Activity 8 - Faraz Younus

...

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

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

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

```
data(SURVEY11)
```

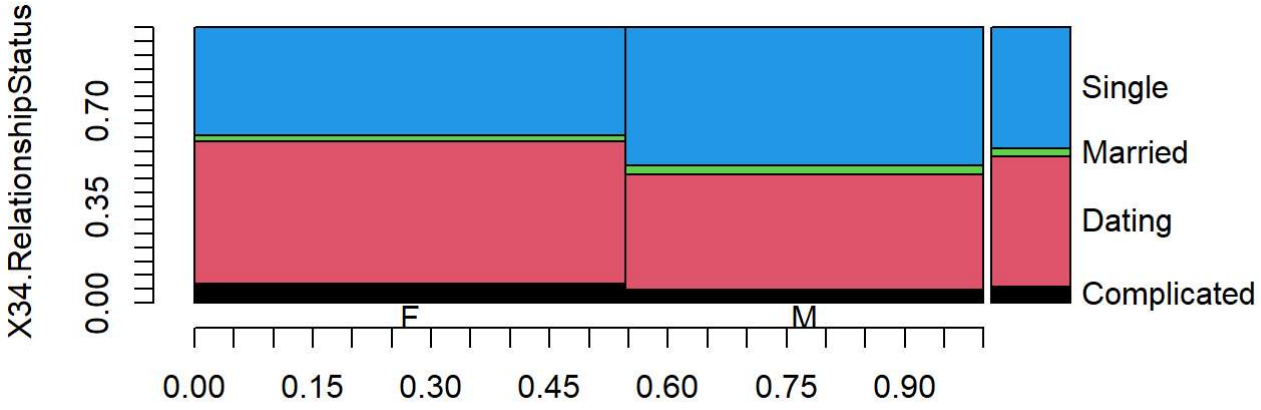
This contains information about students from 2011. Let us study a potential association between someone's relationship status `X34.RelationshipStatus` (the y variable) and someone's gender `X02.Gender`. If an association existed, this would imply that men and women view definitions of relationship statuses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

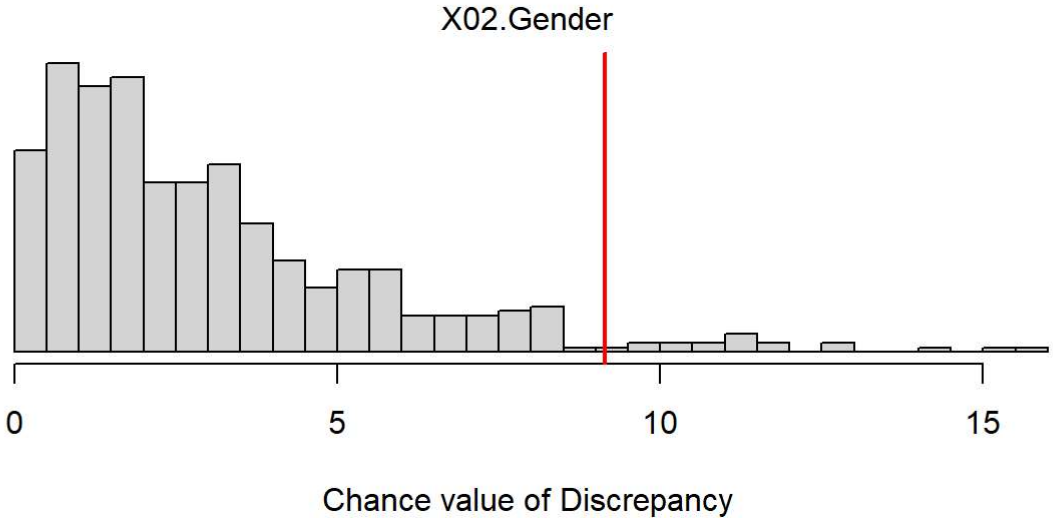
Use `associate`, adding the argument `seed=298` so that the set of 500 (default value) permutation datasets generated are the same for everyone.

```
#Your associate command using a seed of 298

# Set the seed for reproducibility


# Perform the association analysis
associate(X34.RelationshipStatus~ X02.Gender, data=SURVEY11 , seed = 298)
```


