p-value of TIPS data

```
associate(TipPercentage~Weekday,data=TIPS)
Association between Weekday (categorical) and  TipPercentage (numerical)
 using 244 complete cases

Sample Sizesx
  Friday Saturday   Sunday Thursday
      19       87       76       62

Permutation procedure:
                     Friday Saturday Sunday Thursday Discrepancy Estimated p-value
Averages (ANOVA)         17    15.32  16.69    16.13      0.8512             0.47
Mean Ranks (Kruskal)  95.79    128.5  138.3    102.8       1.822            0.606
Medians                15.6     15.2  16.15     15.4        1.44              0.7
With 500 permutations, we are 95% confident that
 the p-value of ANOVA (means) is between 0.426 and 0.515
 the p-value of Kruskal-Wallis (ranks) is between 0.562 and 0.649
 the p-value of median test is between 0.658 and 0.74
Note:  If 0.05 is in a range, change permutations= to a larger number
```

Statistical significance

If the *p*-value is less than 5%, then we say the association is statistically significant.

- There is strong (though not conclusive) evidence that at least two levels of x have the same average/median value of y.
- Test does not tell us WHICH levels may be different however.
- Note: a statistically significant difference may not be large or be of any practical interest

If the *p*-value is at least 5%, then the association is not statistically significant.

- The variability in averages/medians is readily explained by chance alone without invoking the presence of an association.
- If there really is an association, it is too weak to be detected with this data.

p-values for examples

- *p*-value of friendship score vs. smile is 0.22, indicating no association
- *p*-value of friendship score vs. actual sexuality is 0.77, indicating no association
- *p*-value of friendship score vs. glasses is 0.046, indicating an association
- *p*-value of friendship score vs. apparent race is 0.10, indicating no association

In the data, very few associations were statistically significant. Whether the woman was prominently featuring her cleavage, wearing glasses, and whether the picture was a selfie seemed to be associated with friendship potential.

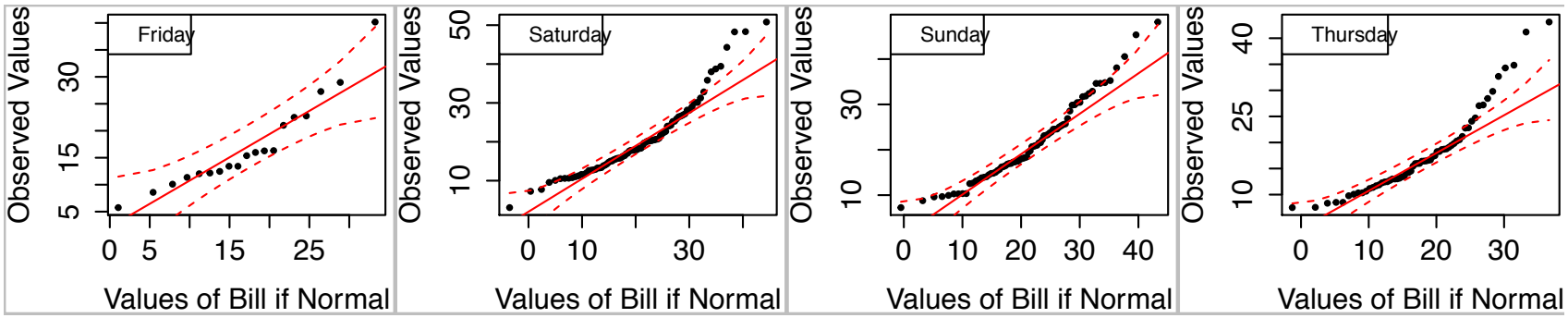
Final example: Bill vs. Weekday

Is there an association between how much parties spend at a restaurant and day of the week?

```
associate(Bill~Weekday,data=TIPS)
```

```
                     Friday Saturday Sunday Thursday Discrepancy Estimated p-value
Averages (ANOVA)      17.15    20.44  21.41    17.68       2.767             0.052
Mean Ranks (Kruskal)  126.3      131  124.1    107.5        10.4             0.022
Medians               15.38    18.24  19.63     16.2       8.566             0.048
With 500 permutations, we are 95% confident that
 the p-value of ANOVA (means) is between 0.034 and 0.075
 the p-value of Kruskal-Wallis (ranks) is between 0.011 and 0.039
 the p-value of median test is between 0.031 and 0.071
Note:  If 0.05 is in a range, change permutations= to a larger number
```

Final example: Bill vs. Weekday



Thursday's distribution has a systematic bend and quite a few points outside the bands, so let's compare medians.

