Anagha Uppal - MW Section - Homework 2

Associations pt 1

by Sept 8, 2015 3:30pm

Note: since knitting and .Rmd file creates its own R session, you will need to have the code `library(regclass)` at some point in the code before referring to any of its commands. In the homework and activities I will do this for you (see the code chunk above), but it's something to keep in mind.

Question 1:

Download HW2-catalog.dat into your BAS320 folder and use `read.csv` to read it in R, calling it `CAT`. This has data from 4000 customers who have purchased items through a mail-order catalog. It is important for businesses to know who is likely to order again so they do not waste time/money sending out catalogs to people who won't buy.

```
#Replace with command reading in HW2-catalog.dat
CAT <- read.csv("HW2-catalog.dat")
```

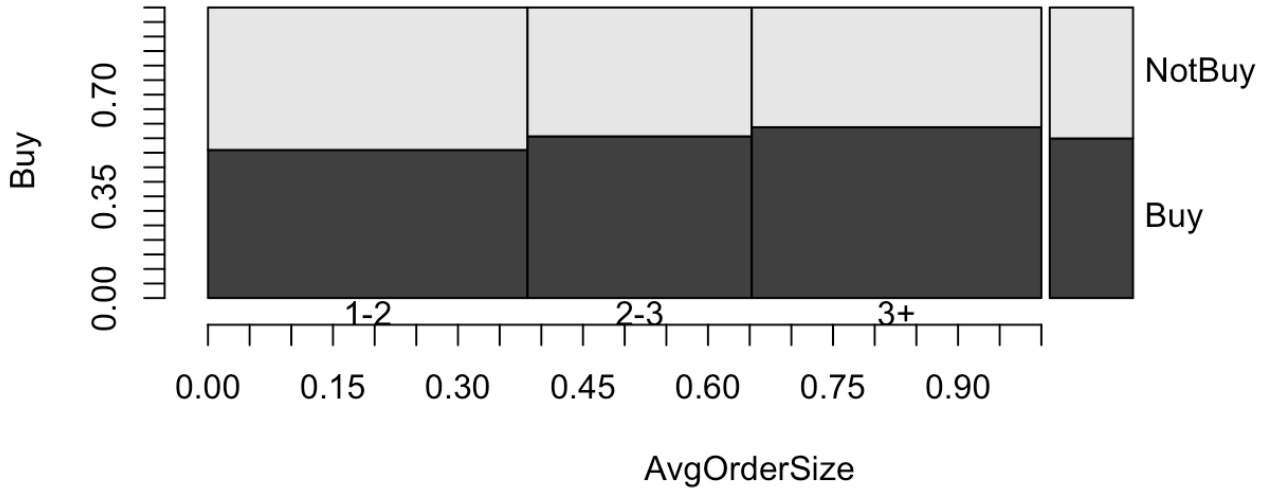
a. Let us study a potential association between the average number of items in previous orders (`AvgOrderSize`) and whether a customer makes a purchase in the next sales quarter (`Buy`). Which variable should play the role of x and which should play the role of y?

