# Chapter 2 - Statistical Analysis of Associations Part 1

Adam Petrie
Department of Business Analytics
University of Tennessee

January 24, 2016

Readings: 2.1 - 2.5, pp. 15 - 52

# Definition of Association

# Relationships Matter

In business analytics, science, engineering, etc., people are interested in studying the nature, structure, and strength of relationships between two or more quantities.

- Chance of donating to UT and choice of college
- Promotion response rate and age
- Attractiveness and hair color
- Amount of money spent at The Home Depot and at Kroger
- Claim on fire insurance policy and policy type
- Churn and number of calls to customer support
- Moneyball - winning and player's on-base percentage

# Association

An association is a general term to describe a relationship between two variables.

> Two quantities are associated when, for whatever reason, knowing the value of one quantity tells you *something* about (i.e. narrows down) the possible values of the other.

In other words, information about one variable can be leveraged to learn something about the possible values of the other.

# Examples

Chance of donating to UT and choice of major

- 60% of alumni donate to UT overall
- However, 70% of alumni of the college of business, 75% of the college of law, and 40% of the college of arts and science donate.
- Knowing major provides additional information about the chance of donating.

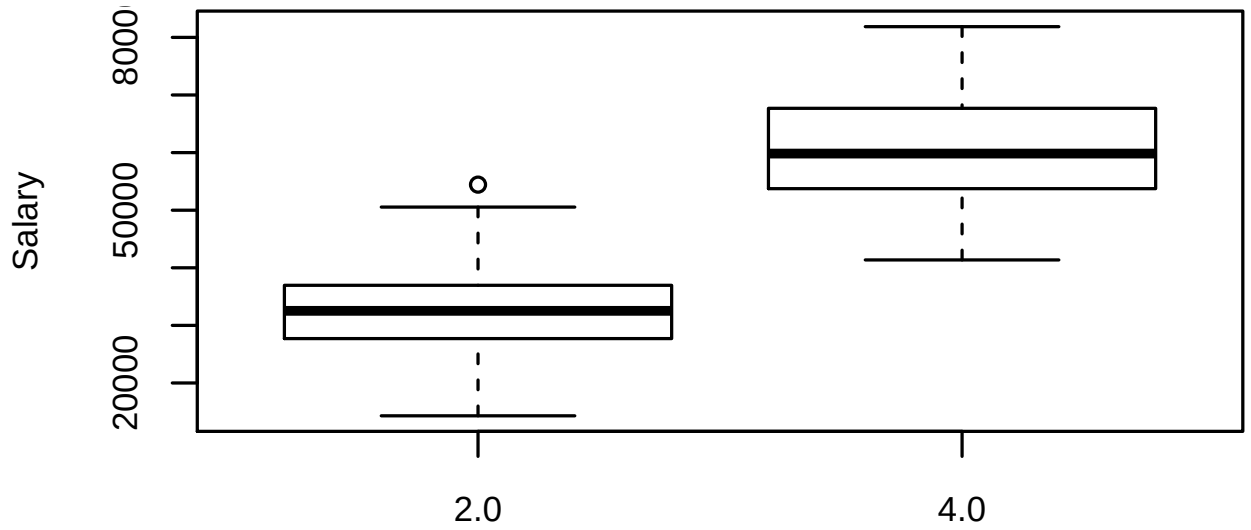
Amount of money spent at The Home Depot and at Kroger

- On average, shoppers spend $200 at Home Depot and $1500 at Kroger per year.
- Among shoppers that spend more than $1500 at Kroger, the average spent at Home Depot is $300
- Among shoppers that spend less than $1500 at Kroger, the average spent at Home Deport is $180
- Knowing amount spent at Kroger narrows down amount spent at Home Depot, so there is an association.

# Association Examples - Quantitative/Categorical

Salary and GPA

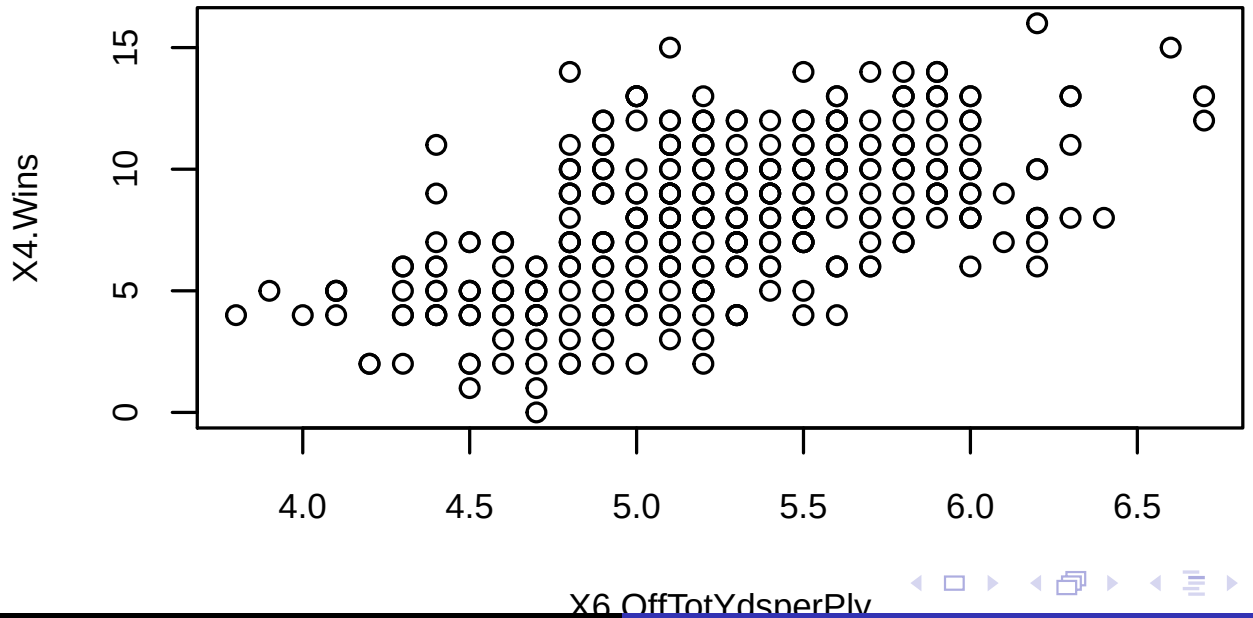
- Starting salaries of UT undergrads may range from \$15,000-\$80,000.
- Among students with 4.0s, salaries may typically range from \$40,000-\$80,000. Among students with 2.0s, salaries may range from \$15,000-\$50,000
- Knowing GPA narrows down the range of salaries so the two quantities have an association.



# Association Examples - Quantitative/Quantitative

Wins and Yards per Play (offense) in the NFL

- # wins ranges from 0 to 16
- Among teams with small values of yards ($\approx 4$) the number of wins typically varies between 3-7. For teams with large values of yards ($\approx 6$), the number of wins typically varies between 8-14.
- Knowing the average number of yards per offensive play narrows down the range of wins, so the two variables have an association.



# Line of analysis

1) Visualize the relationship using the appropriate plot: mosaic plot, side-by-side boxplot, or scatterplot.
2) Assess with your eyes whether the association is of any **practical significance**, i.e., large or noticeable enough to be important based on the context at hand.
3) If the association is strong enough to matter, perform a test to see if the association is **statistically significant**, i.e., unlikely to have been produced "by chance".

# Choosing roles of variables

To make plots, you must assign one variable to be $y$ and one to be $x$.

> $y$ is the quantity you wish to study/predict and $x$ is the quantity that you use to make predictions

While these choices matter for visualization, for the statistical test the choices are arbitrary.

# Choosing roles of variables

Examples:

- Association between gender and eye color? $x$ - gender; $y$ - eye color
- Association between buying and income? $x$ - income; $y$ - buying
- Association between sales and advertising? $x$-advertising; $y$ - sales
- Association between donating and income? $x$-income; $y$-donate
- Association between political outlook and Greek membership? Either assignment could make sense (are Greeks more likely to be conservative? are more liberal people more likely to not go Greek?)

Once $x$ and $y$ have been assigned, the R syntax is `plot(y ~ x,data=)`, where you fill in the name of your dataframe (or leave out that argument entirely if you defined them manually using the left arrow convention).

# Associations Between Two Categorical Variables

# Overview of categorical variables

Categorical variables have a finite number of possible values. These values are words, descriptions, or even numbers.

- Gender (male/female)
- Churn (yes/no) - recall churning means failing to renew a contract
- Marital status (single/married/dating)
- Height (short/average/tall)

It is possible to treat a quantitative (numerical) variable as a categorical variable by grouping values into categories, e.g., amount (0, 1, 2, 3, ...) can be recoded "none", "a few (1-3)", "many (4-7)", "lots (8+)".

# Levels

The possible values of a categorical variable are called **levels**.

# Frequency Table

To numerically summarize a categorical variable, we make a frequency table. A frequency table gives the total number of entities with each level.

For example, let us look at the `TIPS` dataset. Presented is the frequency table of the variable "Weekday" (the day that the data was recorded) and the variable "Time" (whether table was during the day or night).

```
data(TIPS)
table(TIPS$Weekday)
##
##   Friday Saturday   Sunday Thursday
##       19       87       76       62
table(TIPS$Time)
##
##   Day Night
##    68   176
```

# Relative Frequency Table

A **relative** frequency table tabulates the percentage of entities with each level. These percentages may not add up to 100% due to rounding.

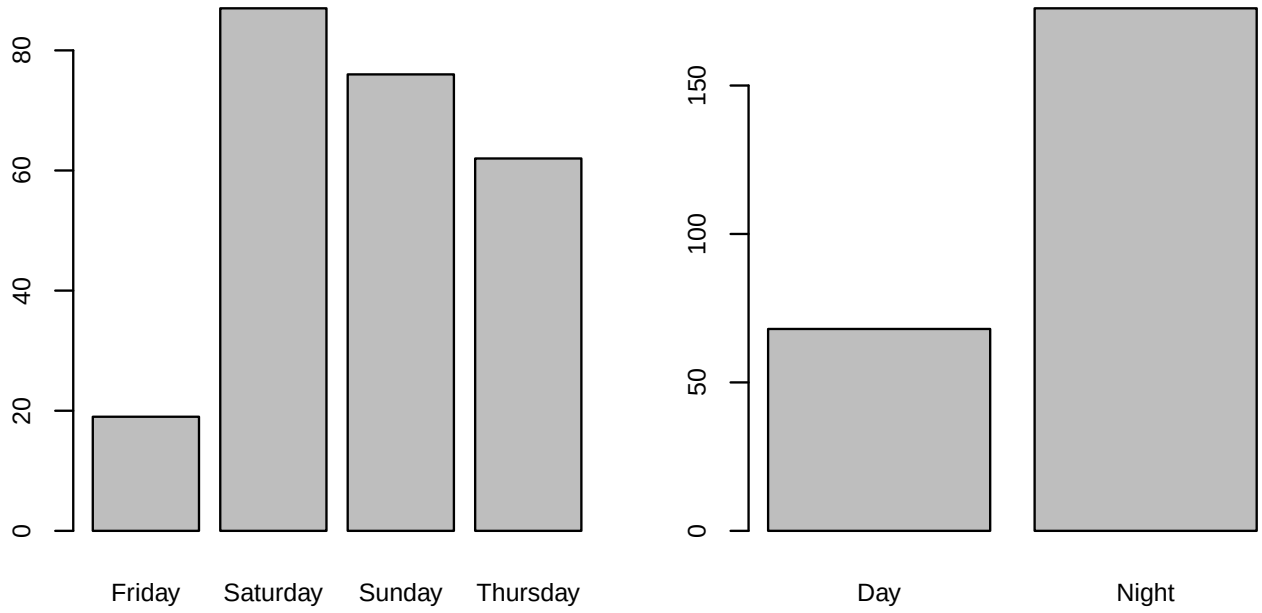
In R, you can make a relative frequency table by making a frequency table and dividing by the sample size (number of entries in the dataframe) by using the command `nrow`.

```
table(TIPS$Weekday)/nrow(TIPS)
##
##     Friday   Saturday     Sunday   Thursday
## 0.07786885 0.35655738 0.31147541 0.25409836
table(TIPS$Time)/nrow(TIPS)
##
##       Day     Night
## 0.2786885 0.7213115
```

# Bar Chart

A bar chart gives a graphical representation of the frequency table. Each level gets a bar, and the height of the bar is equal to the number of entities with that level. You can make a barchart by invoking the command `plot()`

```
plot(TIPS$Weekday)
plot(TIPS$Time)
```