X34.RelationshipStatus
0.00
0.35
0.70
Single
Married
Dating
Complicated
F
M
0.00
0.15
0.30
0.45
0.60
0.75
0.90


X02.Gender
0
5
10
15
Chance value of Discrepancy

```
## 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
```

```
nlevels(SURVEY11$X34.RelationshipStatus)
```

```
## [1] 4
```

```
nlevels(SURVEY11$X02.Gender)
```

```
## [1] 2
```

```
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:

- Let's estimate p-value via theoretical approach for the "discrepancy"" between the conditional distributions of relationship status between genders?

Response:

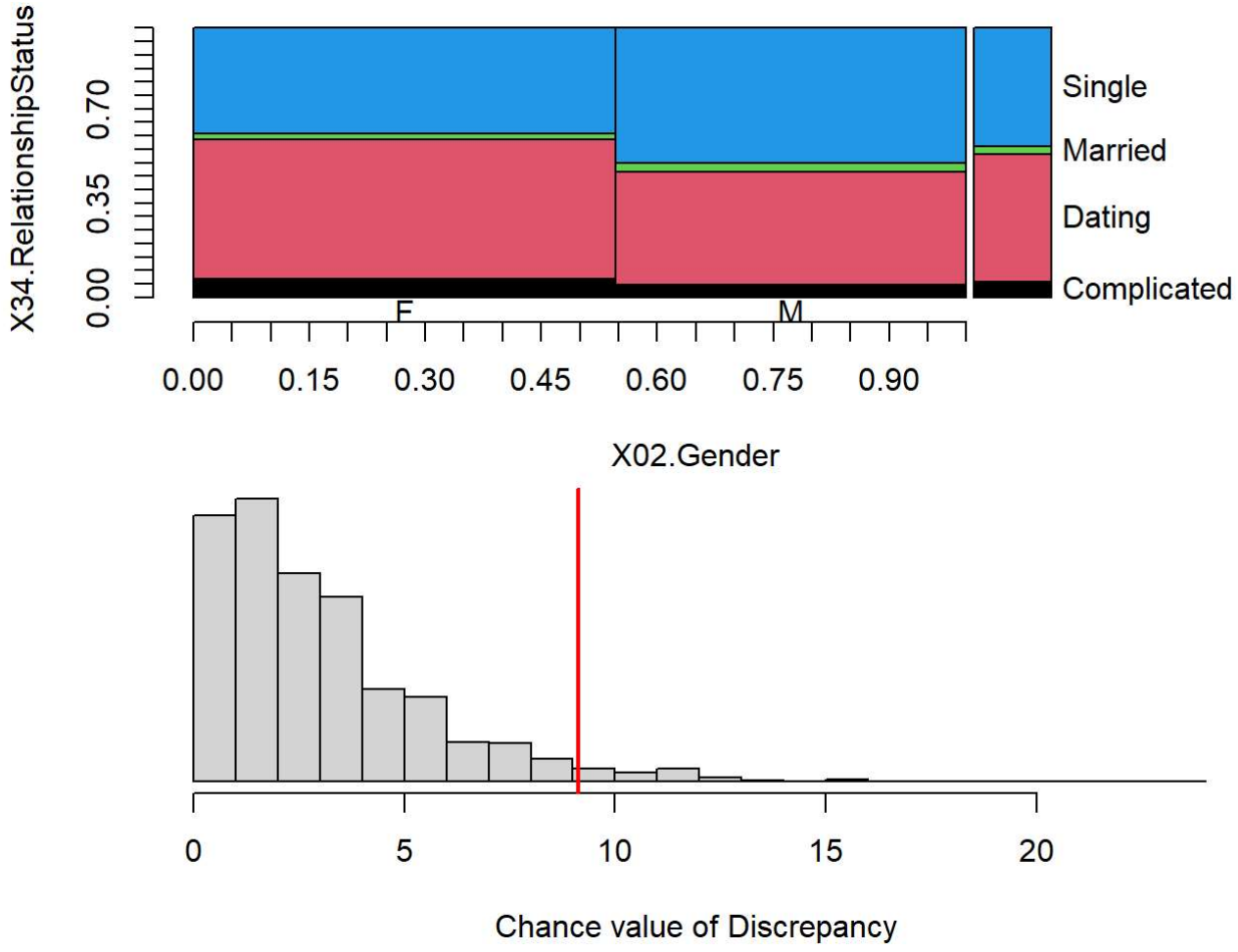
- Let's estimate p-value via permutation approach for the "discrepancy"" between the conditional distributions of relationship status between genders? Why is the test inconclusive?

Response:

- 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:

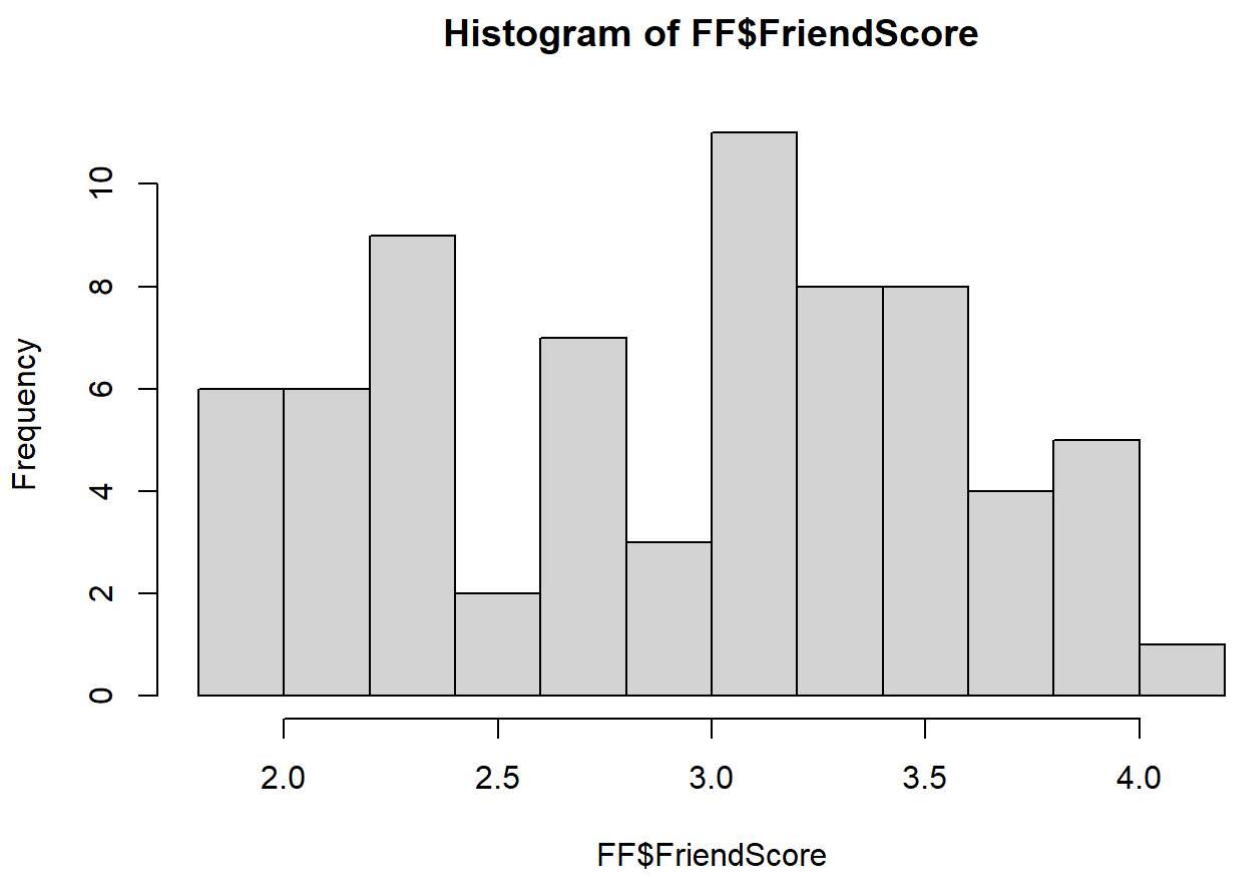
Question 2:

Download Act8-friendpotF.dat to your STA9750 or OPR 9750 folder and read it into R using `read.csv` , calling it `FF` . This is the dataset of friendship potentials generated by students.

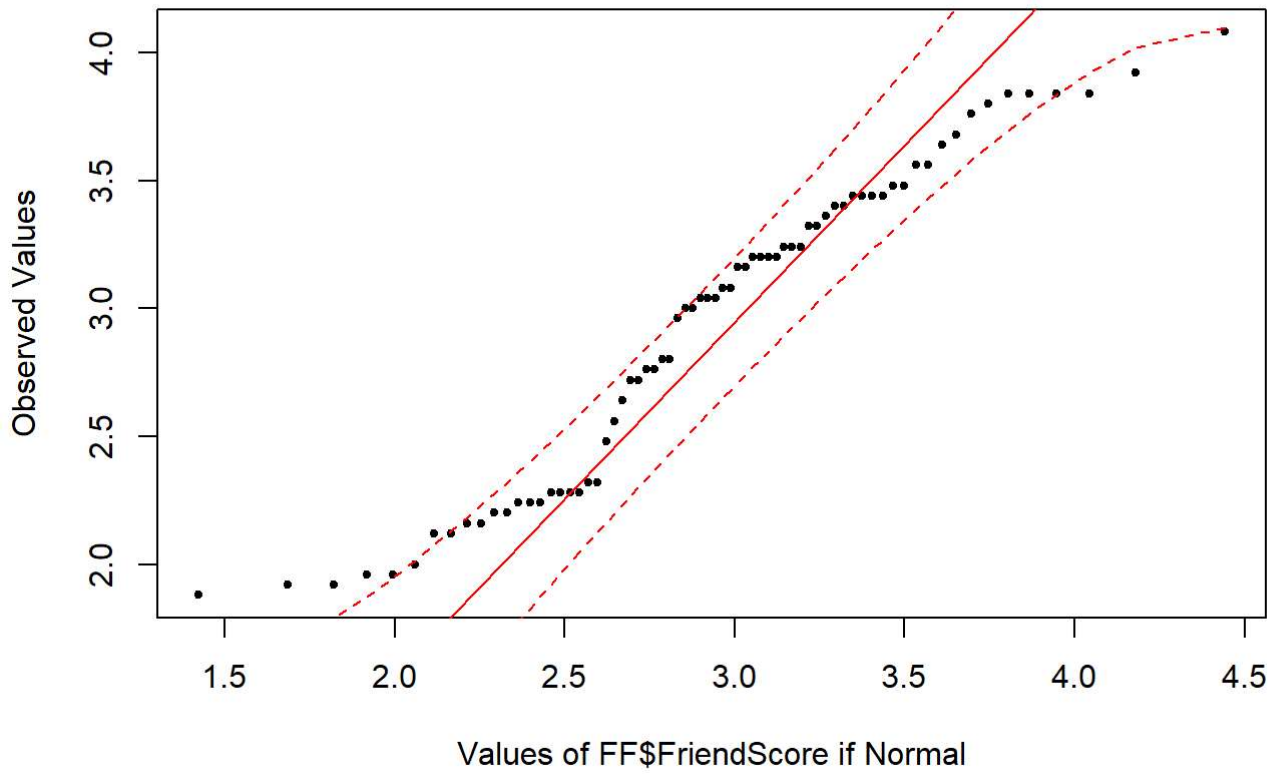
```
FF <- read.csv("Act8-friendpotF.dat")  #Uncomment line and run when file is in the right place
```

- 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)
```