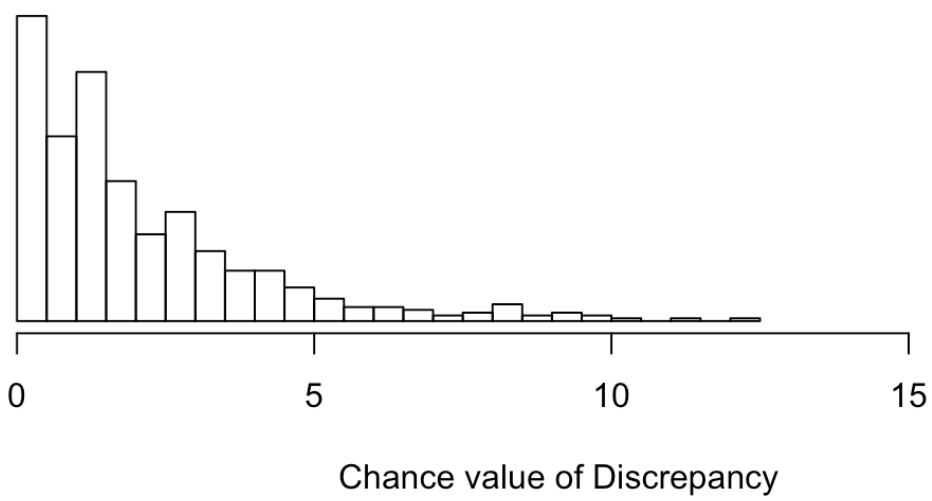
Response: x = AvgOrderSize; y = Buy

---

b. Setup and run the `associate` command. Use the default number of permutations but add `seed=320`.

```
#Code for running associate with seed=320
associate(Buy~AvgOrderSize, data=CAT, seed=320)
```


Buy
0.00
0.35
0.70
NotBuy
Buy
1-2
2-3
3+
0.00
0.15
0.30
0.45
0.60
0.75
0.90
AvgOrderSize


0
5
10
15
Chance value of Discrepancy

```
## Association between AvgOrderSize (categorical) and  Buy (categorical):
##
##  using 4000 complete cases
## Contingency table:
##        y
## x        Buy NotBuy Total
##   1-2    782    752  1534
##   2-3    599    477  1076
##   3+     817    573  1390
##   Total 2198   1802  4000
##
##  Table of Expected Counts:
##       Buy NotBuy
## 1-2 842.9  691.1
## 2-3 591.3  484.7
## 3+  763.8  626.2
##
## Conditional distributions of y (Buy) for each level of x (AvgOrderSize):
## If there is no association, these should look similar to each other and
##  similar to the marginal distribution of y
##                Buy    NotBuy
## 1-2      0.5097784 0.4902216
## 2-3      0.5566914 0.4433086
## 3+       0.5877698 0.4122302
## Marginal 0.5495000 0.4505000
##
## Permutation procedure:
##   Discrepancy Estimated p-value
##       18.2257                 0
## With 500 permutations, we are 95% confident that:
##  the p-value is between 0 and 0.007
## If 0.05 is in this range, change permutations= to a larger number
```

---

c. Why is it important to specify the `seed` argument in `associate`, i.e., what does allow you (or others) to do?

Response: So that the “random” permutations that are used in the simulations are the same and will result in the same conclusions

---

d. Based on the mosaic plot, would you say that the association is strong or weak? Why?

Response:

---

e. The mosaic plot tells an interesting story about the association between buying and the average order size. Describe an insight gained by looking at the plot.

Response: The greater the previous order size, the more likely a customer is to purchase something in the next sales quarter.

---

f. The segmented bar charts in the mosaic plot don't match up exactly (implying that the proportion of customers who buy varies with their average order size), but variation is expected due to the data collection process. Based on histogram of discrepancies from the permutation datasets, is chance alone enough to explain the differences in the segmented bar charts and the differences between the observed and expected counts of the contingency table? Why or why not?

Response: No. The D value of 18.2 was highly unlikely to have occurred by chance.

---

g. Speaking of expected counts, the output of `associate` reveals that we should have expected 626.2 customers who typically buy 3+ items per order to not buy in the next quarter. Show where this number comes from, and what is the actual number of customers with these characteristics from the data?

```
#Show how the expected count is calculated
# If there was no association between the variables, one would expect distribution
for each to equal marginal distribution.
# E = marg dist * total count
# E = .4505 * 1390 = 626.2
# Actual number: 573
```

---

h. Quote the range of p-values of the association based on your simulation. Is the test conclusive? If yes, is the association statistically significant (why or why not)? If no, what needs to be done to make the test conclusive.

Response: At the 95% confidence level, the p-value falls between 0 and 0.7%, which is very low, and therefore, the association is statistically significant and the test conclusive.

---

i. Now examine associations between `Buy` and the other predictors in the dataset (`TimeSinceLast` and `AvgSpend`). Which associations are statistically significant? Which of the three variables considered (`TimeSinceLast` , `AvgOrderSize` , `AvgSpent`) appears to have the strongest relationship with `Buy` ?

```
#Insert code for running associate.  Note:  to save space, this chunk is written s
o that the
#the plots/analysis are not put in the document.
associate(Buy~AvgOrderSize, data=CAT, seed=320)
associate(Buy~TimeSinceLast, data=CAT, seed=320)
associate(Buy~AvgSpent, data=CAT, seed=320)
```

Response: All associations are statistically significant. p-values are all equal, but discrepancy between expected and actual is extremely high for the association between 'Buy' and 'TimeSinceLast', which suggests the association could not have occurred by chance.