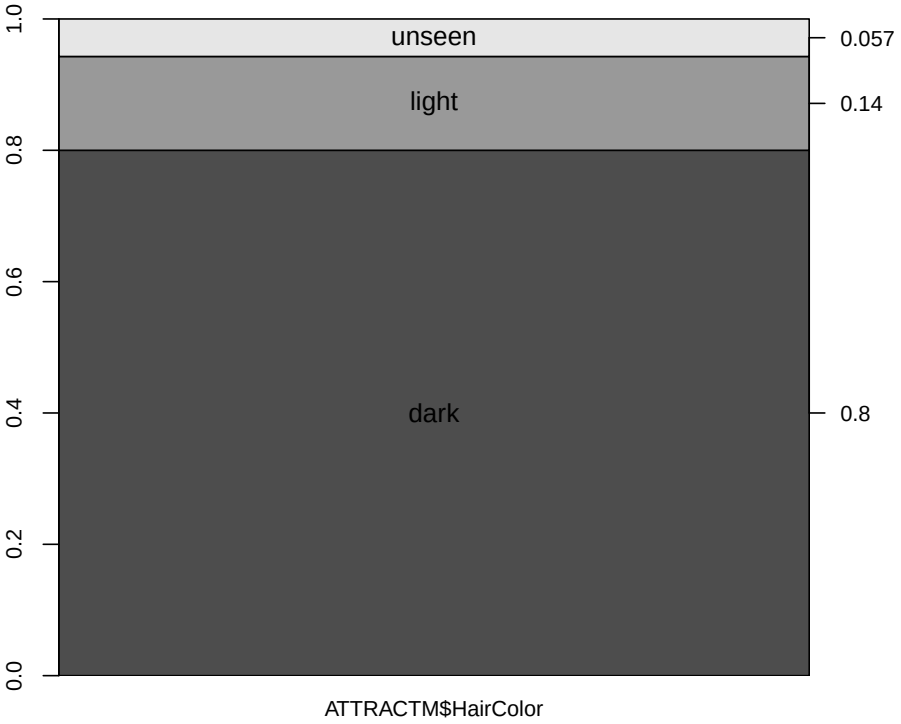
# Segmented Bar Chart and Mosaic Plot

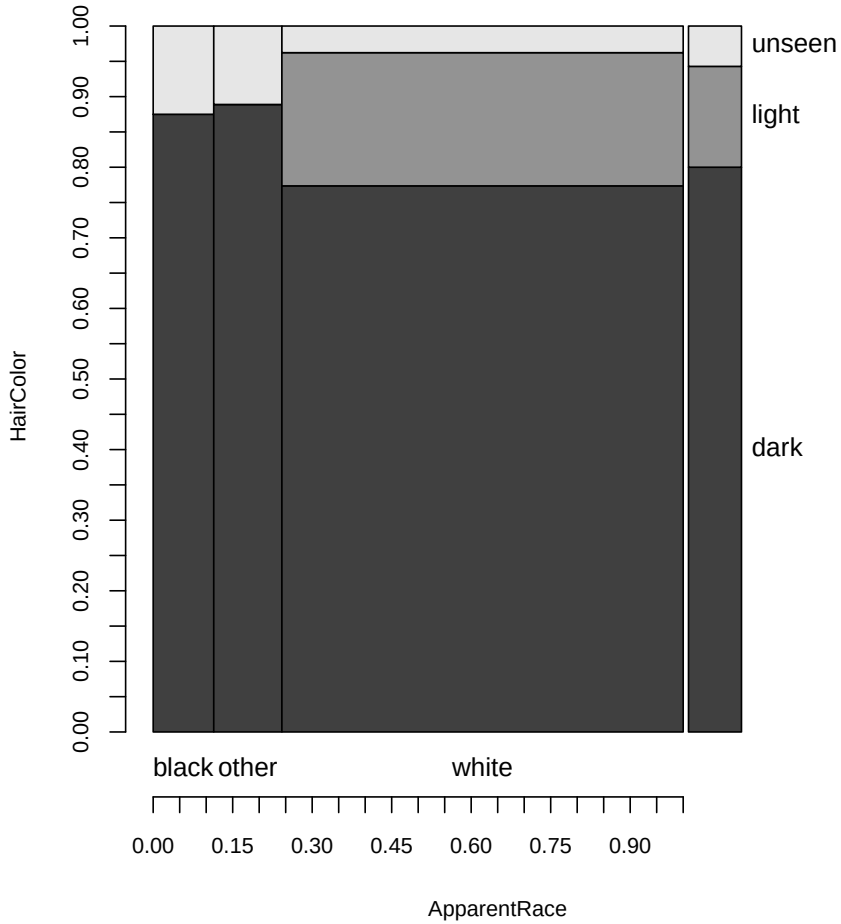
A **segmented bar chart** is another simple way to visualize the distribution of each level. The fraction of the bar devoted to a level is the percentage of individuals who have that level.

A **mosaic plot** consists of the side-by-side segmented bar charts for $y$ for each level of $x$ (this is why assigning $x$ and $y$ correctly is important). The plot allows us to easily discern where the distribution of $y$ varies between levels of $x$ and whether an association is strong.

The next slide shows an association between hair color and apparent race.

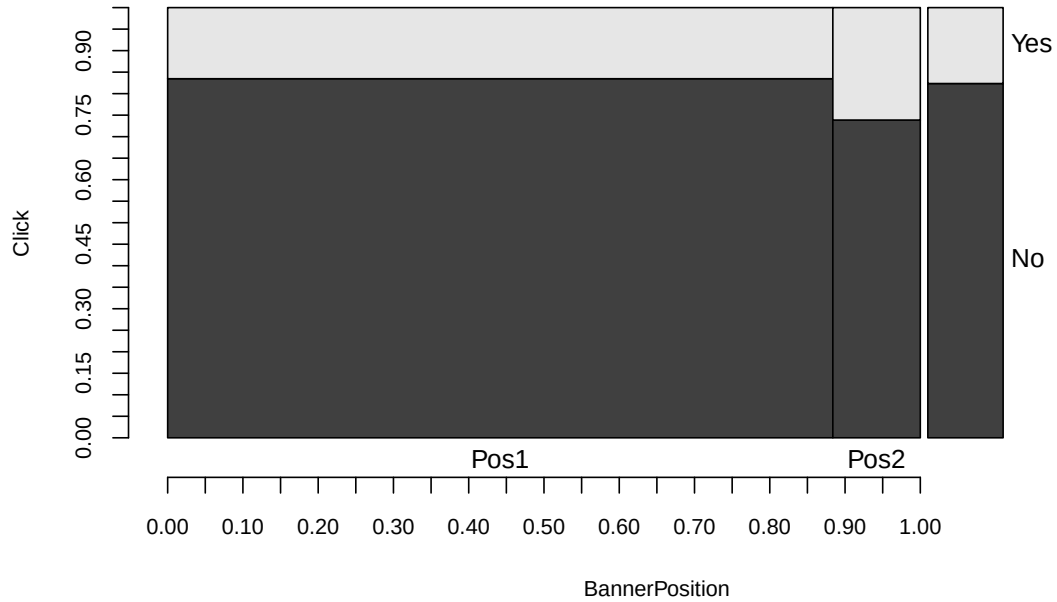
# Segmented Bar Chart and Mosaic Plot



# Example - click-thru

Chance of clicking on an ad on a mobile device and where the ad is placed on the page. Notice the segmented bar chart of the overall distribution on the right.



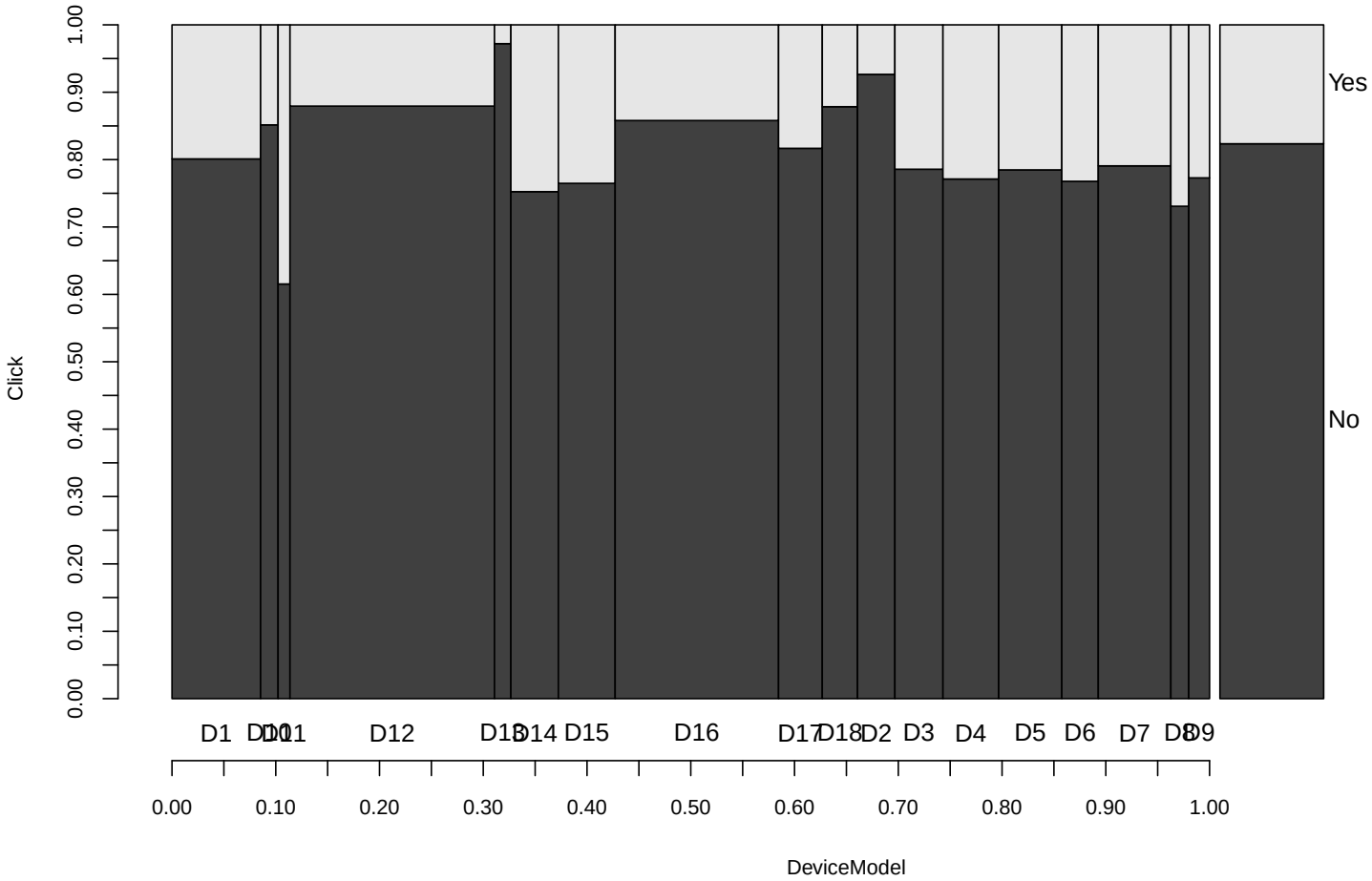
# Example - Click-thru

Click-thru vs. position

- Overall, 17.6% of people in the data clicked in this data.
- When the ad is in position 1, this chance is 16.5%. When the ad is in position 2, this chance is 26.1%.
- Position tells you something about the chance of click, so these two quantities have an association.

Click-thru vs. device type (next slide) - Over the 18 different devices, there does look to be some variation in the chance of clicking (notice the overall segmented bar chart on the right) so we suspect an association exists between click-thru rate and device model.

# Example - click-thru and device model



# Bar widths

The *widths* of each bar tell you the relative fraction of individuals with each level of $x$.

> When we look for the existence of an association, the widths of the bars are not important because they give information regarding the distribution of $x$, **not** whether the distribution of $y$ varies between levels of $x$.

Last slide: the bar for White is the thickest because most of the guys in the survey were white.

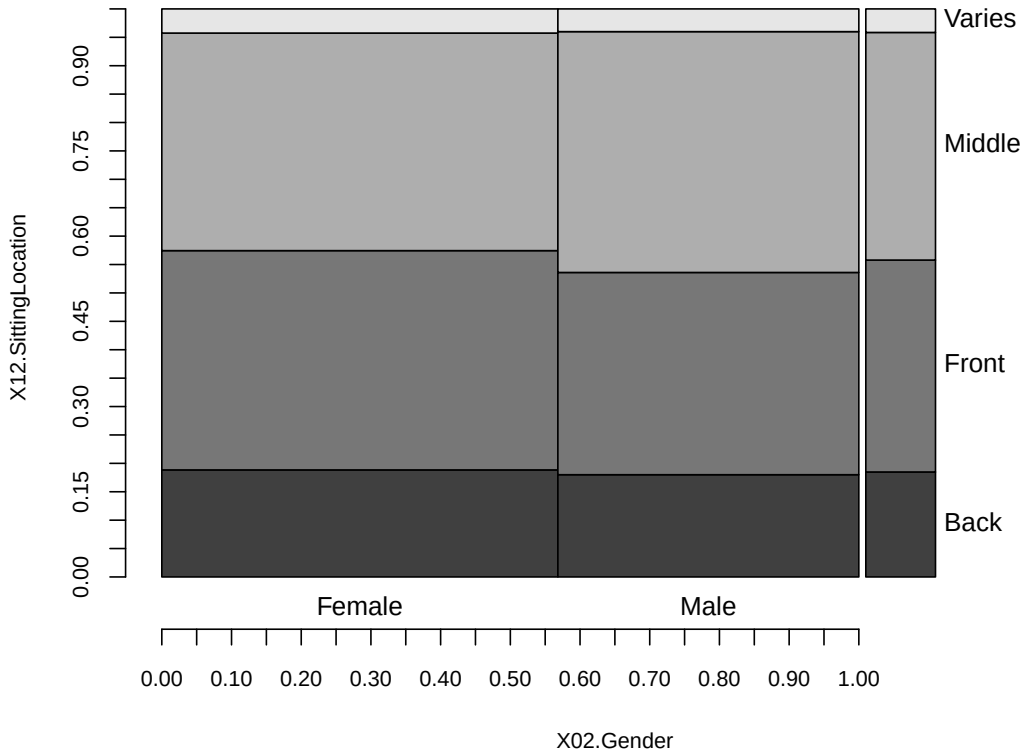
# What to look for

The first step in the analysis is to determine whether the association "looks" meaningful, i.e., it is strong enough to have practical importance. To do this, **look to see if the shadings in the bars are more or less the same**.