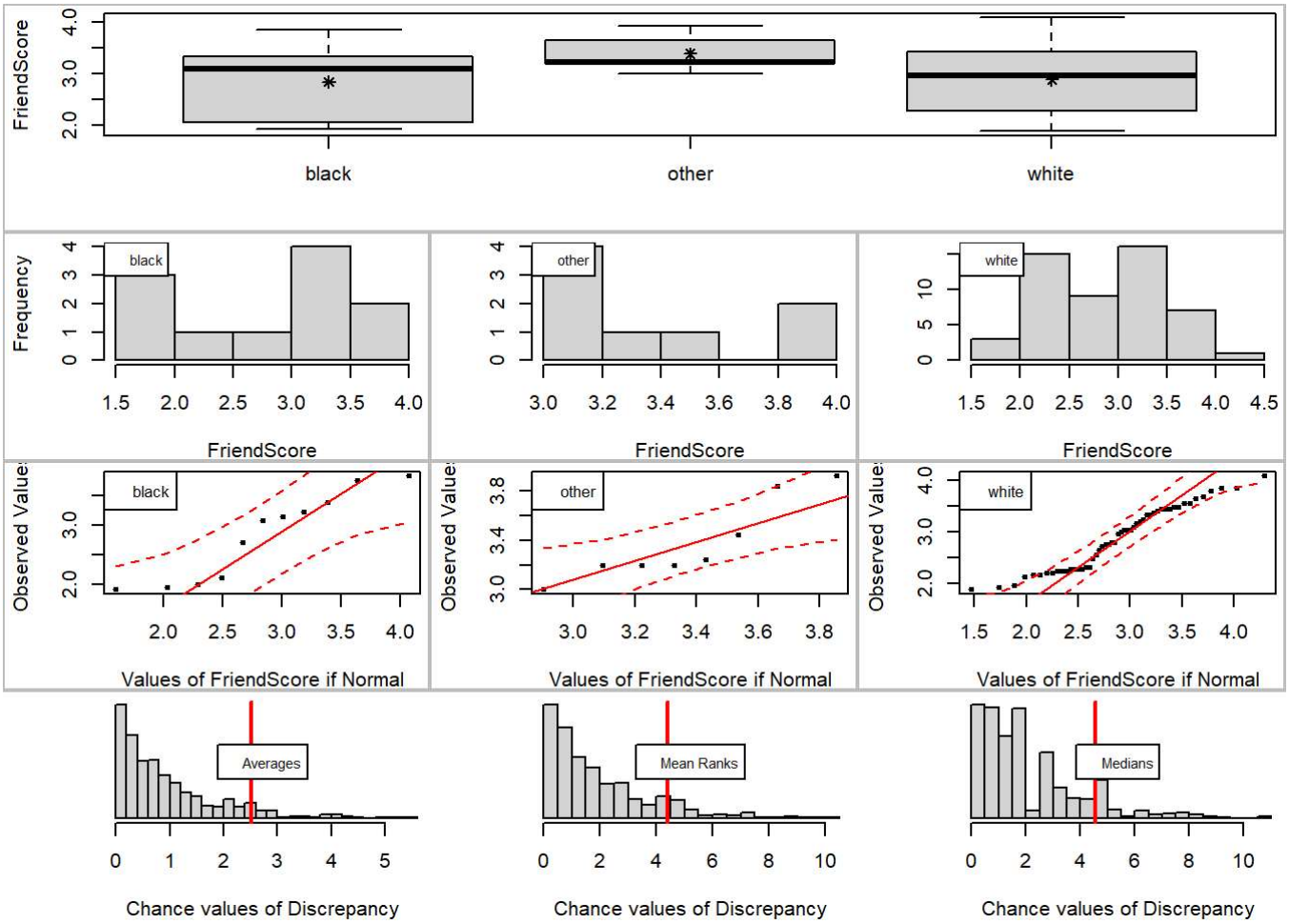

```
qq(FF$FriendScore)
```



Comment:

- 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) #Uncomment and run
```


FriendScore
black
other
white
Frequency
FriendScore
Observed Values
Values of FriendScore if Normal

Averages
Mean Ranks
Medians
Chance values of Discrepancy

```
## 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
```

```
#many more lines using associate, one for each variable, all using seed 2015
```

Comment: ## the QQ plot is good and normally distributed. ## the test result is conculusive. There is is no signicicant relationshop becuase Pvalue>.05 **********

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 p-value via theoretical approach.