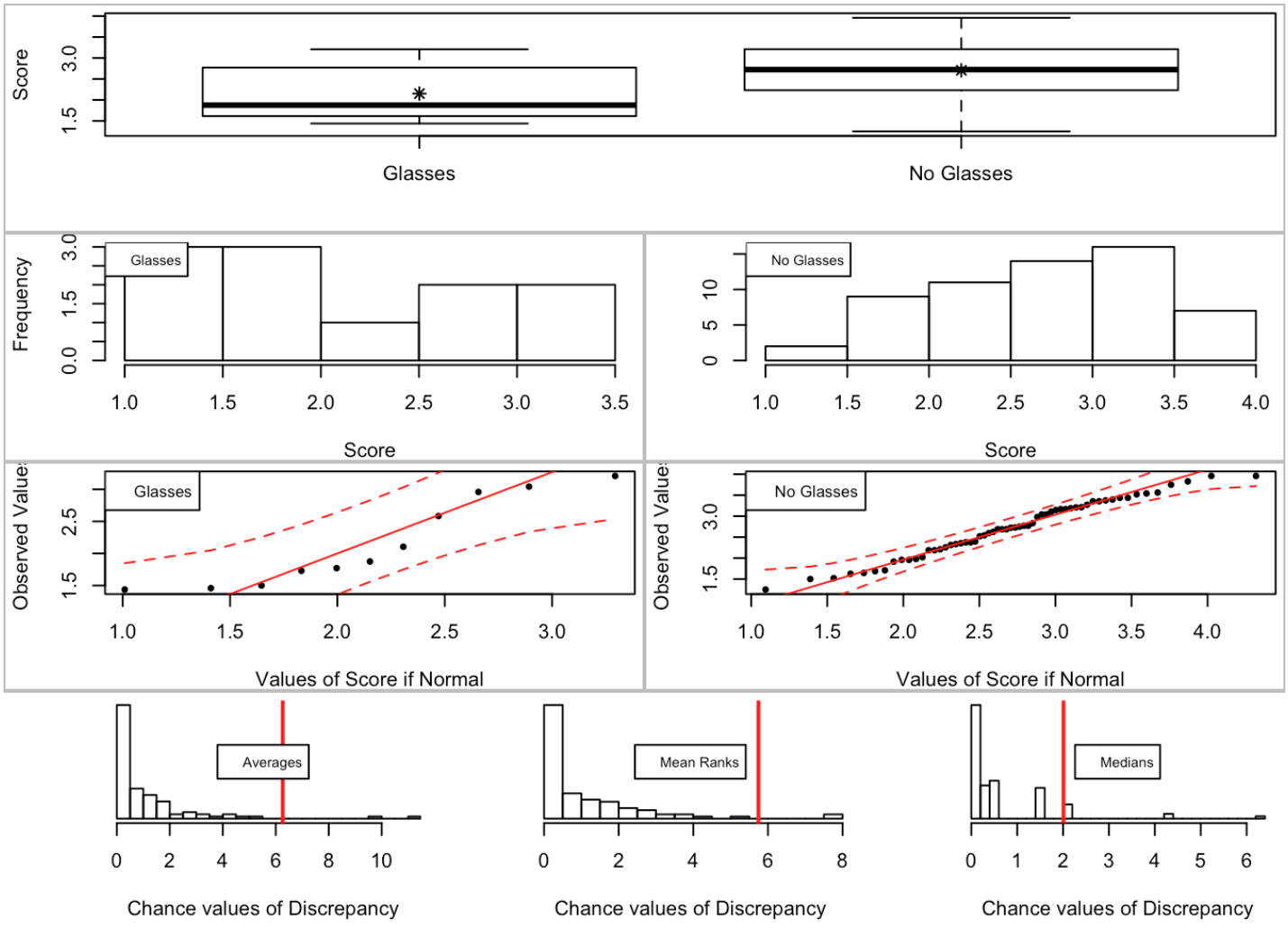
Question 2:

Run `data(ATTRACTF)` to load up the `ATTRACTF` dataset built into `regclass`. Run `?ATTRACTF` to learn about the data. Briefly, it contains the attractiveness ratings of the same 70 people you guys rated regarding friendship potential, and some characteristics of the individuals.

```
#Load in ATTRACTF
data(ATTRACTF)
```

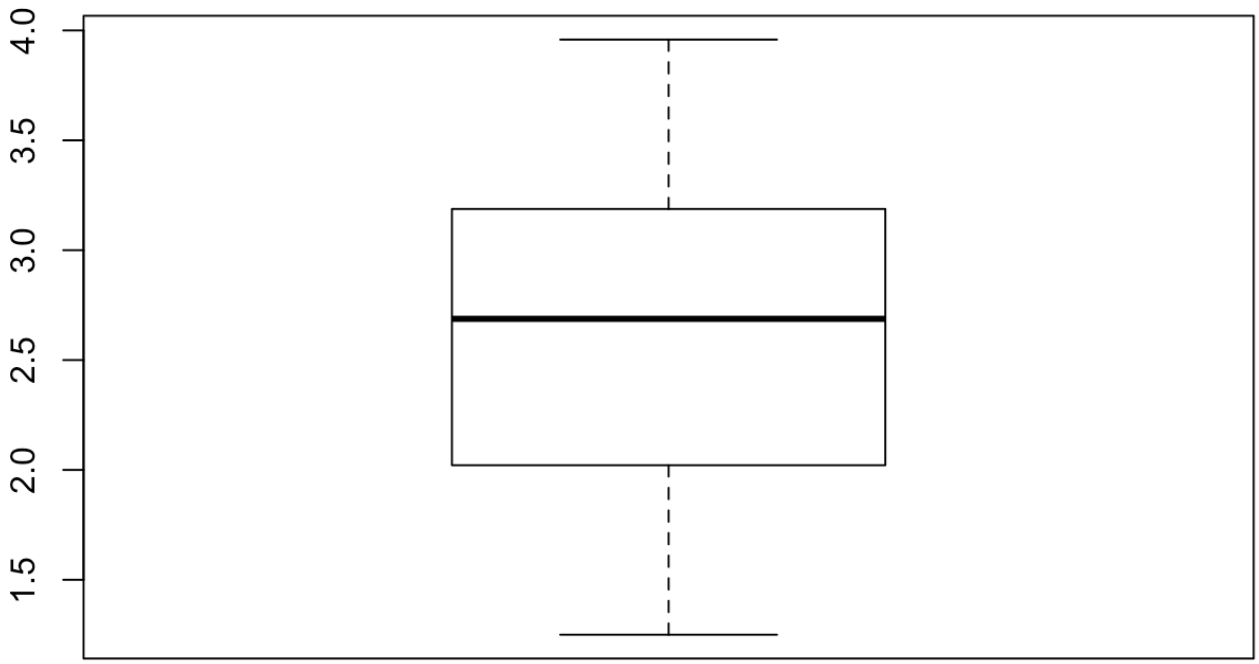
a. Setup and run the `associate` command. Choose the roles of y and x so that we can answer the question "Does wearing glasses influence someone's attractiveness?". Create 100 permutation datasets and use `seed=320`.

```
#Your associate command, with 100 permutations and seed of 320
associate(Score~Glasses, data=ATTRACTF, permutations=100, seed=320)
```

