3          4
##                 0 -2.1128042 -0.5966967  1.1463560  1.3712231
##                 1  3.0976955  0.8748490 -1.6807339 -2.0104236
```

```
chi.test
```

```
##
##  Pearson's Chi-squared test
##
## data:  table1
```

```
## X-squared = 25.242, df = 3, p-value = 1.374e-05
```

To determine whether applicant undergraduate institution rank and rejection status are distributed independently, we can employ a chi-square test. With a chi-square test, we test the null hypotheses $H_0 : P(Rejection|Rank) = P(Rejection)$. Put simply, we are testing whether there is any statistically significant relationship between school rank and admission status. The results of the above test demonstrate that we reject the null hypothesis (at a confidence level $> 99\%$). As such, we accept the alternative hypothesis that admission status is dependent upon undergraduate institution rank. Visual inspection of the residuals suggests that lowest ranked universities are underrepresented (overrepresented) in the admitted category, while top tier universities are underrepresented (overrepresented) in the rejected (accepted) category. This should be fairly intuitive, as we assume that the quality of an undergraduate institution has some effect on the likelihood of admittance to a graduate institution.