Activity 8 - Aryan Khandelwal

…

The main goal of the activity is to make sure package regclass is installed so that you are set up for the rest of the course. This package contains the commands written specifically for association and regression analysis and contains many datasets.

Once regclass is installed, load it up and make it available by running

#install.packages("regclass")  
library(regclass)  
#If not installed, run the line below. (Of course, remove # sign)

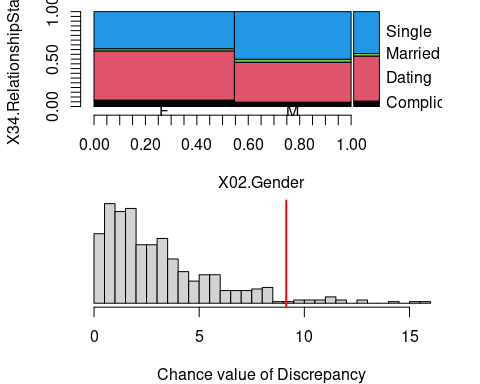
**Question 1:** After the regclass library has been loaded, load in the SURVEY11 dataset from package regclass by running

data(SURVEY11)

This contains information about students from 2011. Let us study a potential association between someone’s relationship status X34.RelationshipStatus (the variable) and someone’s gender X02.Gender. If an association existed, this would imply that men and women view definitions of relationship statuses somewhat differently.

Use associate, adding the argument seed=298 so that the set of 500 (default value) permutation datasets generated are the same for everyone.

#Your associate command using a seed of 298  
associate(X34.RelationshipStatus~X02.Gender, data = SURVEY11, seed = 298)



## Association between X02.Gender (categorical) and X34.RelationshipStatus (categorical):  
##   
## using 628 complete cases  
## Contingency table:  
## y  
## x Complicated Dating Married Single Total  
## F 23 178 8 134 343  
## M 13 120 9 143 285  
## Total 36 298 17 277 628  
##   
## Table of Expected Counts:  
## Complicated Dating Married Single  
## F 19.7 162.8 9.3 151.3  
## M 16.3 135.2 7.7 125.7  
##   
## Conditional distributions of y (X34.RelationshipStatus) for each level of x (X02.Gender):  
## If there is no association, these should look similar to each other and  
## similar to the marginal distribution of y  
## Complicated Dating Married Single  
## F 0.06705539 0.5189504 0.02332362 0.3906706  
## M 0.04561404 0.4210526 0.03157895 0.5017544  
## Marginal 0.05732484 0.4745223 0.02707006 0.4410828  
##   
## Permutation procedure:  
## Discrepancy Estimated p-value  
## 9.138875 0.034  
## With 500 permutations, we are 95% confident that:  
## the p-value is between 0.02 and 0.054   
## If 0.05 is in this range, change permutations= to a larger number

1-pchisq(9.138875, df = (4-1)\*(2-1))

## [1] 0.02750033

* Does the mosaic plot suggest that an association exists? Why or why not?

*Response:* Yes, the mosaic plot suggests an association because there are noticeable differences between the male and female distributions. There are more females that are dating then men.

* Let’s estimate -value via theoretical approach for the “discrepancy”” between the conditional distributions of relationship status between genders?

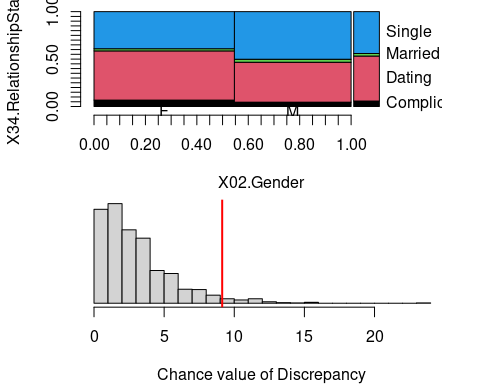
*Response:* The p value via theoretical approach is 0.02750033.

* Let’s estimate -value via permutation approach for the “discrepancy”” between the conditional distributions of relationship status between genders? Why is the test inconclusive?