- If each bar is similar, the distribution of $y$ is about the same for all levels of $x$. Knowing $x$ doesn't narrow down the possible values of $y$ much, so there is little to no association.
- If there are noticeable differences among bars, the distribution of $y$ differs between at least two levels of $x$. Knowing $x$ narrows down the possible values of $y$ a lot, so there is an association.

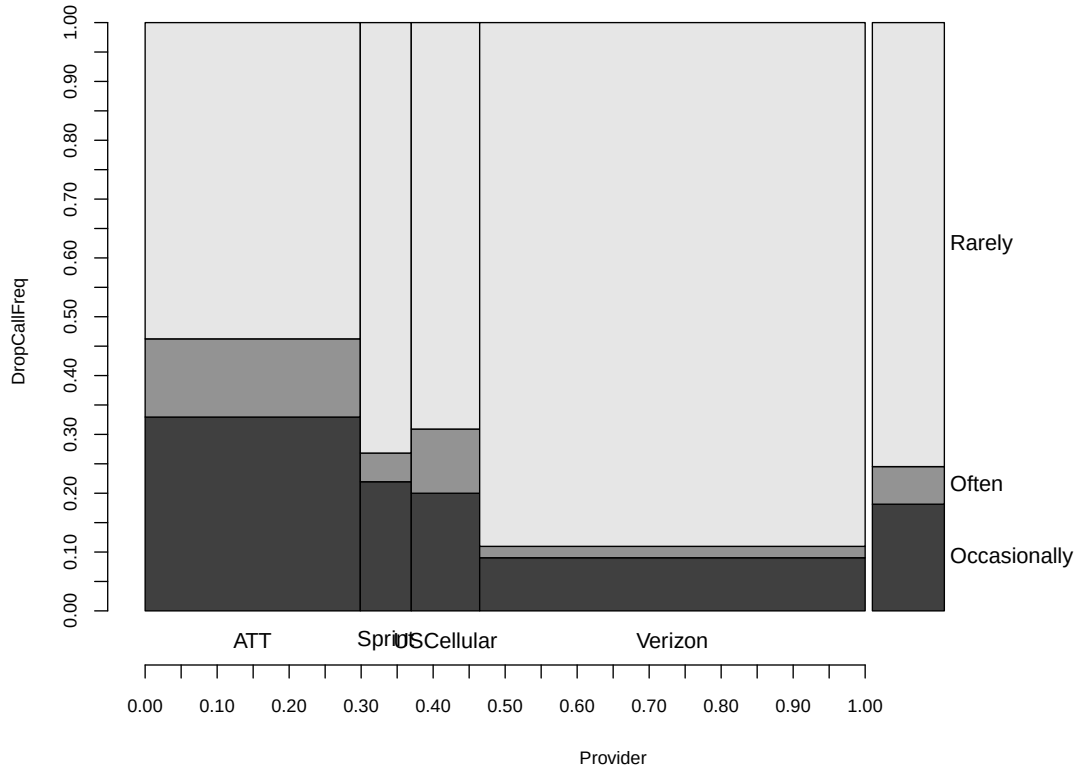
# Another example - UT first choice?

```
data(SURVEY09); mosaic(X12.SittingLocation~X02.Gender, data=SURVEY09)
```




The distribution of sitting location is nearly identical for both genders, so we do not suspect an association.

# Example - Dropped Calls




The distribution of "Dropped Call Frequency" looks different between providers (e.g., Verizon has a big chunk for "Rarely" while ATT does not).

The bars look *very* different indicating a strong association.

# Contingency table

To determine if there is a statistically meaningful association, we need to quantify just **how** the distributions of $y$ vary between levels of $x$. How should we operationalize the definition of "difference in distributions"?

We start with the **contingency table**. This gives the number of individuals who have each combination of $x$ and $y$ values.

| | Occasionally | Often | Rarely | Total |
|---|---|---|---|---|
| ATT | 57 | 23 | 93 | 173 |
| Sprint | 9 | 2 | 30 | 41 |
| USCellar | 11 | 6 | 38 | 55 |
| Verizon | 28 | 6 | 276 | 310 |
| Total | 105 | 37 | 437 | 579 |

# Contingency Table Analysis

The numbers in a contingency table are the **observed counts** in the data.

| | Occasionally | Often | Rarely | Total |
|---|---|---|---|---|
| ATT | 57 | 23 | 93 | 173 |
| Sprint | 9 | 2 | 30 | 41 |
| USCellular | 11 | 6 | 38 | 55 |
| Verizon | 28 | 6 | 276 | 310 |
| Total | 105 | 37 | 437 | 579 |

- 38 USCellular customers report Rarely having dropped calls (cell inside table)
- There are a total of 41 Sprint customers (right margin)
- There are 105 people who report Occasionally having dropped calls (bottom margin)
- The sample size is 579 people (bottom right)

# Marginal Distribution

The **marginal** (or overall) distribution of a variable is the overall frequency distribution of its levels.

The marginal distribution can be obtained by looking at the numbers in the margins of the contingency table. These give the row and column totals and show the overall (marginal) distribution of $x$ and of $y$, respectively.

The numbers are often converted into percentages by dividing by the sample size.

# Marginal Distribution of Cell Phone Carrier

| | Occasionally | Often | Rarely | Total |
|---|---|---|---|---|
| ATT | 57 | 23 | 93 | **173** |
| Sprint | 9 | 2 | 30 | **41** |
| USCellular | 11 | 6 | 38 | **55** |
| Verizon | 28 | 6 | 276 | **310** |
| Total | 105 | 37 | 437 | 579 |

The marginal distribution of Cell Phone Carrier is

| ATT | Sprint | USCellular | Verizon |
|---|---|---|---|
| 173 | 41 | 55 | 310 |
| 0.30 | 0.07 | 0.09 | 0.54 |

The percentages come from taking the count and dividing by the sample size, i.e. $55/549 = 0.09$, after rounding.

# Marginal Distribution of Dropped Call Frequency

| | Occasionally | Often | Rarely | Total |
|---|---|---|---|---|
| ATT | 57 | 23 | 93 | 173 |
| Sprint | 9 | 2 | 30 | 41 |
| USCellular | 11 | 6 | 38 | 55 |
| Verizon | 28 | 6 | 276 | 310 |
| Total | **105** | **37** | **437** | 579 |

The marginal distribution of Dropped Calls is

| Occasionally | Often | Rarely |
|---|---|---|
| 105 | 37 | 437 |
| 0.18 | 0.06 | 0.75 |

The percentages are found by taking the counts and dividing by the sample size, i.e. $105/549 = 0.18$, after rounding.

# Conditional distribution

The **conditional** distribution of $y$ (given a particular level of $x$) is the percentages of individuals with that particular level of $x$ who have each level of $y$. In other words, the conditional distribution of $y$ given $x$ is the relative frequency distribution of $y$ for a **particular** level of $x$.

When there is no association, the distribution of $y$ for **each** level of $x$ should be similar to the marginal distribution. In other words, knowing $x$ does NOT narrow down the possible values of $y$.

# Conditional Distribution for ATT customers

| | Occasionally | Often | Rarely | Total |
|---|---|---|---|---|
| ATT | 57 | 23 | 93 | 173 |
| Sprint | 9 | 2 | 30 | 41 |
| USCellular | 11 | 6 | 38 | 55 |
| Verizon | 28 | 6 | 276 | 310 |
| Total | 105 | 37 | 437 | 579 |

The conditional distribution of dropped call frequency for ATT is

| Occasionally | Often | Rarely |
|---|---|---|
| 57 | 23 | 93 |
| 0.33 | 0.13 | 0.54 |

The percentages are found by taking the counts and dividing by the number of ATT customers, e.g., $57/173 = 0.33$, after rounding.

# Conditional Distributions and the Mosaic Plot

The mosaic plot shows the conditional distribution of $y$ for each level of $x$.

If there is no association, then the conditional distribution of $y$ for each level of $x$ should closely resemble the marginal (overall) distribution of $y$.

In other words, when $y$ and $x$ have no association then $y$ has the same distribution regardless of $x$ (no information about $x$ can be leveraged to learn anything about $y$).

# Conditional Distributions for dropped calls

Calculating the conditional distributions of dropped call frequency for each carrier we get:

| Carrier | Occasionally | Often | Rarely |
|---|---|---|---|
| ATT | 0.33 | 0.13 | 0.54 |
| Sprint | 0.22 | 0.05 | 0.73 |
| USCellular | 0.20 | 0.11 | 0.69 |
| Verizon | 0.09 | 0.02 | 0.89 |
| Marginal | 0.18 | 0.06 | 0.75 |

There are *substantial* differences between the marginal distribution of $y$ and the conditional distributions of $y$. Knowing the cell phone carrier greatly narrows down the possibilities for dropped call frequency. We thus suspect an association (this is the same information we gained in the mosaic plot).

# Defining discrepancy between distributions

There are many ways operationalize the definition of "difference" in conditional distributions. We will use the popular method of comparing the observed counts ($O$) in the contingency table to the counts we would have *expected* ($E$) had their been no association.

We will call the difference in conditional distribution by the word **discrepancy** and define it as:

$$D = \text{sum over every value in table of } \frac{(O-E)^2}{E}$$

$O$ is given from the table. $E$ must be computed.

# Expected counts

The formula for the discrepancy between distribution involves the number we would *expect* in the contingency table had there been no association. There is a formula which gives us this number.

For each value in the table, $E$ is:

$$E = \frac{\text{Row Total} \times \text{Column Total}}{n}$$

# Finding expected counts

Let's illustrate where this formula comes from with the dropped call data by answering the question:

How many Verizon customers would we expect to Rarely have dropped calls if there was no association between carrier and dropped call frequency?

- If there was no association, then the distribution of dropped call frequency should be the same for each carrier.
- The distribution for Verizon would be the same as the overall, marginal distribution.

| Occasionally | Often | Rarely |
|---|---|---|
| 0.18134715 | 0.06390328 | 0.7547496 |

- Overall, 75% of people rarely experience dropped calls. If there were no association, this should be true for Verizon customers as well. There are 310 Verizon customers, so we expect 75.475% of them, or $0.75475 \times 310 = 234.0$ to report rarely having dropped calls.

# Finding expected counts

The expected count was essentially $0.75 \times 310 = 234.0$.

- The 75% can be seen to be the column total for Rarely divided by the sample size (437/579)
- 310 is the row total for this Verizon.
- Putting this together we get the row total times column total divided by the "grand total" (sample size).

$$E = \frac{\text{Row Total} \times \text{Column Total}}{n}$$

# Calculating D

Once each expected count has been calculated, the tedious process of calculating $D$ can occur. Software will do this for us.

| Observed | Occasionally | Often | Rarely | Total |
|---|---|---|---|---|
| ATT | 57 | 23 | 93 | 173 |
| Sprint | 9 | 2 | 30 | 41 |
| USCellular | 11 | 6 | 38 | 55 |
| Verizon | 28 | 6 | 276 | 310 |
| Total | 105 | 37 | 437 | 579 |

| Expected | Occasionally | Often | Rarely |
|---|---|---|---|
| ATT | 31.4 | 11.1 | 130.6 |
| Sprint | 7.4 | 2.6 | 30.9 |
| USCellular | 10.0 | 3.5 | 41.5 |
| Verizon | 56.2 | 19.8 | 234.0 |

$$D = \frac{(57-31.4)^2}{31.4} + \frac{(23-11.1)^2}{11.1} + \ldots + \frac{(6-19.8)^2}{19.8} + \frac{(276-234.0)^2}{234.0}$$

$$D = 78.7$$

# How big a D is evidence of an association?

The numerical value of $D$ tells us how much of a discrepancy exists between the conditional distributions of $y$ and the marginal distribution of $y$, and can be calculated from the counts we observed and the counts we'd expect if there is no association.

Values of $D$ farther from zero indicate larger discrepancies and thus stronger associations. The key question is:

> How large does $D$ need to be for us to be convinced that there is an association?