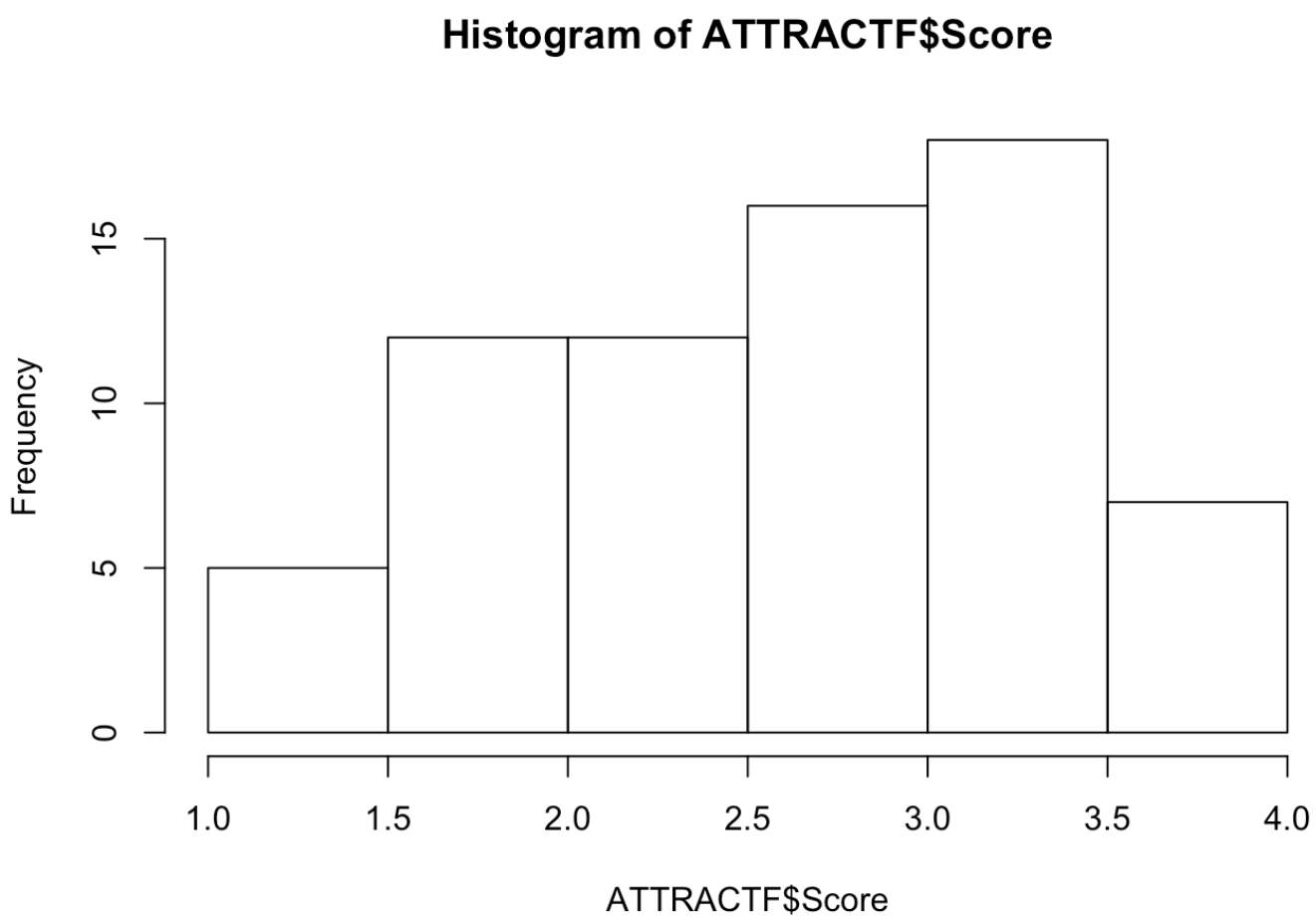


```
## Association between Glasses (categorical) and  Score (numerical)
##  using 70 complete cases
##
## Sample Sizesx
##    Glasses No Glasses
##         11         59
##
## Permutation procedure:
##                      Glasses No Glasses Discrepancy Estimated p-value
## Averages (ANOVA)       2.152      2.706       6.263              0.02
## Mean Ranks (Kruskal)      25      37.46       5.745              0.02
## Medians                1.875      2.723       2.009              0.09
## With 100 permutations, we are 95% confident that
##  the p-value of ANOVA (means) is between 0.002 and 0.07
##  the p-value of Kruskal-Wallis (ranks) is between 0.002 and 0.07
##  the p-value of median test is between 0.042 and 0.164
## Note:  If 0.05 is in a range, change permutations= to a larger number
##
##
##
## Advice: If it makes sense to compare means (i.e., no extreme outliers and the
## distributions aren't too skewed), use the ANOVA.  If there there are
## some obvious extreme outliers but the distributions are roughly symmetric, use
## Rank test.  Otherwise, use the Median test or rerun the test using, e.g., log1
0(y)
## instead of y
```