*Response:* The p value via permutation is 0.034. This test is inconclusive because 0.05 is inside the range of p values.

* Add the argument permutations=1500 to make 1500 permutation datasets instead of the default 500 (have the seed still be 298) with seed number equals to 298. The test will now be conclusive. Is the association statistically significant?

#Your associate command using 1500 permutations and a seed of 298  
associate(X34.RelationshipStatus~X02.Gender, data = SURVEY11, seed = 298, permutations = 1500)



## Association between X02.Gender (categorical) and X34.RelationshipStatus (categorical):  
##   
## using 628 complete cases  
## Contingency table:  
## y  
## x Complicated Dating Married Single Total  
## F 23 178 8 134 343  
## M 13 120 9 143 285  
## Total 36 298 17 277 628  
##   
## Table of Expected Counts:  
## Complicated Dating Married Single  
## F 19.7 162.8 9.3 151.3  
## M 16.3 135.2 7.7 125.7  
##   
## Conditional distributions of y (X34.RelationshipStatus) for each level of x (X02.Gender):  
## If there is no association, these should look similar to each other and  
## similar to the marginal distribution of y  
## Complicated Dating Married Single  
## F 0.06705539 0.5189504 0.02332362 0.3906706  
## M 0.04561404 0.4210526 0.03157895 0.5017544  
## Marginal 0.05732484 0.4745223 0.02707006 0.4410828  
##   
## Permutation procedure:  
## Discrepancy Estimated p-value  
## 9.138875 0.034  
## With 1500 permutations, we are 95% confident that:  
## the p-value is between 0.025 and 0.044   
## If 0.05 is in this range, change permutations= to a larger number

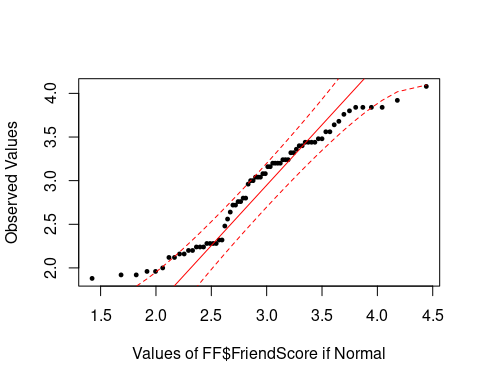
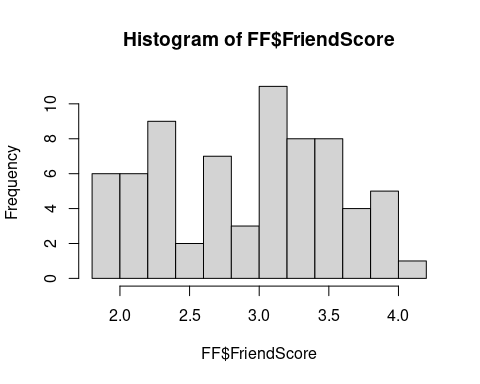
*Comment:* The association is statistically significant as the p-value is below 5%.

**Question 2:**  
Download Act8-friendpotF.dat to your STA 3000 folder and read it into R using read.csv, calling it FF. This is the dataset of friendship potentials generated by students.

FF <- read.csv("Act8-friendpotF.dat") #Uncomment line and run when file is in the right place.

* Look at a histogram and a QQ-plot of FriendScore using hist() and qq(). Remember to refer to a specific column of the FF dataframe you will need to write it as FF$FriendScore. Does the mean or does the median provide the better summary of the typical value of FriendScore?