# How big a D is evidence of an association?

With the dropped call data, $D = 78.7$. Is this a large value for the discrepancy? Could we have observed this value when all carriers have the same fundamental distribution of dropped call frequency?

We can find the values of $D$ we expect to find "by chance" with a simulation called the **permutation procedure**.

# Permutation procedure

To determine the possible values of $D$ that may arise when there is no association we will use the **permutation procedure**.

- Create an artificial **permutation** dataset where the observed values of $x$ and $y$ are *randomly* paired together. By design, there is no association in this dataset.
- Calculate $D$ for this permutation sample and record it.
- Repeat this process a "lot" of times and make a histogram of the values of $D$ to see what happens "by chance" when the variables have no association.

The distribution of possible values of $D$ that arise from this simulation (i.e., through natural variation when no association exists) is called the **sampling distribution** of $D$.

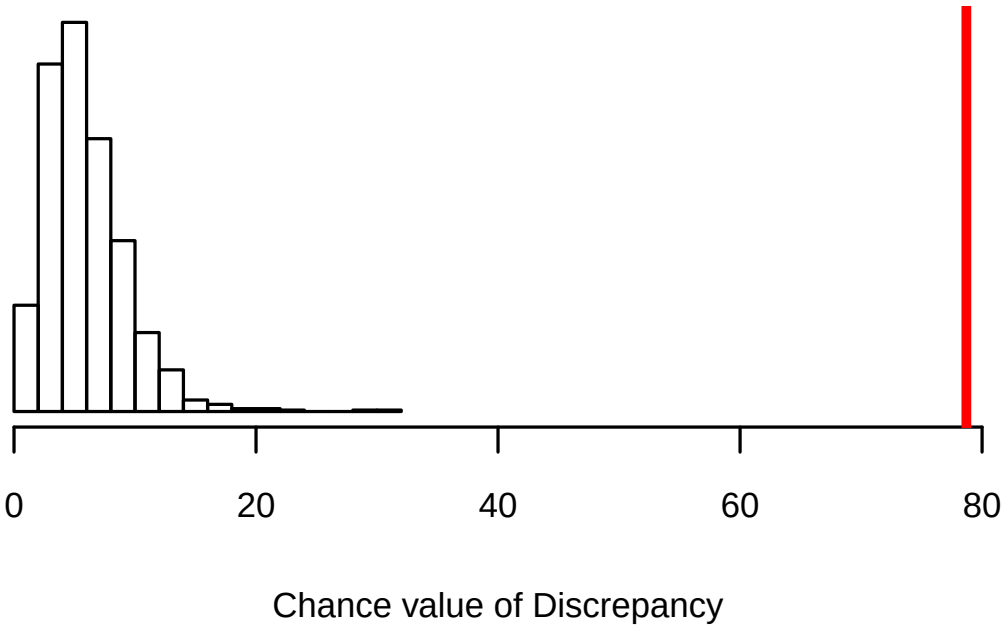
# Permutation sample

The **permutation procedure** is most easily illustrated when $x$ and $y$ are both numerical. Below is a table of the original observed values and of three permutation datasets where the values of $x$ and $y$ have been randomly paired together (values in each sample have been sorted on $x$).

| Individual | Observed Data | | Permutation 1 | | Permutation 2 | | Permutation 3 | |
|---|---|---|---|---|---|---|---|---|
| | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
| 1 | 2.2 | 0.2 | 2.2 | 3.4 | 2.2 | 5.6 | 2.2 | 3.4 |
| 2 | 2.7 | 3.4 | 2.7 | 6.0 | 2.7 | 3.4 | 2.7 | 0.2 |
| 3 | 2.8 | 3.4 | 2.8 | 5.6 | 2.8 | 6.0 | 2.8 | 5.6 |
| 4 | 4.9 | 5.6 | 4.9 | 3.4 | 4.9 | 3.4 | 4.9 | 3.4 |
| 5 | 7.2 | 6.0 | 7.2 | 0.2 | 7.2 | 0.2 | 7.2 | 6.0 |

# Simulation results for dropped call data

The histogram shows the values of $D$ observed over 1000 permutation datasets. The value of $D$ observed in the data is 78.7, which is *highly* out of line with what occurs "by chance".




The discrepancy, and thus the association, is "statistically significant" since it is unlikely to have arisen by chance.

# p-value of discrepancy

To perform the statistical analysis, we calculate the $p$-value of the association.

> The **p-value** of the observed discrepancy $D$ is the probability that a permutation sample (created by randomly pairing $x$ and $y$ together so that they do not have an association by design) would produce a value of $D$ *at least as large as what was observed in the original data*.

In other words, the $p$-value of the association is the probability of seeing at least as large a discrepancy between the conditional distributions of $y$ and the marginal distribution of $y$ "by chance" alone.

# Statistical Significance

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

- Such a large discrepancy is unlikely to occur by chance (i.e., it happens less than 5% of the time).
- Note: a small *p*-value does not mean the association is strong.

If the *p*-value is 5% or greater, we say that the association is **not statistically significant**.

- An observed discrepancy of this magnitude happens "all the time" (with a greater than 5% chance) when variables do not have an association.
- There is insufficient evidence to suggest the variables are related.
- Note: a large *p*-value does not mean an association for sure does not exist; the association may have been too weak to be detected by the data.

# Calculation of p-value

The $p$-value of the discrepancy $D$ can be approximated by the permutation procedure

- Count up the number of permutation datasets which had a discrepancy of $D$ or greater.
- Divide this by the number of permutations datasets that were generated.

For example, imagine you observed a value of $D = 51.1$ from your data and you made 50,000 permutation datasets. If you find that 23 of them had a value of $D \geq 51.1$, the approximate $p$-value would be $23/50000 = 0.00046$.

# The test in 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.

# Example: dropped call data

Let us make 1000 permutation datasets and, for reproducibility, set the random number seed (you will get the exact same results if you )

```
associate(DropCallFreq~Provider,data=CALLS,permutations=1000,seed=2015)
```

# Example: dropped call data output 1

The text output sent to the Console gives you the contingency table (observed counts) and the table of expected counts if $x$ and $y$ did not have an association.

```
Association between Provider (categorical) and  DropCallFreq (categorical) usin
Contingency table:
          y
x          Occasionally Often Rarely Total
  ATT                57    23     93   173
  Sprint              9     2     30    41
  USCellar           11     6     38    55
  Verizon            28     6    276   310
  Total             105    37    437   579

   Table of Expected Counts:
         Occasionally Often Rarely
ATT              31.4  11.1  130.6
Sprint            7.4   2.6   30.9
USCellar         10.0   3.5   41.5
Verizon          56.2  19.8  234.0
```

# Example: dropped call data output 2

The text output sent to the Console gives you the conditional/marginal distributions.

```
Conditional distributions of y (DropCallFreq) for each level of x (Provider):
If there is no association, these should look similar to each other and
 similar to the marginal distribution of y
           Occasionally      Often    Rarely
ATT          0.32947977 0.13294798 0.5375723
Sprint       0.21951220 0.04878049 0.7317073
USCellular   0.20000000 0.10909091 0.6909091
Verizon      0.09032258 0.01935484 0.8903226
Marginal     0.18134715 0.06390328 0.7547496
```

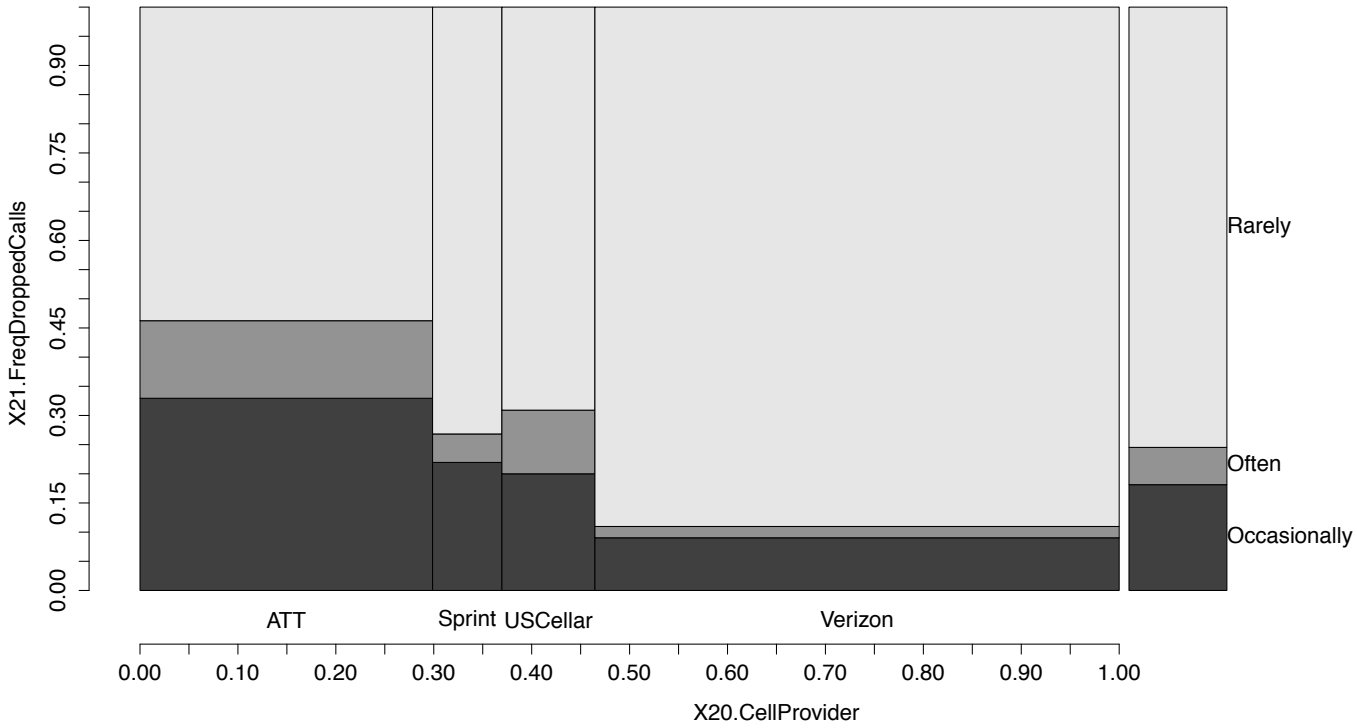
# Example: dropped call data output 3

The text output sent to the Console gives you the discrepancy between observed and expected counts and the approximate $p$-value as found with the permutation procedure.

```
Permutation procedure:
  Discrepancy Estimated p-value
     78.65499                 0
With 1000 permutations, we are 95% confident that:
 the p-value is between 0 and 0.004
If 0.05 is in this range, change permutations= to a larger number
```

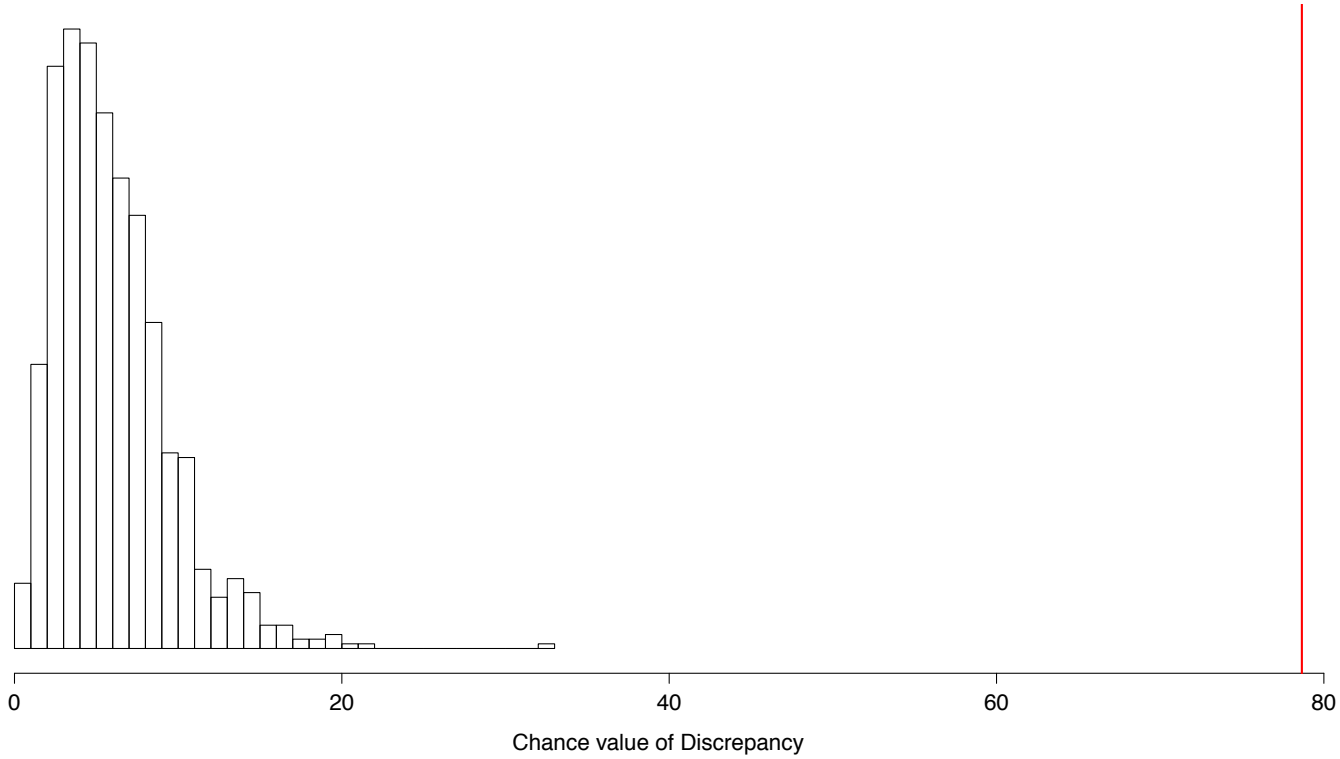
# Example: dropped call data output 4

The top plot is the mosaic plot along with a segmented bar chart of the marginal distribution of $y$.