Final example: Bill vs. Weekday

The p-value of the median test is 0.034. This is less than 5%, indicating a statistically significant association (the median for Friday is \$15.38 compared to a median of \$19.63, which is pretty large).

Not so fast! Since the p-value is estimated from the permutation procedure, this test is INCONCLUSIVE. The range of p-values consistent with our simulation is between 0.02 and 0.054, so we'd need to up the number of permutations from the default value of 500. When this is done, the p-value is between 0.032 and 0.043, so the association is indeed significant.

Using R

Loading data built into R

R has many datasets built in which can be loaded in with the command `data`.

- `data(faithful)` loads up information on eruption/waiting times for Old Faithful
- `data(airquality)` loads up information about daily air quality measurements in New York, May to Sept 1973

Once you have installed the `regclass` package, there are many datasets you can load this way.

- `library(regclass)` will load up the library and give you access to the routines/data
- `data(CALLS)` loads up dropped call data
- `data(CHURN)` loads up information on customers and whether they renewed their contracts at a cell phone company when it expired.

For all datasets you can load in this way, you do `?DATA` (replacing `DATA` by the name of the data frame) to get a help file telling you exactly what every column in and what the dataset is about.

Loading data with read.csv

Most of the datasets we use in lecture are built into R via package `regclass`. To read in data from a file:

```
DATA <- read.csv("filename with extension")
```

Basic R Commands

The command `associate` (available once you have installed package `regclass` and done `library(regclass)`) will perform all aspects of the analysis. It's good to know the more basic commands as well. Let `x` and `y` be the column names in the data frame `DATA`.

- `plot(y~x,data=DATA)` - makes a mosaic plot or side-by-side barcharts
- `table(DATA$x,DATA$y)` - makes a contingency table
- `hist(DATA$y)` - makes a histogram of y
- `aggregate(DATA$y,by=list(DATA$y),mean)` - finds average value of y for each level of x (replace `mean` with `median` to get medians)
- `qq(DATA$y)` - QQ plot (from package `regclass`). Also available by doing `qqnorm(DATA$y)`.
- `mosaic(y~x,data=DATA)` - a mosaic plot (from package `regclass`)

Using R

We will use the (custom) command `associate()` to perform the test. You will have to load up library `regclass` first.

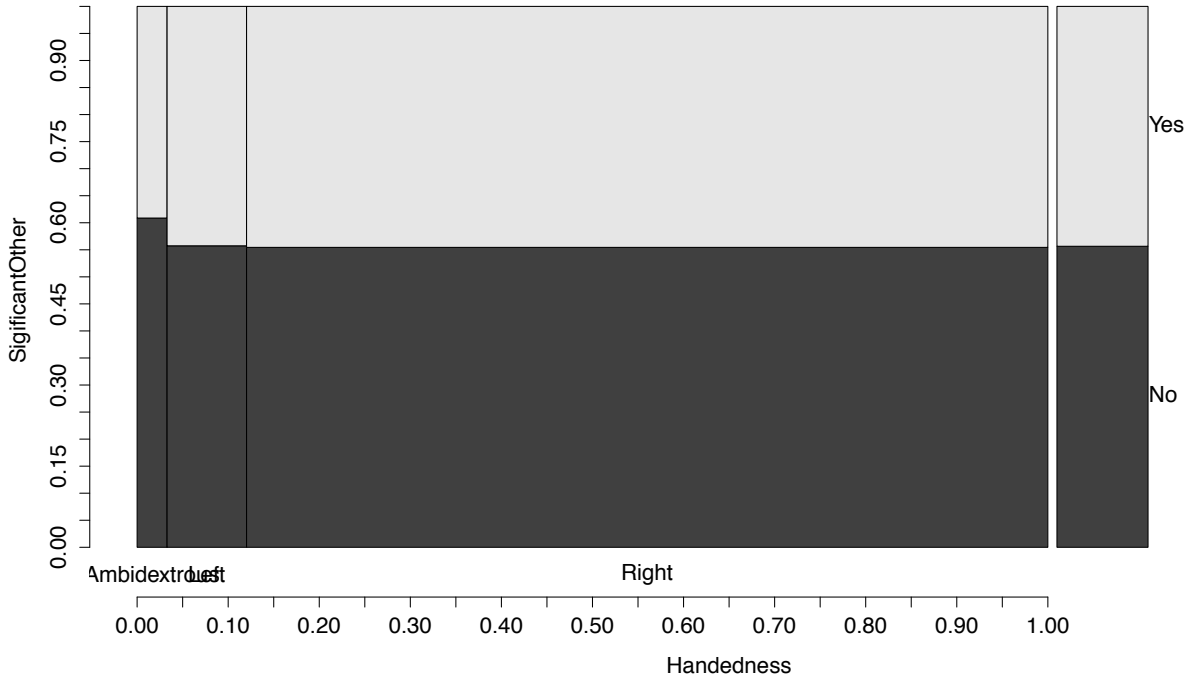
```
associate(y~x,data=...,permutations=500,seed=...)
```