```
names(EX2.TIPS)
```

```
## [1] "Tip.Percentage" "Bill_in_USD"    "Tip_in_USD"     "Gender"
## [5] "Smoker"         "Weekday"        "Day_Night"      "Size_of_Party"
```

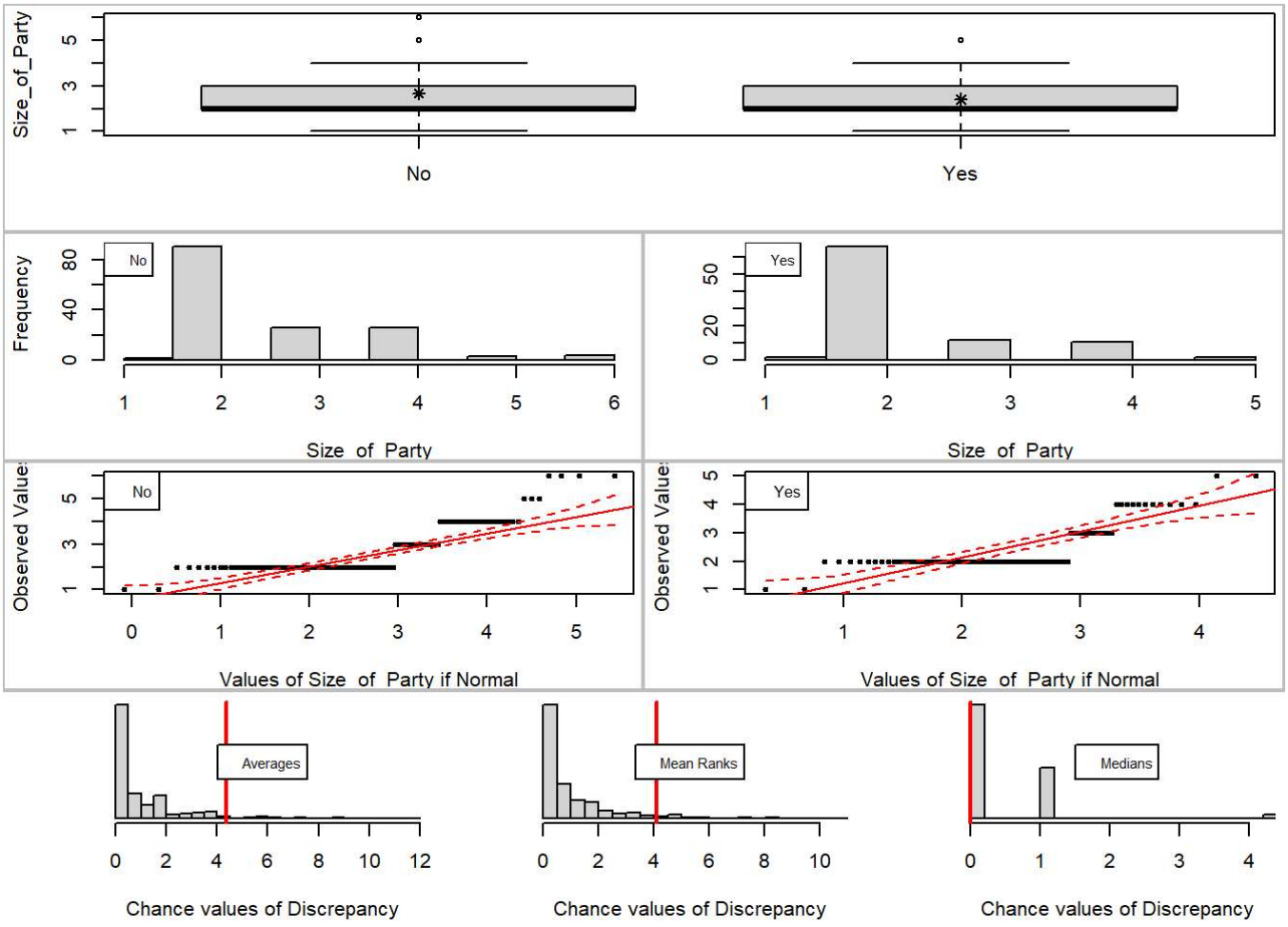
```
1-pf(4.37, df1=(2-1),df2= nrow(EX2.TIPS)-2)
```

```
## [1] 0.03762077
```

Comment:

- We are curious whether there is an association between the size of the party and whether they smoke with the seed number equals to 9750. 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= 9750)
```