# Example: dropped call data output 5

The bottom plot is the distribution of the values of discrepancy calculated on the permutation datasets (where $x$ and $y$ by design have no association). The red line indicates the value of the discrepancy observed in your data.



# Example: dropped call data conclusion

Since the $p$-value is less than 5%, we conclude that the association between provider and dropped call frequency is statistically significant. The implication is that the distribution of dropped call frequency is somehow different between carriers (i.e., not all carriers are created equal).

Looking at the mosaic plot, the differences in carriers is large and of practical significance. It looks like back in 2009 Knoxville, Verizon was definitely the provider to have. Things may have changed since then.

# Statistical vs. Practical significance

# Statistical vs. Practical significance

An association is **statistically significant** if it is unlikely that a pair of unrelated variables would exhibit such a large difference in conditional distributions (i.e., at least as large a discrepancy between observed and expected counts) as the ones that we observed in our data.

- This does *not* mean that the association is strong, interesting, or important
- For very large sample sizes, even extremely weak associations can be statistically significant.

Always look at the plots for signs of **practical significance**, i.e., a large enough difference that matters and is meaningful to you (subjective).

# Note about *p*-values

**Important note about *p*-values of test**

The *p*-value is *estimated* from the permutation procedure as the fraction of datasets where $D$ exceeded the observed value by chance. Due to natural variability (you don't expect a fair coin to land heads exactly 250 out of 500 flips), *there is still a little uncertainty in the quoted p-value* (the number may be different if you change the random number seed).

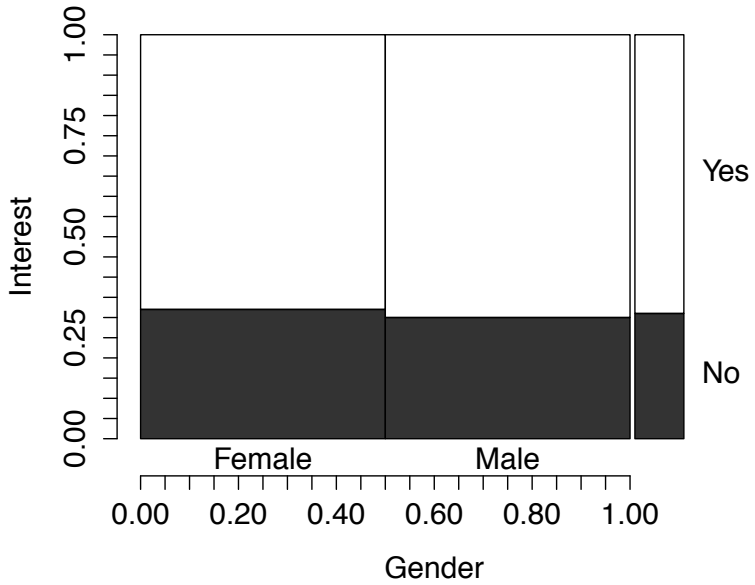
The output in R gives a *range* for the *p*-value of the test. If the 0.05 is inside the range of quoted *p*-values, the test is inconclusive. Increase the number of permutations until you are confident that the *p*-value is either below 5% or above 5%.

# Example - frequency flier interest

A survey of 50 people asked whether they would be interested in a new frequent flier program. Very small difference in distributions and $p$-value says no significant association.

```
data(SMALLFLYER)
associate(Interest~Gender,data=SMALLFLYER)
```

```
                Discrepancy    pvalue
Permutation test  0.04675082 1.0000000
```



# Example - frequency flier interest (cont)

Now imagine just duplicating the original dataset 1000 times and rerunning the test – the discrepancy is highly significant even though the difference in distributions still looks VERY small.

```
data(LARGEFLYER)
associate(Interest~Gender,data=LARGEFLYER)
```

```
  Discrepancy Estimated p-value
     46.65736                 0
```

# Example - clicking

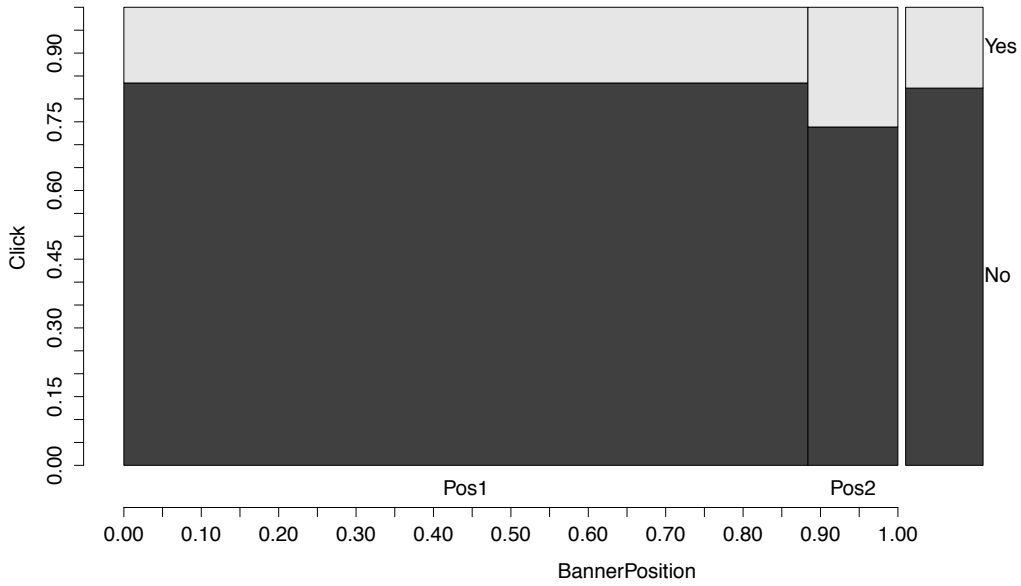
Earlier we saw that the fraction of users who click on an ad seemed to vary somewhat on the type of device they were using. What about the position of the ad?

```
data(EX6.CLICK)
associate(Click~BannerPosition,data=EX6.CLICK)
```

```
  Discrepancy Estimated p-value
     88.04572                 0
```

# Example - clicking

The association is highly statistically significant, so it looks like the position "matters". However, with nearly 14000 observations even very weak associations will be statistically significant. Indeed, the mosaic plot shows that the difference is relatively small, so it doesn't matter "much" (the practical significance is somewhat low, but may be exploitable).



# Associations Between a Categorical and Quantitative Variable

# Friendship analytics

In advertising (and many other things), the physical appearance of a spokesperson, actor, actress, etc., matters.

- If selling beauty, style, fashion products, person should be attractive.
- If selling insurance, person should look trustworthy and authoritative.
- If selling household products, person should look relatable and likely to actually use the product.

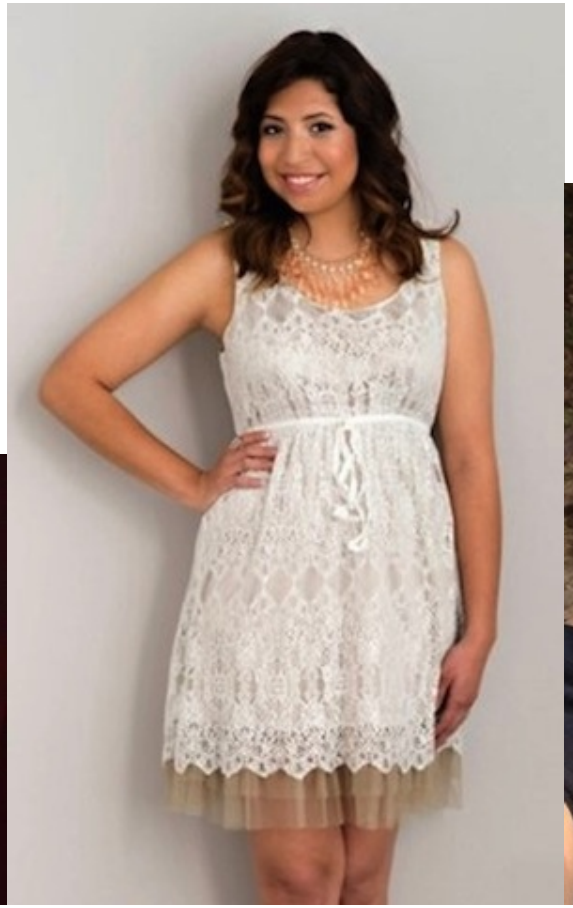
It is imperative to figure out "what matters" when determining how people perceive someone. Take the case of finding factors associated with someone's "friendship potential".

- hair color, eye color
- smile, glasses
- weight, complexion

# Friendship survey

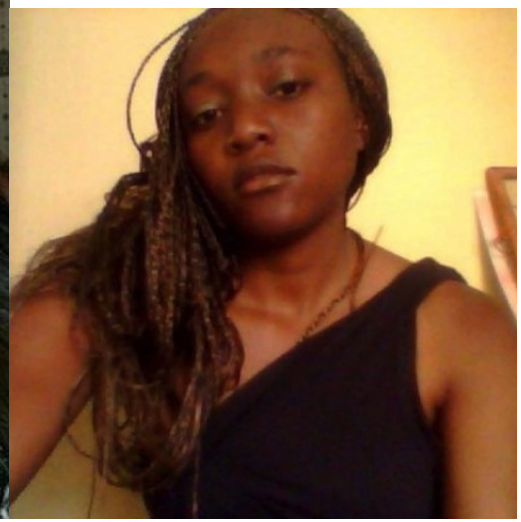
You were asked to rate, on a scale of 1 (low) to 5 (high), how likely it is that you could be friends with 70 people. You were also asked to rate a few characteristics of people that may influence scores (nerdiness, professionalism, etc.). So how can we tell what factors matter?

# Girls with most friendship potential



# Girls with least friendship potential





# Guys with most friendship potential

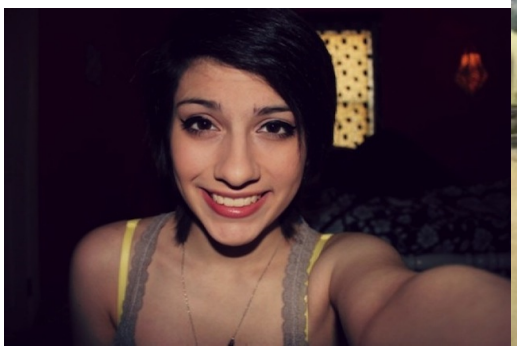
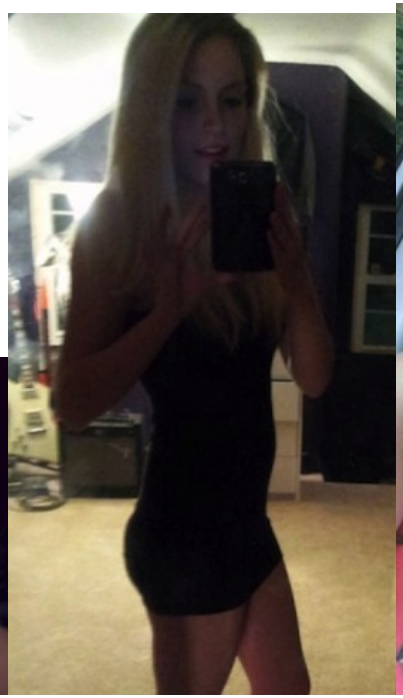
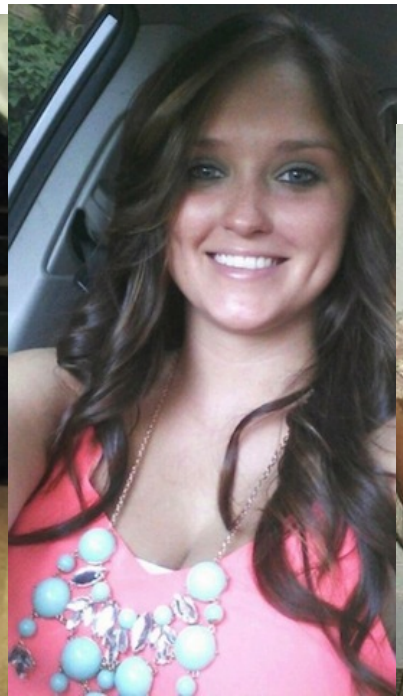



# Guys with least friendship potential

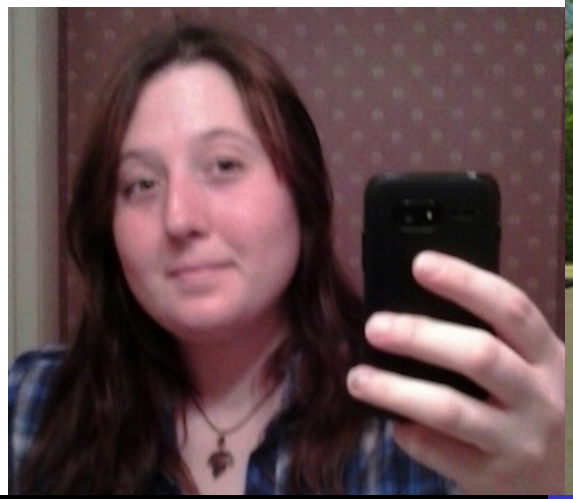


# Most attractive girls

Classes in the past rated attractiveness instead of friendship potential.