- `y` and `x` are the column names in the data frame
- fill in `data=` with the name of the data frame. This argument can be omitted if you defined x and y manually using the left arrow convention.
- `permutations` gives the number of permutation datasets to produce. If the argument is omitted, 500 will be made.
- `seed` is an optional argument that provides the random number seed. Since the p-value is approximated by randomly pairing x and y values, it can/will differ if you run the command again. Setting `seed` to any positive integer will allow you to reproduce the results.

Running associate (2 categorical variables)

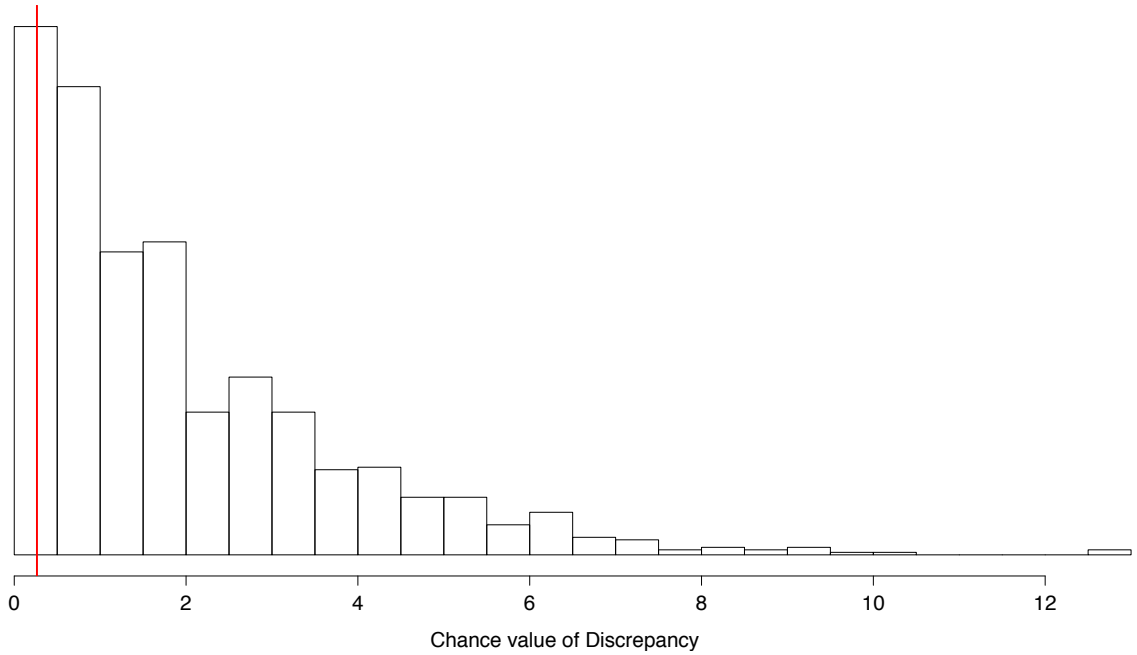
```
library(regclass) #need to load up regclass to use associate
data(SURVEY10) #loads up this dataset built-in to regclass
associate(SigificantOther~Handedness,data=SURVEY10,permutations=1000)
```

Mosaic plot - visualize gauge existence the strength of the association.



Running associate (2 categorical variables)

Sampling distribution of D - values of the discrepancy between observed and expected values (i.e., the discrepancy in the segmented bar charts in the mosaic plot) that can occur "by chance". Red line marks observed discrepancy. Check to see if it's out of line with what happens naturally when x and y are unrelated.