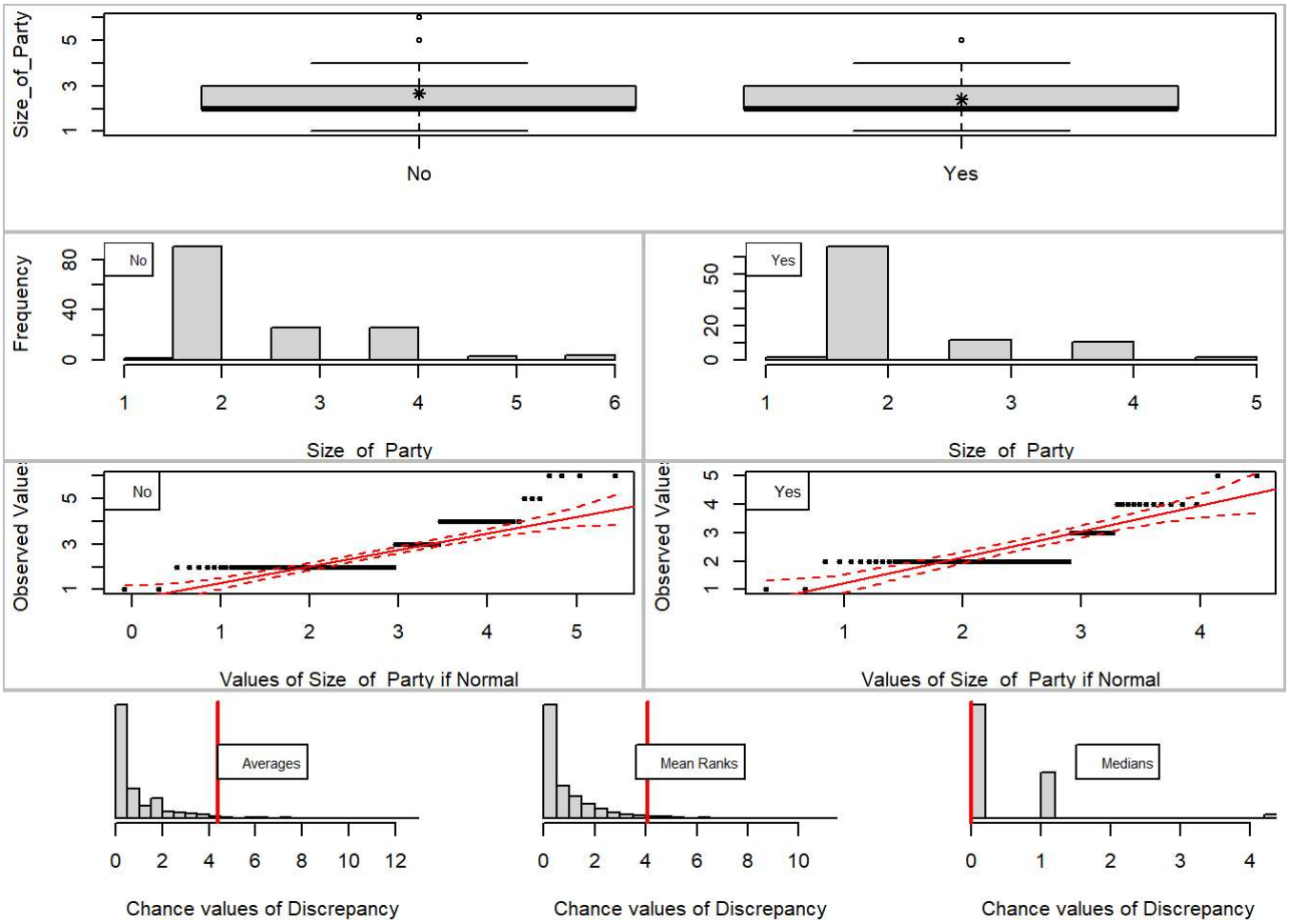


```
## Association between Smoker (categorical) and  Size_of_Party (numerical)
##  using 244 complete cases
##
## Sample Sizesx
##  No Yes
## 151  93
##
## Permutation procedure:
##                          No   Yes Discrepancy Estimated p-value
## Averages (ANOVA)     2.669 2.409        4.37             0.044
## Mean Ranks (Kruskal)   123 121.8       4.085             0.058
## 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 0.039 and 0.082
##  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:

- Change the number of permutations to 5000 with the seed number equals to 9750. 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= 9750 ,permutations= 5000 )
```


Size_of_Party
No
Yes
Frequency
Size of Party
Observed Value
Values of Size of Party if Normal