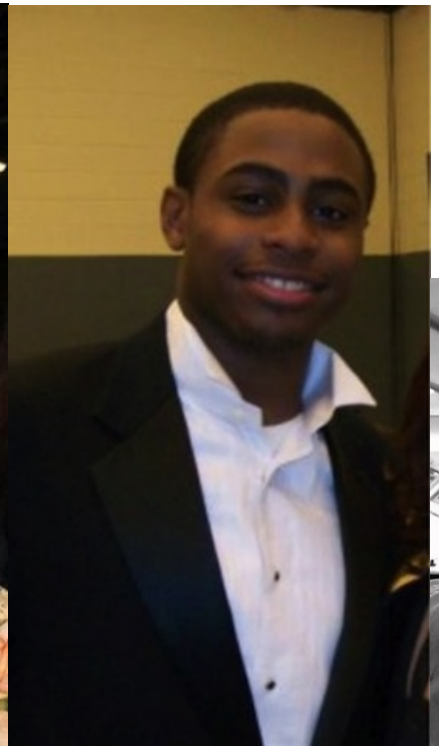





# Least attractive girls

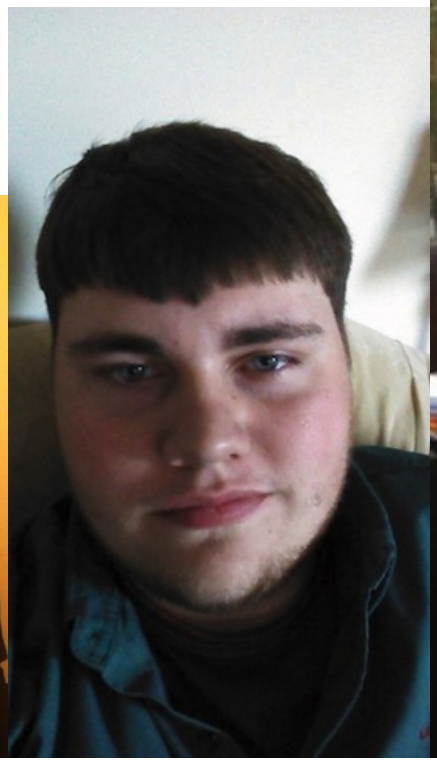
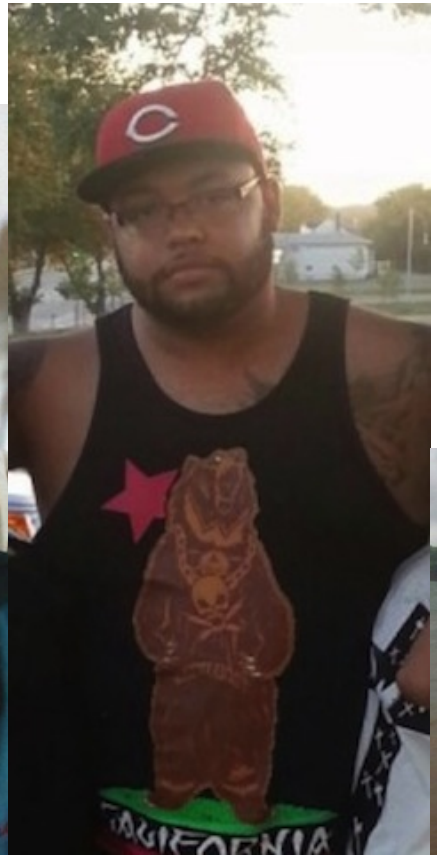
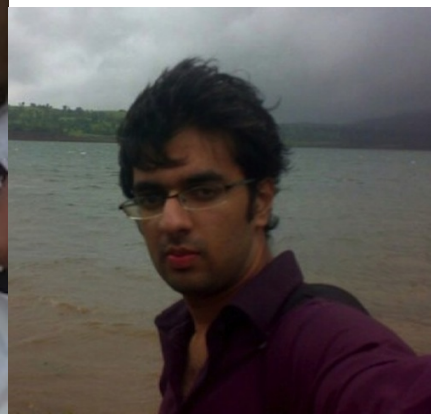


# Most attractive guys

Past classes rated attractiveness instead of friendship potential. Notice that 3 of the 4 top guys are the same!

# Least attractive guys

# Formal Definition of Association

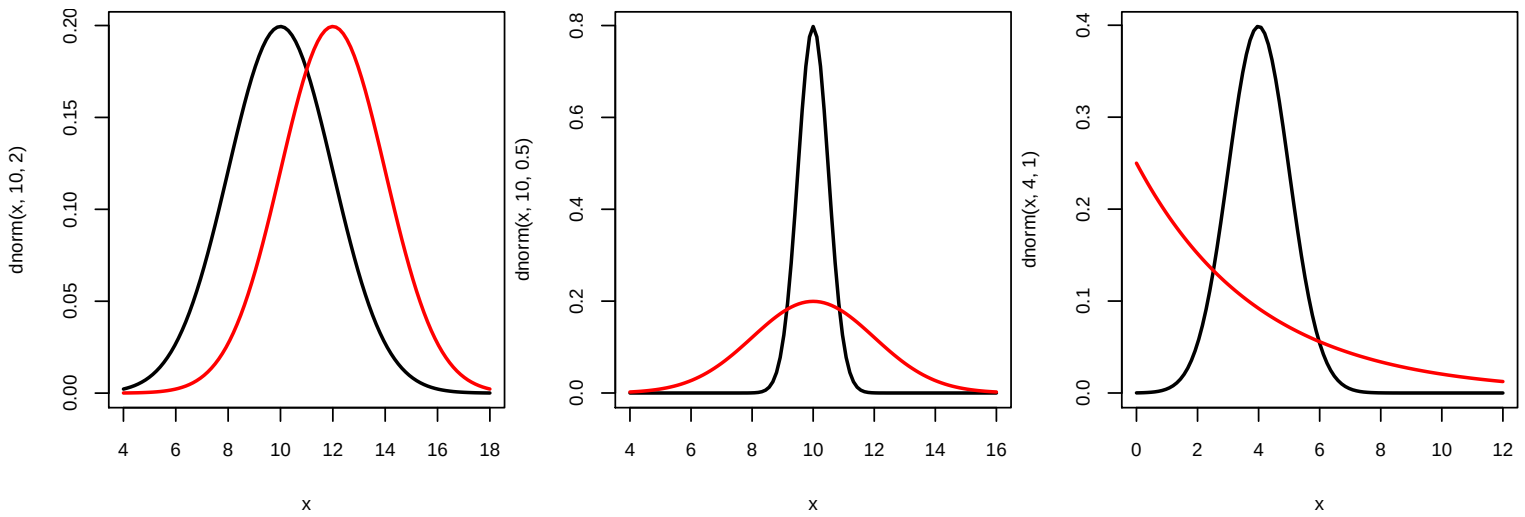
Recall that an association is a general term to describe a relationship between two variables.

> Two quantities are associated when, for whatever reason, knowing the value of one quantity tells you *something* about (i.e. narrows down) the possible values of the other.

**In this discussion, we will always assign the role of $x$ to be the categorical variable and $y$ to be the quantitative variable**. An association exists between $y$ and $x$ if the distribution of $y$ varies between levels of $x$.

# General Illustration of Association

When $y$ and $x$ have an association, the distribution of $y$ is somehow different between the levels of $x$.




Left: the two groups differ in terms of their average values of $y$. Middle and right: the two groups have the same average, but overall distributions of $y$ are different.

# Practical definition of association

Checking whether the *distribution* of $y$ is the same for each level of $x$ is a hard problem. Since people are usually more interested in whether the *typical value* of $y$ is different between levels, we will compare only the **average** or **median** values (whichever is appropriate for the distribution).

- Is the *average* friendship potential different between smilers/non-smilers?
- Is the *median* donation amount among all majors the same?

# Practical definition of association

In this course, we will only **test** whether the difference in *typical* values of $y$ (rather than the whole of its distribution) between levels of $x$ is statistically significant.

Side-by-side boxplots or histograms give an informal assessment of whether the *distribution* of $y$ as a whole varies between levels of $x$, but we will not formally test if this difference is significant.

# Which typical value?

What number best summarizes the typical value of $y$?

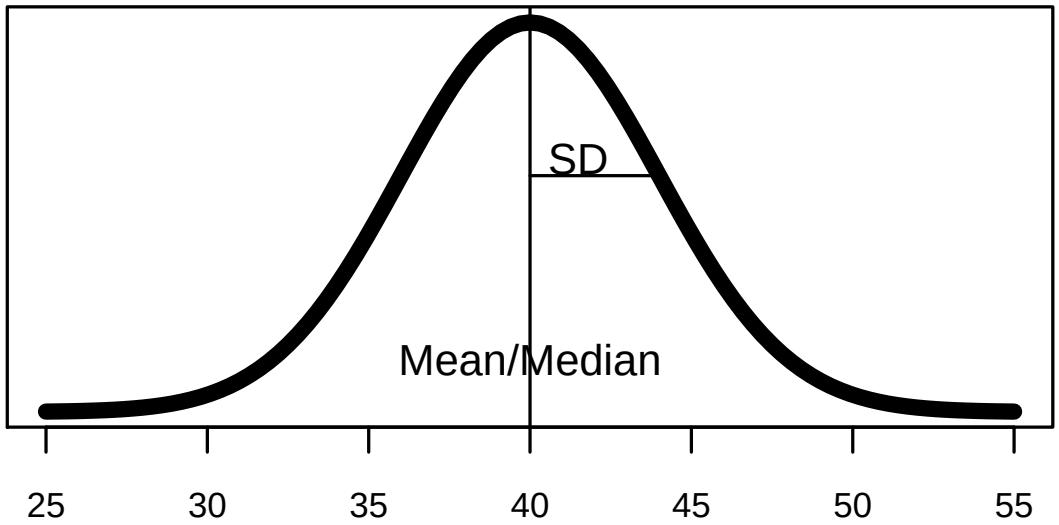
> Use average (mean) when: distributions are roughly symmetric with no extreme outliers.
>
> Use median when: distributions are noticeably skewed and/or have extreme outliers.

The resemblance of the distribution of $y$ to a Normal distribution provides an excellent reference point for deciding which of these quantities to use.

# Normal distribution

The **Normal distribution** is the standard "bell-curve" that provides an approximate shape for many distributions. It is symmetric and unimodal.

- The mean/median (typical value) are both located at the peak of the distribution.
- The standard deviation or SD (typical difference between values and the mean) is the "half-width" of the curve.



# Comparing to a Normal distribution

If the Normal distribution provides a good description of the shape of the distribution, then the average/mean provides a good summary of the typical value of the distribution. If not, the median provides a better summary.

The histogram provides an informal assessment of how much it looks like a Normal distribution, but the **QQ plot** is specialized to do this.