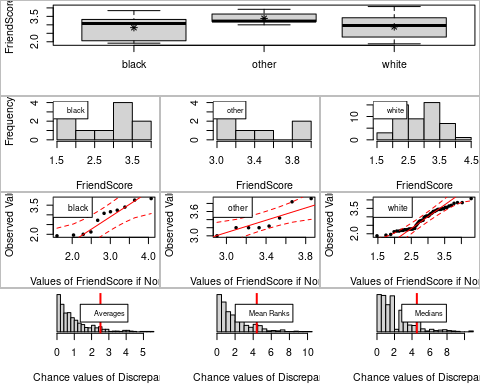
#R code for histogram and qqplot  
hist(FF$FriendScore) ; qq(FF$FriendScore)



*Comment:* Mean provides a better summary of typical value in the data, because the distribution can be described as a normal distribution.

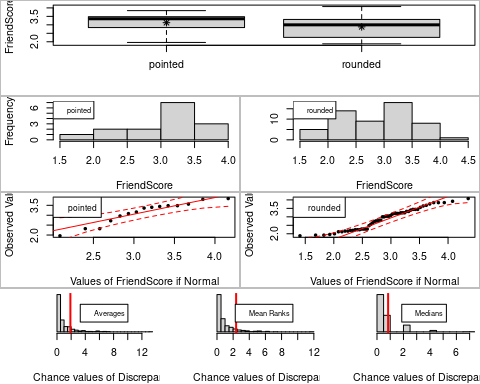
* Go through some of the other categorical variables and determine if associations exist with the seed number equals to 2015 (the test with the permutation procedure). If an association exists, which level has the highest average friendship potential? Possible choices: ApparentRace, Chin, Cleavage, ClothingStyle, Glasses, HairColor, LookingAtCamera, Selfie, Piercings. If a test is inconclusive with the default 500 permutations (e.g., Glasses), up the number of permutations.

associate(FriendScore~ApparentRace,data=FF,seed=2015) #No associations



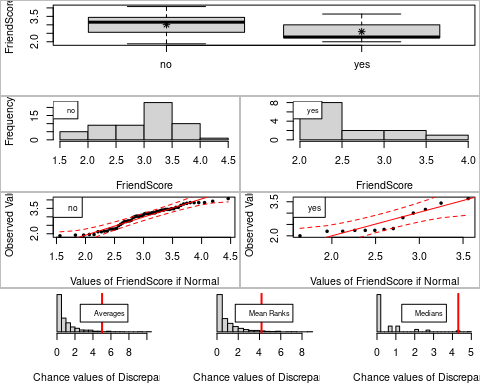
## Association between ApparentRace (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## black other white   
## 11 8 51   
##   
## Permutation procedure:  
## black other white Discrepancy Estimated p-value  
## Averages (ANOVA) 2.836 3.38 2.883 2.513 0.082  
## Mean Ranks (Kruskal) 33.27 31.25 36.65 4.393 0.114  
## Medians 3.08 3.22 2.96 4.554 0.1  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.059 and 0.11   
## the p-value of Kruskal-Wallis (ranks) is between 0.087 and 0.145   
## the p-value of median test is between 0.075 and 0.13   
## 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., log10(y)   
## instead of y

associate(FriendScore~Chin,data=FF,seed=2015) # No associations



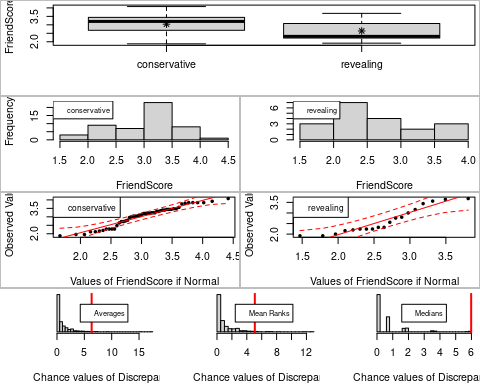
## Association between Chin (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## pointed rounded   
## 15 55   
##   
## Permutation procedure:  
## pointed rounded Discrepancy Estimated p-value  
## Averages (ANOVA) 3.125 2.88 1.893 0.15  
## Mean Ranks (Kruskal) 31.4 36.62 2.37 0.098  
## Medians 3.36 3 0.8408 0.288  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.12 and 0.184   
## the p-value of Kruskal-Wallis (ranks) is between 0.073 and 0.127   
## the p-value of median test is between 0.249 and 0.33   
## 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., log10(y)   
## instead of y