Running associate (2 categorical variables)

```
Association between Handedness (categorical) and  SigificantOther (categorical)
 using 699 complete cases
Contingency table:
              y
x                No Yes Total
  Ambidextrous   14   9    23
  Left           34  27    61
  Right         341 274   615
  Total         389 310   699

 Table of Expected Counts:
                  No   Yes
Ambidextrous    12.8  10.2
Left            33.9  27.1
Right          342.3 272.7
```

Running associate (2 categorical variables)

```
Conditional distributions of y (SigificantOther) for each group of x (Handednes
If there is no association, these should look similar to each other and
 similar to the marginal distribution of y
                    No       Yes
Ambidextrous 0.6086957 0.3913043
Left         0.5573770 0.4426230
Right        0.5544715 0.4455285
Marginal     0.5565093 0.4434907

Permutation procedure:
  Discrepancy Estimated p-value
    0.2643293             0.899
With 1000 permutations, we are 95% confident that:
 the p-value is between 0.879 and 0.917
If 0.05 is in this range, change permutations= to a larger number
```

Summary for Categorical/Categorical associations

After running `associate()`

- Look at the mosaic plot to see if the differences in segmented bar charts for the levels of x have noticeable, interesting differences that would carry practical significance. If not, no need to do statistically analysis.
- Check the p-value and it's 95% confidence interval to confirm enough permutations were run (i.e. there is no doubt of whether it is above 0.05 or below 0.05).
- Make a conclusion about the statistical and practical significance based on whether the p-value is < 0.05 (significant) or ≥ 0.05 (not significant). Note: pay attention to range of p-values given (since we are estimating it with a simulation). If 0.05 is inside the range the test is inconclusive and the command needs to be run again with a higher number of permutations (add `permutations=1000` or something).

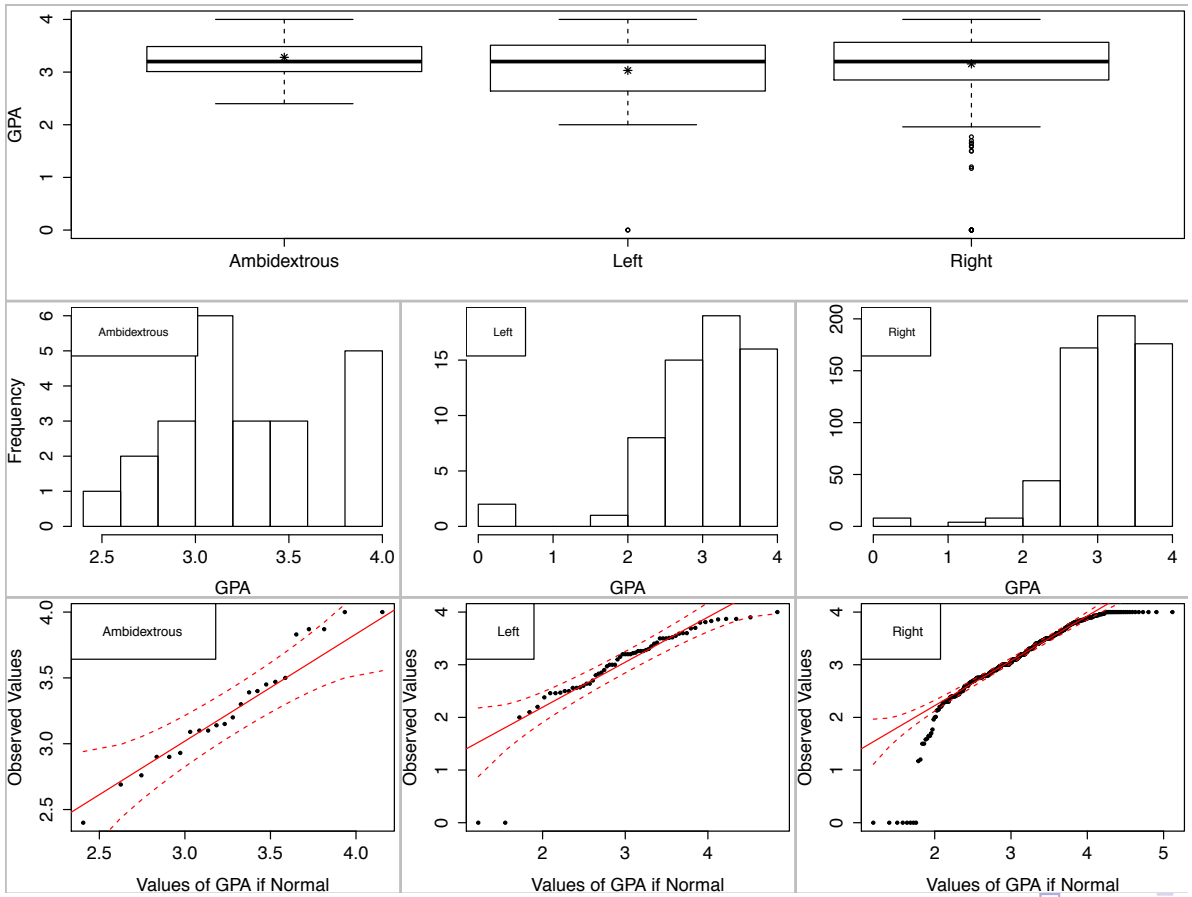
Running associate (1 categorical and 1 quantitative variable)

```
library(regclass) #if not already loaded up
data(SURVEY10) #if not already loaded up
associate(GPA~Handedness,data=SURVEY10,permutations=100,seed=1313)
```

Warning: there are a LOT of plots to see. Make sure the plotting window is large!