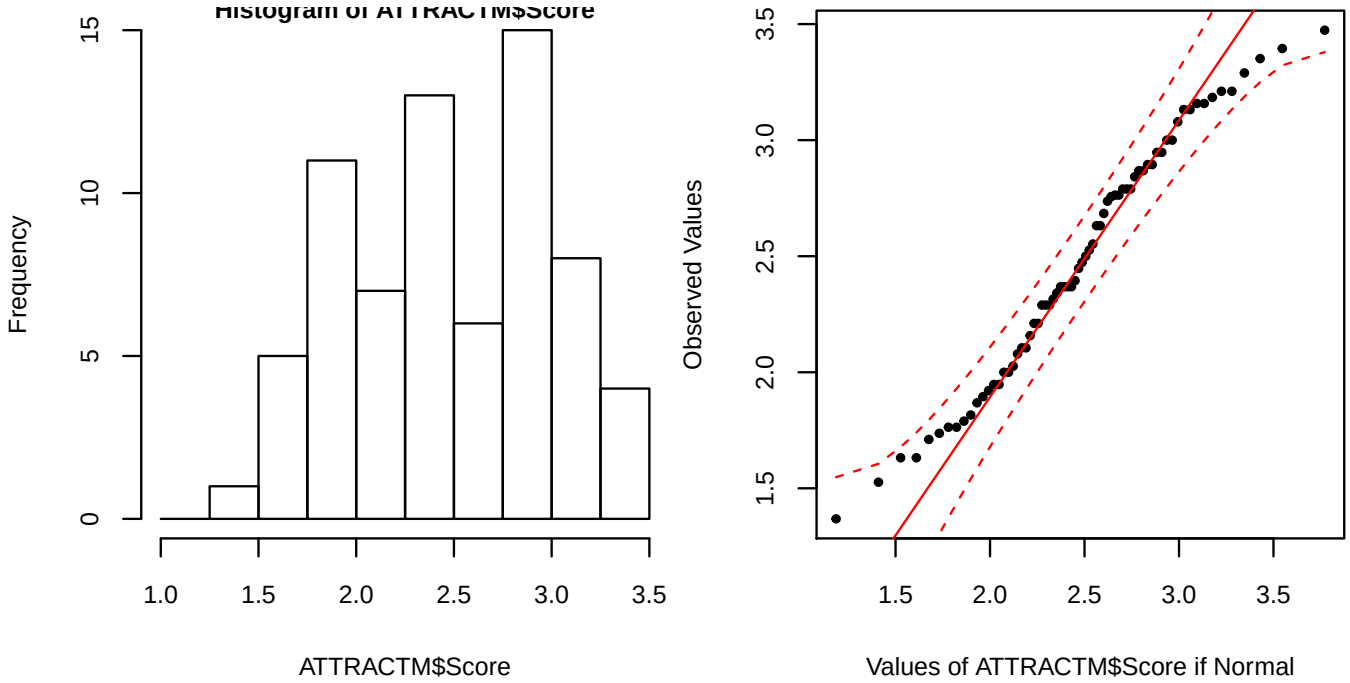
# QQ plot overview

Our version of the QQ plot (and there are others) will compared the observed values in the data to the values we would have expected them to be had been generated at random from a Normal distribution.

If the observed and expected values match up well, then the distribution is well-described by a Normal distribution, and the average/mean provides a useful summary for the typical value. If not, the median provides a better summary.

# QQ plot example1 (very close to Normal)

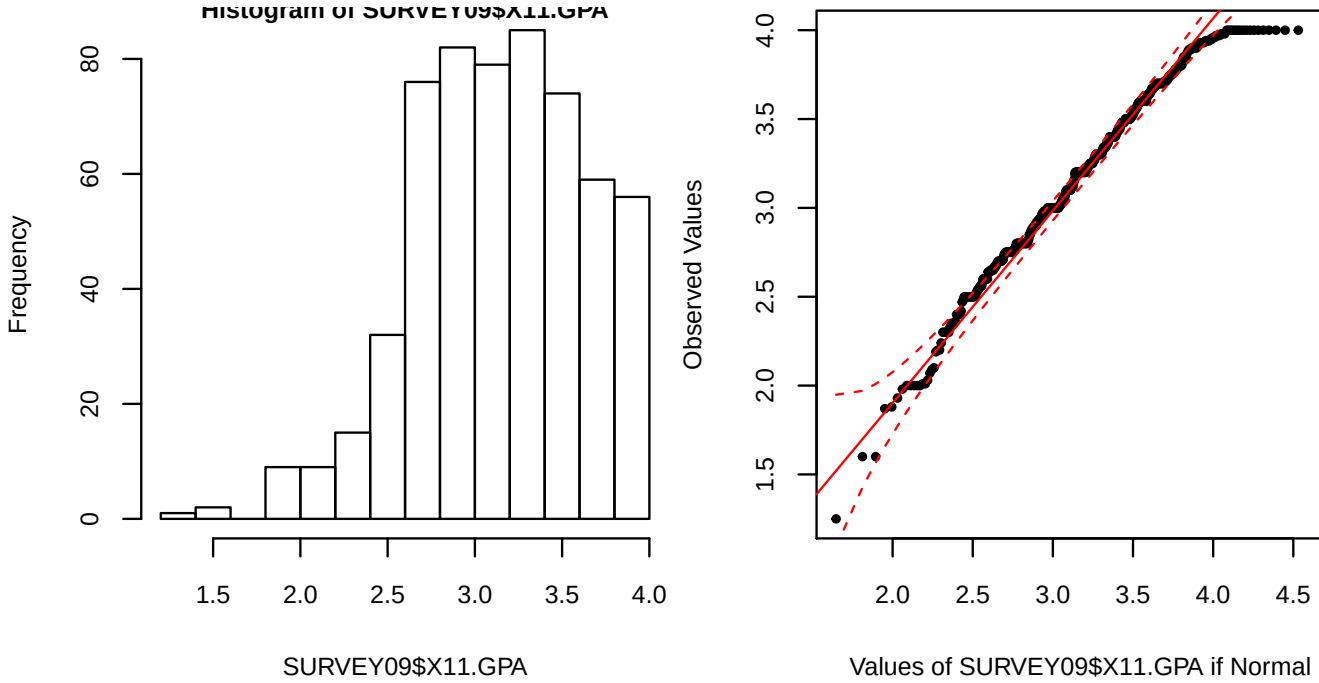
```
hist(ATTRACTM$Score,breaks=seq(1,3.5,.25)); qq(ATTRACTM$Score)
```




About as close to a Normal distribution you'll find, though the histogram doesn't have an "obvious" bell-shape.

# QQ plot example2 (small but systematic difference from Normal)

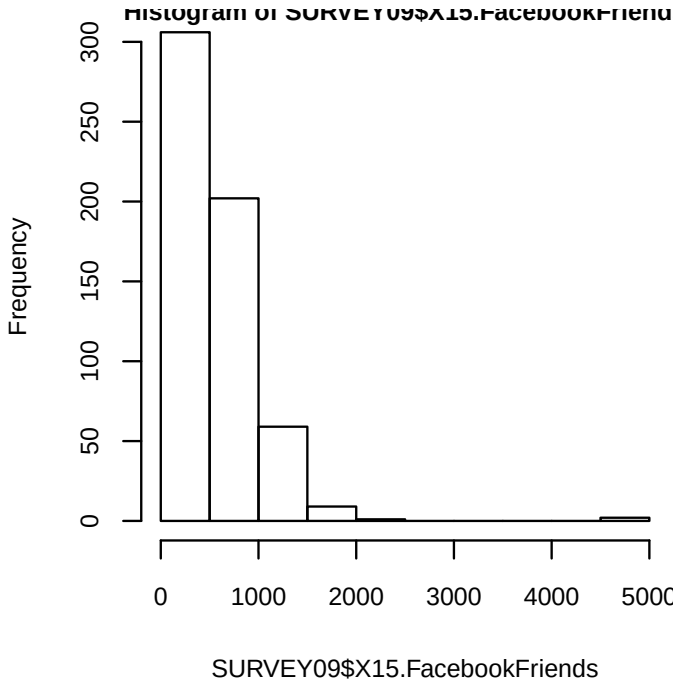
```
hist(SURVEY09$X11.GPA); qq(SURVEY09$X11.GPA)
```

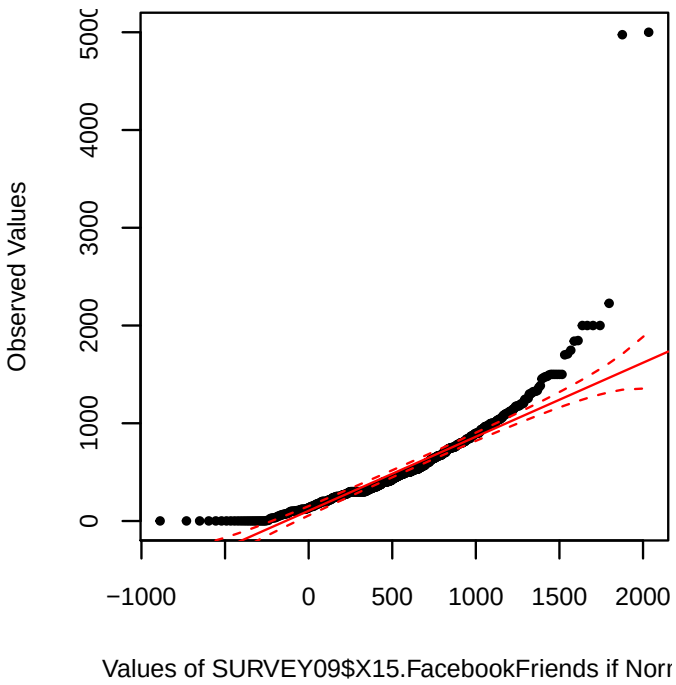



Decent match except for at large GPAs. Some systematic departure from Normality.

# QQ plot example3 (not Normal)

```
hist(SURVEY09$X15.FacebookFriends); qq(SURVEY09$X15.FacebookFriends)
```

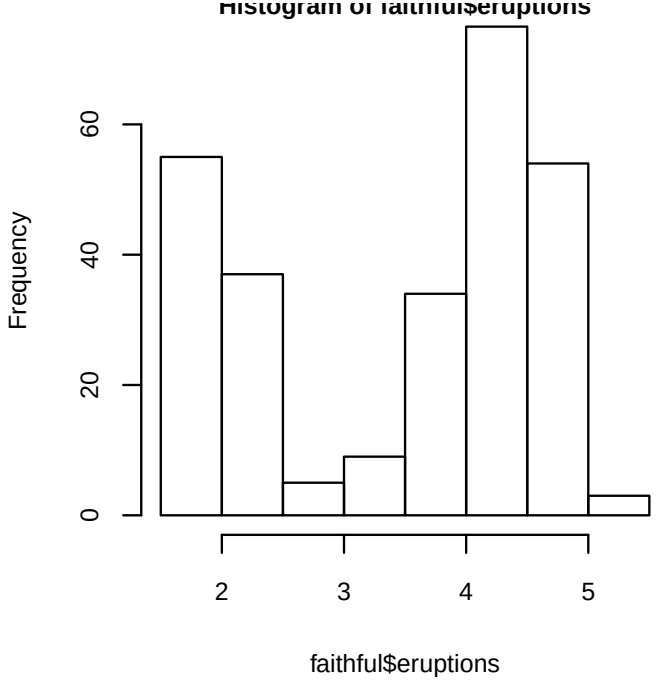


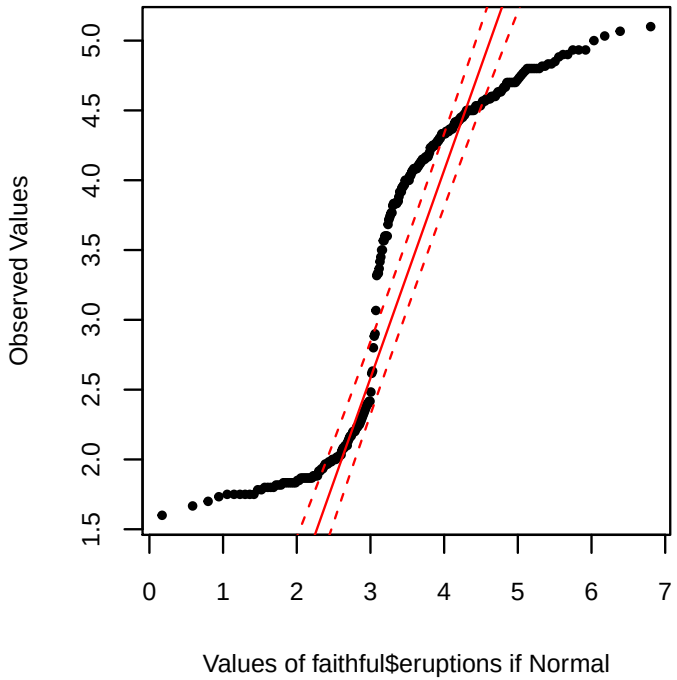




Bad match, clearly not Normal.

# QQ plot example 4 (not Normal)

```
data(faithful)
hist(faithful$eruptions); qq(faithful$eruptions)
```




Bad match due to bimodality.