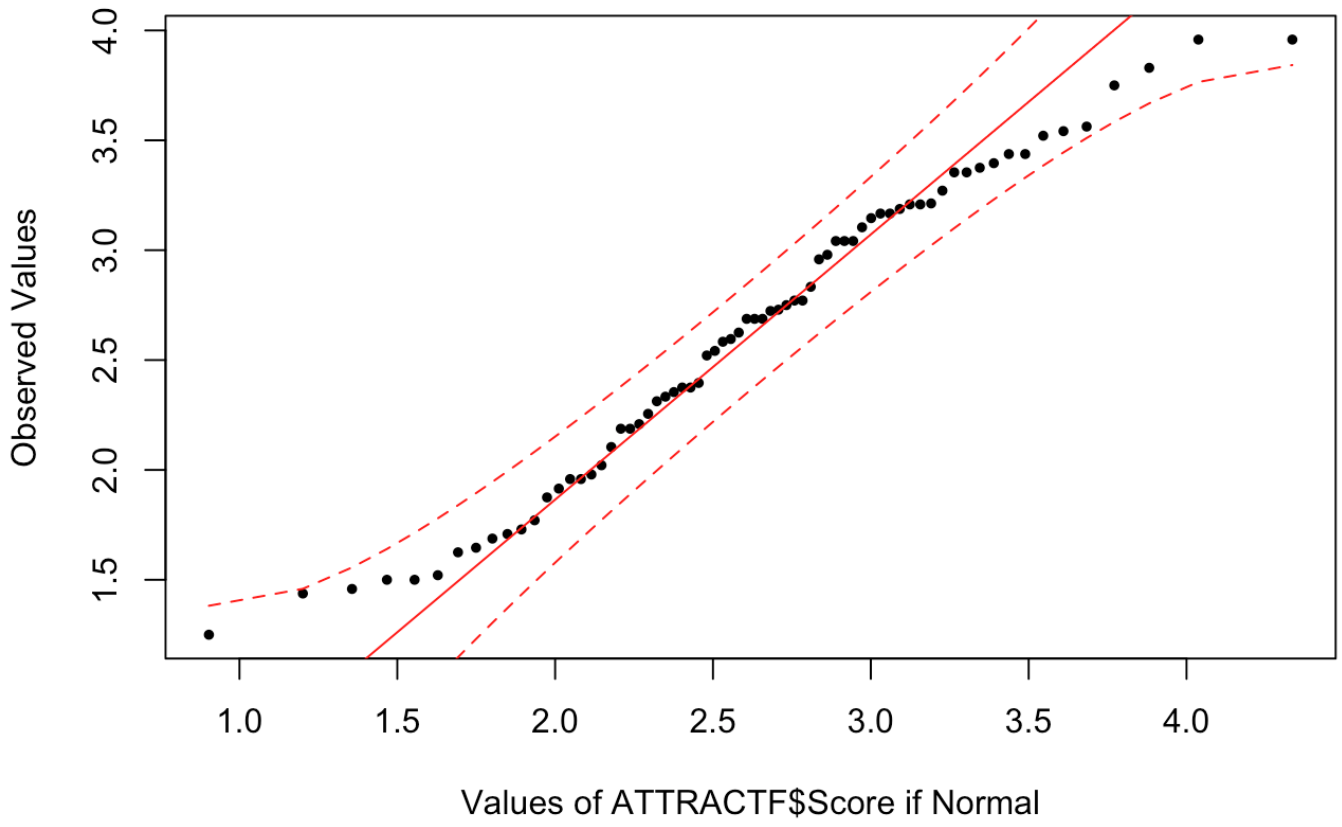
```
boxplot(ATTRACTF$Score)
```




```
hist(ATTRACTF$Score)
```




```
qq(ATTRACTF$Score)
```



b. Examine the boxplots, histograms, and QQ plots. What number provides the most meaningful summary of the typical value of `Score` here: the mean or the median? Justify your answer by referencing the appropriate plot.