Averages
Chance values of Discrepancy

Mean Ranks
Chance values of Discrepancy

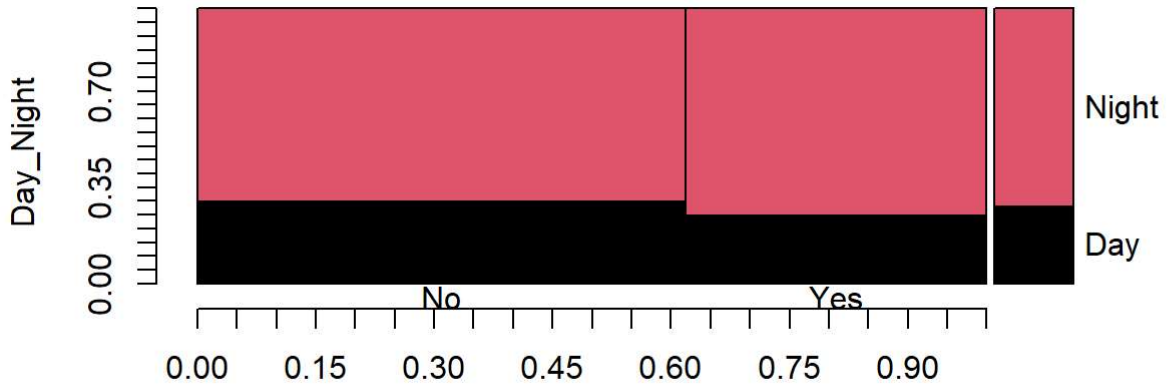
Medians
Chance values of Discrepancy

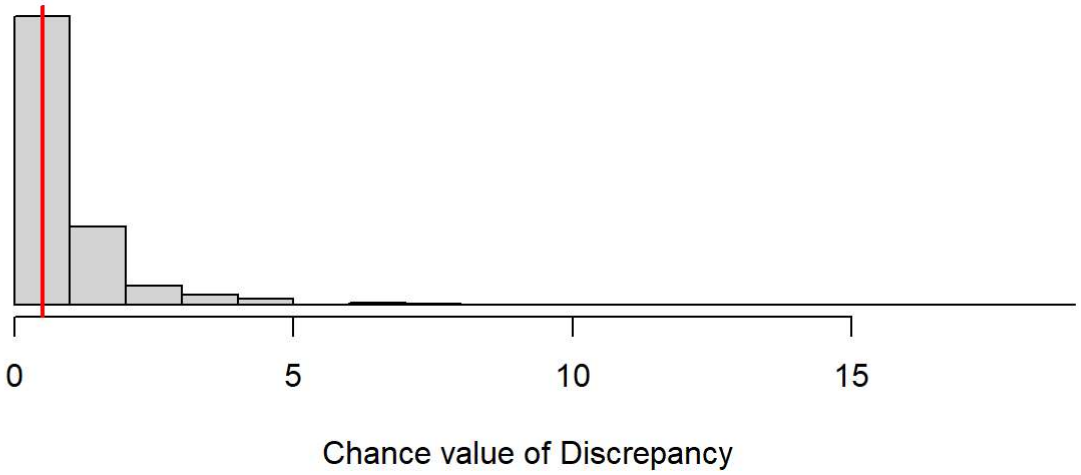
```
## Association between Smoker (categorical) and  Size_of_Party (numerical)
##  using 244 complete cases
##
## Sample Sizesx
##  No Yes
## 151  93
##
## Permutation procedure:
##                          No   Yes Discrepancy Estimated p-value
## Averages (ANOVA)     2.669 2.409        4.37            0.0386
## Mean Ranks (Kruskal)   123 121.8       4.085             0.042
## Medians                  2     2   8.162e-31                 1
## With 5000 permutations, we are 95% confident that
##  the p-value of ANOVA (means) is between 0.033 and 0.044
##  the p-value of Kruskal-Wallis (ranks) is between 0.037 and 0.048
##  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 P value of means is between .033 and .044 thwerefore it's conslusive! **********

- 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 9750. 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= 9750 ,permutations= 5000 )
```


Day_Night
0.00
0.35
0.70
Night
Day
No
Yes
0.00
0.15
0.30
0.45
0.60
0.75
0.90
Smoker


0
5
10
15
Chance value of Discrepancy

```
## 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            0.4538
## With 5000 permutations, we are 95% confident that:
##  the p-value is between 0.44 and 0.468
## If 0.05 is in this range, change permutations= to a larger number
```

Comment: #If P value is less then .05 there is a statistically significiant relationship. **********