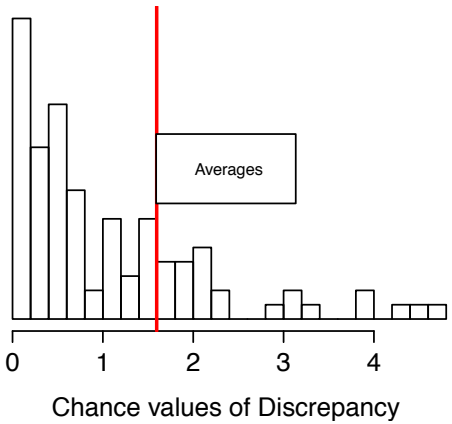
Running associate (1 categorical and 1 quantitative variable)

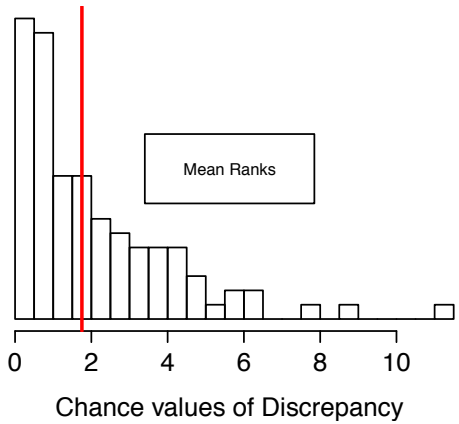
Visually gauge whether there is an association by comparing averages (*'s in the boxplots) if the distributions look approximately Normal in the QQ plots or medians (horizontal bars in boxplots) otherwise.

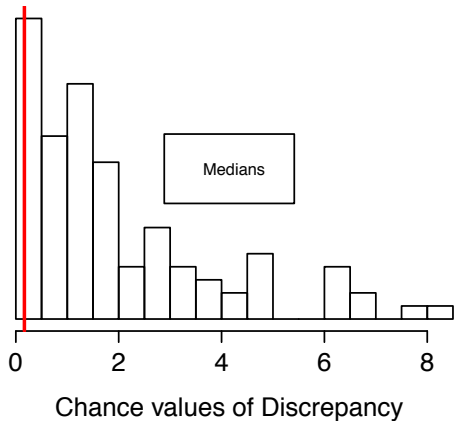


Running associate (1 categorical and 1 quantitative variable)

Sampling distribution of discrepancy in averages and medians that can occur "by chance". Red line marks observed discrepancy. Check to see if it's out of line with what happens naturally when x and y are unrelated.







Running associate (1 categorical and 1 quantitative variable)

```
Association between Handedness (categorical) and  GPA (numerical)
 using 699 complete cases

Sample Sizesx
Ambidextrous          Left         Right
          23            61           615

Permutation procedure:
                    Ambidextrous  Left Right Discrepancy Estimated p-value
Averages (ANOVA)             3.28 3.031 3.156       1.596                0.24
Mean Ranks (Kruskal)        367.8 363.4   348       1.753                0.41
Medians                       3.2   3.2   3.2      0.1666                0.87
With 100 permutations, we are 95% confident that
 the p-value of ANOVA (means) is between 0.16 and 0.336
 the p-value of Kruskal-Wallis (ranks) is between 0.313 and 0.513
 the p-value of median test is between 0.788 and 0.929
Note:  If 0.05 is in a range, change permutations= to a larger number
```

Note: make need to increase # permutations if the test is inconclusive (0.05 is inside the interval of p-values).

Summary for Quantitative/Categorical associations

After running `associate()`

- Look at the side-by-side boxplots and decide if you are comparing averages or medians. Also determine if the difference in typical values between levels of x is large enough to be of any practical significance (if not, no need to do statistical analysis).
- Look at the differences in means (or medians)
- Check the p-value and it's 95% confidence interval to confirm enough permutations were run (i.e. there is no doubt of whether it is above 0.05 or below 0.05).
- Make a conclusion about the statistical and practical significance based on whether the p-value is < 0.05 (significant) or ≥ 0.05 (not significant). Note: pay attention to range of p-values given (since we are estimating it with a simulation). If 0.05 is inside the range the test is inconclusive and the command needs to be run again with a higher number of permutations (add `permutations=1000` or something).