associate(FriendScore~Cleavage,data=FF,seed=2015) # Associations



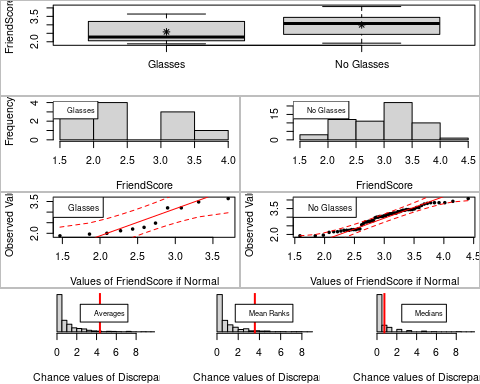
## Association between Cleavage (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## no yes   
## 57 13   
##   
## Permutation procedure:  
## no yes Discrepancy Estimated p-value  
## Averages (ANOVA) 3.009 2.597 5.014 0.02  
## Mean Ranks (Kruskal) 34.44 40.15 4.193 0.032  
## Medians 3.16 2.28 4.309 0.022  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.01 and 0.036   
## the p-value of Kruskal-Wallis (ranks) is between 0.018 and 0.051   
## the p-value of median test is between 0.011 and 0.039   
## 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., log10(y)   
## instead of y

associate(FriendScore~ClothingStyle,data=FF,seed=2015) # Associations



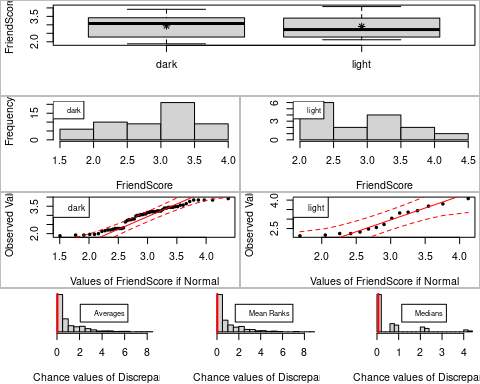
## Association between ClothingStyle (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## conservative revealing   
## 51 19   
##   
## Permutation procedure:  
## conservative revealing Discrepancy Estimated p-value  
## Averages (ANOVA) 3.042 2.64 6.335 0.014  
## Mean Ranks (Kruskal) 35.73 34.89 5.077 0.02  
## Medians 3.2 2.32 5.983 0.008  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.006 and 0.029   
## the p-value of Kruskal-Wallis (ranks) is between 0.01 and 0.036   
## the p-value of median test is between 0.002 and 0.02   
## 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., log10(y)   
## instead of y

associate(FriendScore~Glasses,data=FF,seed=2015, permutations = 1000) #Associations



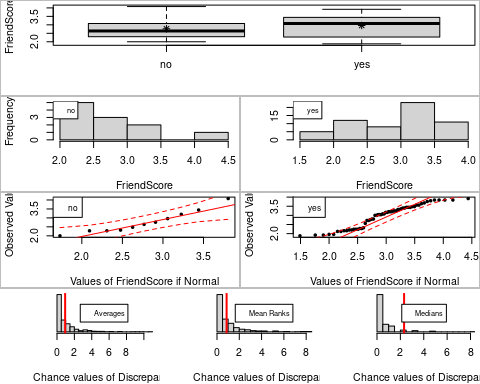
## Association between Glasses (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## Glasses No Glasses   
## 11 59   
##   
## Permutation procedure:  
## Glasses No Glasses Discrepancy Estimated p-value  
## Averages (ANOVA) 2.585 2.997 4.343 0.031  
## Mean Ranks (Kruskal) 35.91 35.42 3.569 0.051  
## Medians 2.28 3.08 0.7477 0.299  
## With 1000 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.021 and 0.044   
## the p-value of Kruskal-Wallis (ranks) is between 0.038 and 0.067   
## the p-value of median test is between 0.271 and 0.328   
## 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., log10(y)   
## instead of y