Response: Mean - All points fall in boundaries of qq plot and histogram looks fairly normal.

c. Quote this “typical”" value of attractiveness score for girls with glasses and girls without glasses. Who is generally considered more attractive: girls with or girls without glasses? Is this difference “large” and of practical significance? Note: practical significance is in the eye of the beholder, so justify your thoughts.

Response: 2 for girls with, and 3 for girls without. Without. There are not very many observations for girls with glasses as compared to girls without, and given the unreliable nature of personal preference ratings, we cannot practically say that those without glasses are more attractive than those with glasses.

d. The test of statistical significance is inconclusive. Why? Make the appropriate change to `associate` so that the test gives a more definitive answer as to whether the association is or is not statistically significant and state your finding. Note: add the argument `plot=FALSE` to `associate` so that the plots are not re-displayed.

```
#Your updated associate command. Note:  add the argument plot=FALSE so that plots aren't included.
associate(Score~Glasses, data=ATTRACTF, permutations=500, seed=320, plot=FALSE)
```

```
## Association between Glasses (categorical) and  Score (numerical)
##  using 70 complete cases
##
## Sample Sizesx
##    Glasses No Glasses
##         11         59
##
## Permutation procedure:
##                      Glasses No Glasses Discrepancy Estimated p-value
## Averages (ANOVA)       2.152      2.706       6.263             0.024
## Mean Ranks (Kruskal)      25      37.46       5.745             0.026
## Medians                1.875      2.723       2.009              0.12
## With 500 permutations, we are 95% confident that
##  the p-value of ANOVA (means) is between 0.012 and 0.042
##  the p-value of Kruskal-Wallis (ranks) is between 0.014 and 0.044
##  the p-value of median test is between 0.093 and 0.152
## Note:  If 0.05 is in a range, change permutations= to a larger number
##
##
##
## Advice: If it makes sense to compare means (i.e., no extreme outliers and the
## distributions aren't too skewed), use the ANOVA.  If there there are
## some obvious extreme outliers but the distributions are roughly symmetric, use
## Rank test.  Otherwise, use the Median test or rerun the test using, e.g., log1
0(y)
## instead of y
```

```
# Since the p-value falls between 1.2 and 4.2, it looks like the association here is statistically significant. To reiterate, the sample size for girls with glasses is quite small.
```

---

e. Determine whether any of the associations of `Score` with the 9 other categorical variables in `ATTRACTF` are statistically significant by running the `associate` command on each. However, let's me more stringent and consider an association to be statistically significant only if the p-value is less than 0.01. Since we are performing so many tests, we run the risk when using the standard procedure of "discovering" an association that doesn't really exist by getting a p-value less than 5% by chance. Changing the threshold to 0.01 helps control the "false discovery" rate with so may tests.

```
#Your associate commands. Note: I have added eval=FALSE to this code so that it wi
ll not
#include the output in the final document to save space.
associate(Score~Actual.Sexuality, data=ATTRACTF, permutations=500, seed=320, plo
t=FALSE) #NO
associate(Score~ApparentRace, data=ATTRACTF, permutations=500, seed=320, plot=FALS
E) #NO
associate(Score~Chin, data=ATTRACTF, permutations=500, seed=320, plot=FALSE) #no
associate(Score~Cleavage, data=ATTRACTF, permutations=500, seed=320, plot=FALSE)
#NO
associate(Score~ClothingStyle, data=ATTRACTF, permutations=500, seed=320, plot=FAL
SE) #NO
associate(Score~HairColor, data=ATTRACTF, permutations=500, seed=320, plot=FALSE)
#no
associate(Score~Smile, data=ATTRACTF, permutations=500, seed=320, plot=FALSE) #NO
associate(Score~Selfie, data=ATTRACTF, permutations=500, seed=320, plot=FALSE) #NO
associate(Score~LookingAtCamera, data=ATTRACTF, permutations=500, seed=320, plot=T
RUE) #NO
```

Response: None of the other categorical variables have a statistically significant association with the final attractiveness score.