associate(FriendScore~HairColor,data=FF,seed=2015) # No Associations



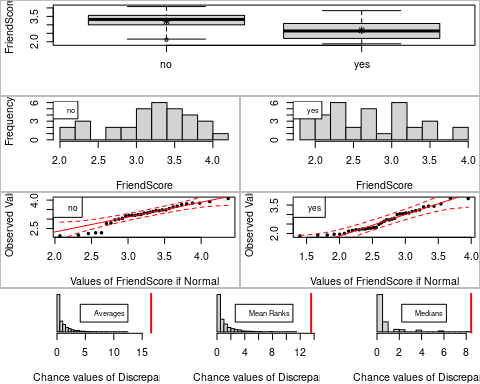
## Association between HairColor (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## dark light   
## 55 15   
##   
## Permutation procedure:  
## dark light Discrepancy Estimated p-value  
## Averages (ANOVA) 2.94 2.904 0.04048 0.846  
## Mean Ranks (Kruskal) 36.96 30.13 0.03205 0.872  
## Medians 3.08 2.72 0.06254 0.774  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.811 and 0.877   
## the p-value of Kruskal-Wallis (ranks) is between 0.84 and 0.9   
## the p-value of median test is between 0.735 and 0.81   
## 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., log10(y)   
## instead of y

associate(FriendScore~LookingAtCamera,data=FF,seed=2015) #No Associations



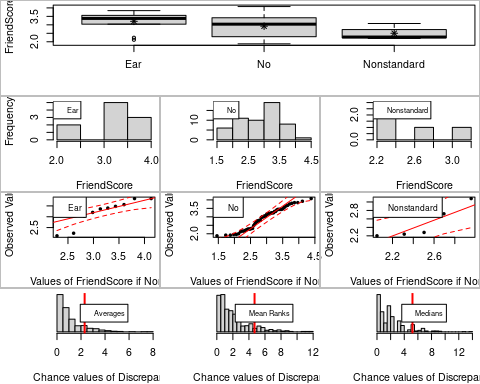
## Association between LookingAtCamera (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## no yes   
## 11 59   
##   
## Permutation procedure:  
## no yes Discrepancy Estimated p-value  
## Averages (ANOVA) 2.767 2.963 0.9385 0.336  
## Mean Ranks (Kruskal) 33.09 35.95 0.862 0.358  
## Medians 2.64 3.08 2.318 0.07  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.295 and 0.379   
## the p-value of Kruskal-Wallis (ranks) is between 0.316 and 0.402   
## the p-value of median test is between 0.049 and 0.096   
## 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., log10(y)   
## instead of y

associate(FriendScore~Selfie,data=FF,seed=2015, permutations = 1000) # Associations



## Association between Selfie (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## no yes   
## 33 37   
##   
## Permutation procedure:  
## no yes Discrepancy Estimated p-value  
## Averages (ANOVA) 3.219 2.677 16.59 0  
## Mean Ranks (Kruskal) 33.24 37.51 13.58 0  
## Medians 3.32 2.64 8.441 0.002  
## With 1000 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0 and 0.004   
## the p-value of Kruskal-Wallis (ranks) is between 0 and 0.004   
## the p-value of median test is between 0 and 0.007   
## 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., log10(y)   
## instead of y

associate(FriendScore~Piercings,data=FF,seed=2015) # No Associations



## Association between Piercings (categorical) and FriendScore (numerical)  
## using 70 complete cases  
##   
## Sample Sizesx  
## Ear No Nonstandard   
## 10 55 5   
##   
## Permutation procedure:  
## Ear No Nonstandard Discrepancy Estimated p-value  
## Averages (ANOVA) 3.208 2.921 2.504 2.301 0.132  
## Mean Ranks (Kruskal) 33.9 35.75 36 4.702 0.118  
## Medians 3.38 3.04 2.28 5.207 0.104  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.104 and 0.165   
## the p-value of Kruskal-Wallis (ranks) is between 0.091 and 0.15   
## the p-value of median test is between 0.079 and 0.134   
## 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., log10(y)   
## instead of y

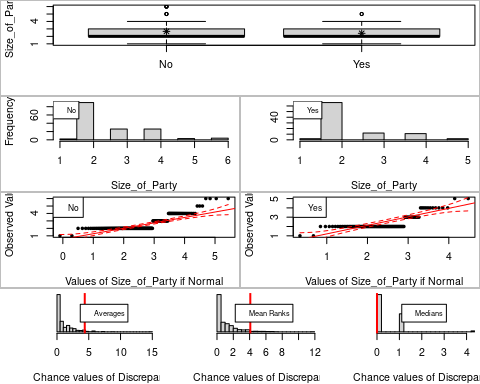
*Comment:*Cleavage, Clothing, Glasses, Selfie all show strong Assosiations as their p values are less than 0.05. The highest average potential is Selfie.

**Question 3:**  
EX2.TIPS dataset records the bill and tip amounts (along with tip percentage) as well as the party size, who was at the table, whether they smoked, and the time of visit.

data(EX2.TIPS)

* We are curious whether there is an association between the size of the party and whether they smoke. Party sizes range between 1 and 6, so in this case, it is appropriate to compare the average between groups (when y only takes on a few different values, using the average is typically the best way to go regardless of what the distribution looks like). Estimate the -value via theoretical approach.

associate(Size\_of\_Party~Smoker, data=EX2.TIPS, seed = 3000)



## Association between Smoker (categorical) and Size\_of\_Party (numerical)  
## using 244 complete cases  
##   
## Sample Sizesx  
## No Yes   
## 151 93   
##   
## Permutation procedure:  
## No Yes Discrepancy Estimated p-value  
## Averages (ANOVA) 2.669 2.409 4.37 0.044  
## Mean Ranks (Kruskal) 123 121.8 4.085 0.038  
## Medians 2 2 8.162e-31 1  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.028 and 0.066   
## the p-value of Kruskal-Wallis (ranks) is between 0.023 and 0.059   
## the p-value of median test is between 0.993 and 1   
## 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., log10(y)   
## instead of y

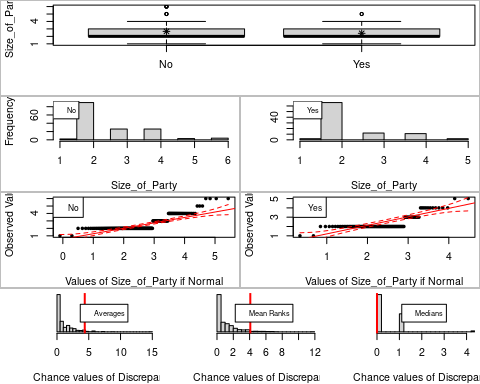
1-pf(4.37, df1 = (2-1), df2 = (244-2))

## [1] 0.03762077

*Comment:* The Estimate p value via theoretical approach is 0.03762077

* We are curious whether there is an association between the size of the party and whether they smoke with the seed number equals to 3000. Run associate() to perform a test of significance using the default number of permutations. Why is the test inconclusive?

associate(Size\_of\_Party~Smoker, data=EX2.TIPS, seed = 3000)

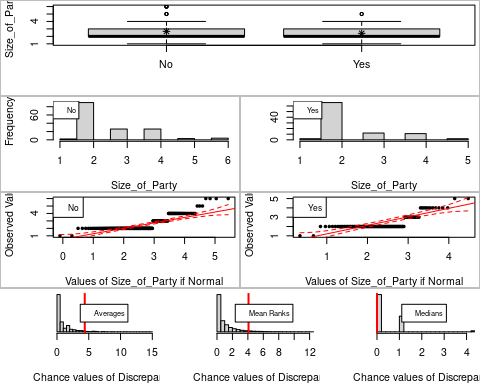


## Association between Smoker (categorical) and Size\_of\_Party (numerical)  
## using 244 complete cases  
##   
## Sample Sizesx  
## No Yes   
## 151 93   
##   
## Permutation procedure:  
## No Yes Discrepancy Estimated p-value  
## Averages (ANOVA) 2.669 2.409 4.37 0.044  
## Mean Ranks (Kruskal) 123 121.8 4.085 0.038  
## Medians 2 2 8.162e-31 1  
## With 500 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.028 and 0.066   
## the p-value of Kruskal-Wallis (ranks) is between 0.023 and 0.059   
## the p-value of median test is between 0.993 and 1   
## 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., log10(y)   
## instead of y

*Comment:* This test is inconclusive because 0.05 is inside the range of quoted p values.

* Change the number of permutations to 5000 with the seed number equals to 3000. The test should now be conclusive. Is there a statistically significant difference in average party size between smokers/non-smokers?

associate(Size\_of\_Party~Smoker, data=EX2.TIPS, seed = 3000, permutations = 5000)

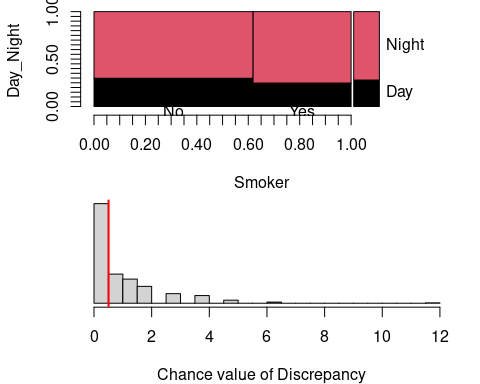


## Association between Smoker (categorical) and Size\_of\_Party (numerical)  
## using 244 complete cases  
##   
## Sample Sizesx  
## No Yes   
## 151 93   
##   
## Permutation procedure:  
## No Yes Discrepancy Estimated p-value  
## Averages (ANOVA) 2.669 2.409 4.37 0.0434  
## Mean Ranks (Kruskal) 123 121.8 4.085 0.0442  
## Medians 2 2 8.162e-31 1  
## With 5000 permutations, we are 95% confident that  
## the p-value of ANOVA (means) is between 0.038 and 0.049   
## the p-value of Kruskal-Wallis (ranks) is between 0.039 and 0.05   
## the p-value of median test is between 0.999 and 1   
## 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., log10(y)   
## instead of y

*Comment:* The tet should be statistically significant beacuse the p values are less than 0.05.

* Examine the association between y = Day\_Night (whether the table was seated for lunch vs. dinner) and x = Smoker with the seed number equals to 3000. Examine graphical output and the result of the statistical test. Is the fraction of non-smokers who dine at lunch larger or smaller than for smokers? Is the association statistically significant? Explain.

associate(Day\_Night~Smoker, data = EX2.TIPS, seed = 3000 )



## Association between Smoker (categorical) and Day\_Night (categorical):  
##   
## using 244 complete cases  
## Contingency table:  
## y  
## x Day Night Total  
## No 45 106 151  
## Yes 23 70 93  
## Total 68 176 244  
##   
## Table of Expected Counts:  
## Day Night  
## No 42.1 108.9  
## Yes 25.9 67.1  
##   
## Conditional distributions of y (Day\_Night) for each level of x (Smoker):  
## If there is no association, these should look similar to each other and  
## similar to the marginal distribution of y  
## Day Night  
## No 0.2980132 0.7019868  
## Yes 0.2473118 0.7526882  
## Marginal 0.2786885 0.7213115  
##   
## Permutation procedure:  
## Discrepancy Estimated p-value  
## 0.5053734 0.48  
## With 500 permutations, we are 95% confident that:  
## the p-value is between 0.435 and 0.525   
## If 0.05 is in this range, change permutations= to a larger number

*Comment:* The fraction of non-smoker who dine at lunch is larger than smokers. The test is not statistically significant beacuse the p value is greater than 0.05 or 5%.