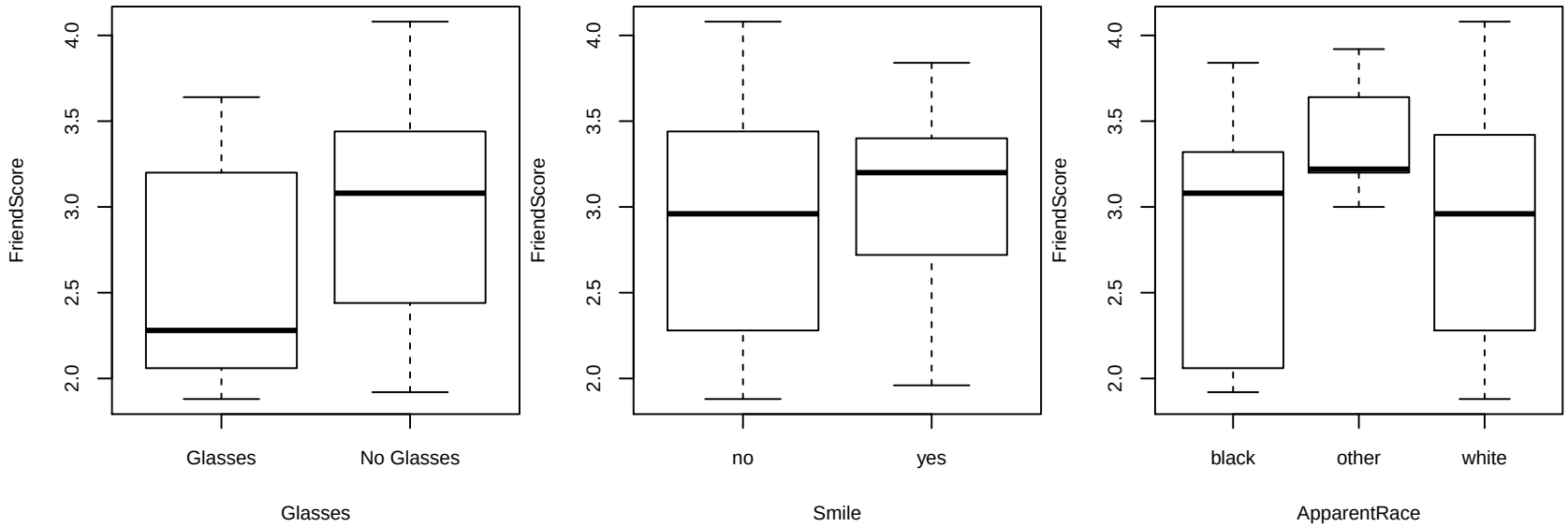
# Determining Normality from QQ plots

If a Normal distribution is a good approximation for the data, then the observed data values and expected (had the data been Normal) values should closely match up and the points should fall near the diagonal red line. There is some leeway, so the distribution can be considered approximately Normal if:

- There is no systematic global curvature of the stream of points away from the diagonal red line.
- Almost all the points fall within the upper and lower dotted red bands. It is ok if a few points at the outskirts fall outside by a little.
- No points are extremely far from the upper/lower dotted bands (these are outliers, which ruin the value of the mean).

# Side-by-side boxplots

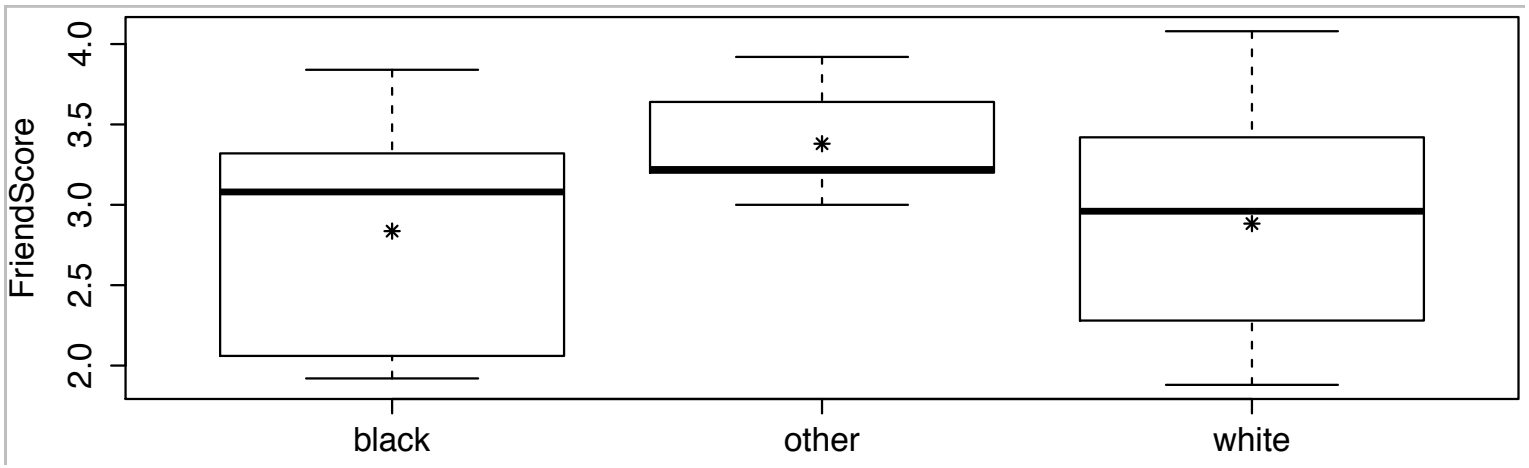
Recall that side-by-side boxplots are convenient ways of showing the overall distribution of $y$ for two or more levels of $x$. The *shape* of the overall distribution remains hidden, but the median values are represented by the bar through the box.



# Boxplots in associate

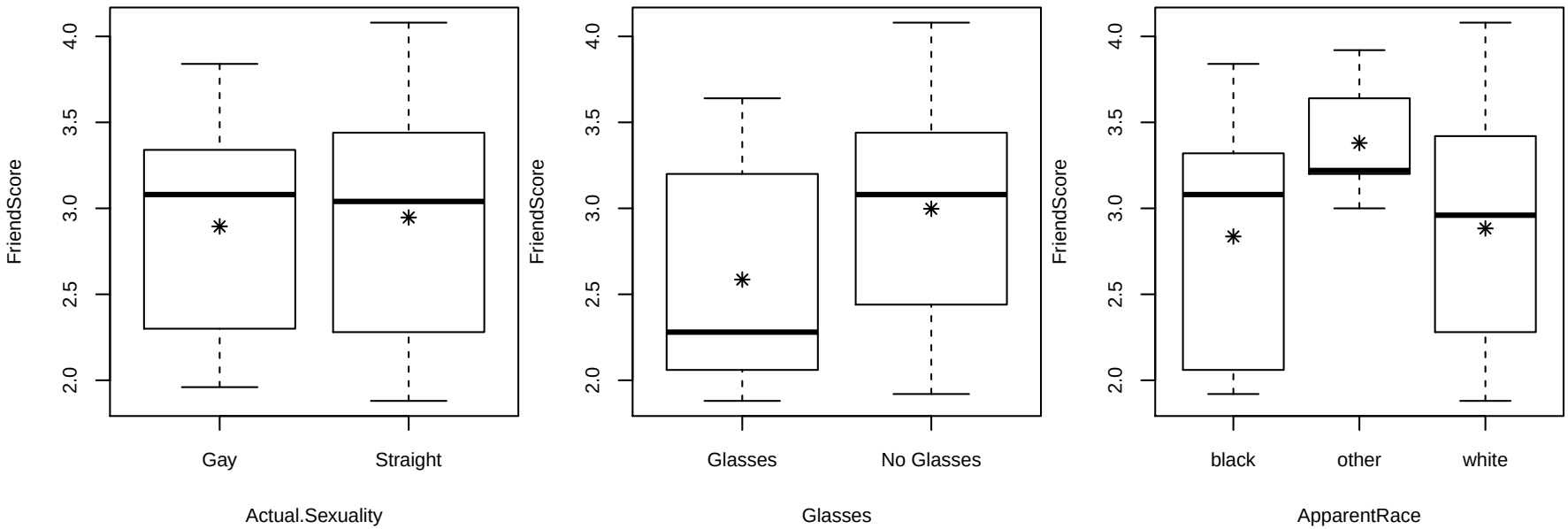
R's default presentation of boxplots is somewhat lacking because the average is not displayed. We will be analyzing association using the associate command in package regclass, which does show the averages for easy comparison.

```
associate(FriendScore~ApparentRace,data=FFRIEND)
```



# What to look for in the boxplots

Look to see if the means and/or medians noticeably differ between levels of $x$.
If they do, we suspect an association.

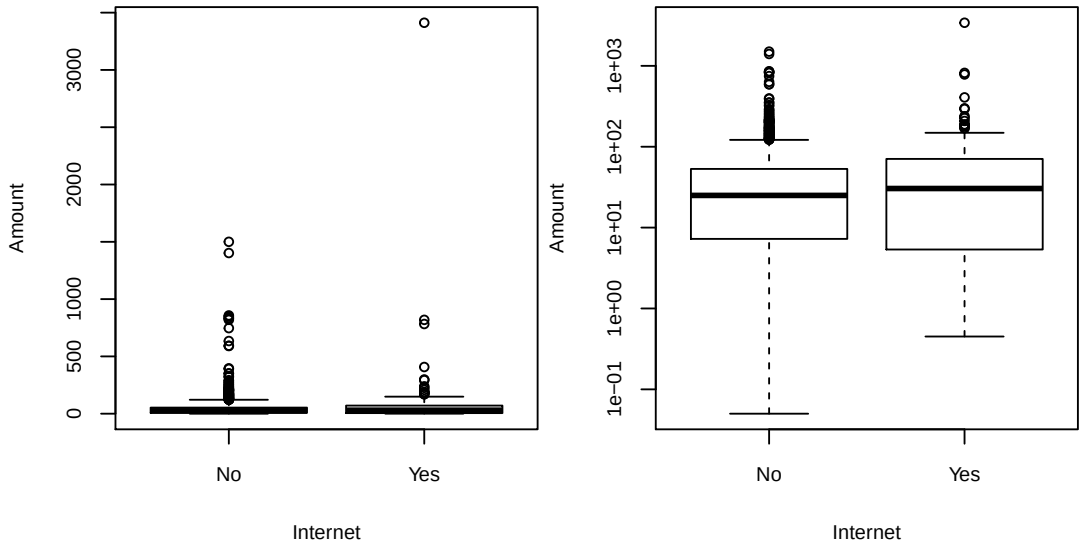



Left - do not suspect association since the means/medians are about the same.
Right/Middle - suspect an association since means/medians look different.

# Reminder: consider logarithmic plots

If the distribution is highly skewed, the boxplots may not be very informative. Try plotting $\log_{10} y$ for each group instead. In the plots below, the potential association between the amount of money a customer spends and whether it is an internet purchase is examined. We do not suspect an association.

```
CUST <- read.csv("Customer37783.dat")
plot(Amount~Internet,data=CUST) #or associate(Amount~Internet,data=CUST)
plot(Amount~Internet,log="y",data=CUST) #or associate(log10(Amount)~Internet,data=CUST)
```



# Philosophy

When we look at side-by-side boxplots, we never expect means or medians to match up *exactly*, even if the distributions of $y$ for each level of $x$ are fundamentally the same. Since the data represent a sample from a much larger population, we expect there to be some variation.

When both variables are categorical, we quantified the difference in the distribution of $y$ between levels of $x$ with the discrepancy between observed and expected counts. Here, we will use the discrepancy in the averages or medians between the levels of $x$.

When comparing averages is appropriate, we compare the *observed variability* in averages between levels of $x$ to the *expected variability* had there been no association (each level of $x$ has the same fundamental underlying average value of $y$). The test is called an **ANOVA** (analysis of variance).

# Test for means: F-statistic

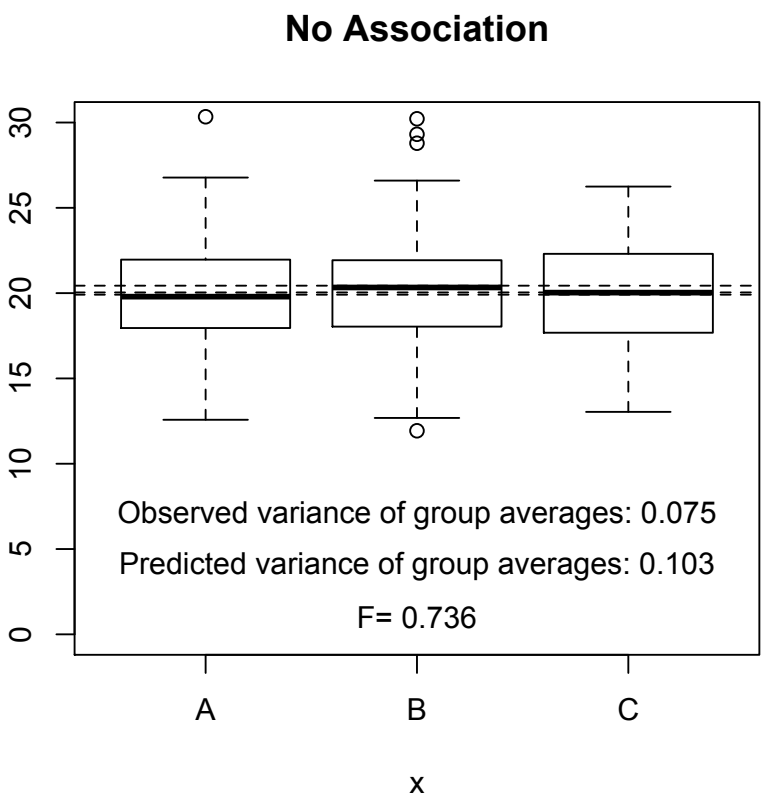
The discrepancy in the average value of $y$ between levels of $x$ is called the $F$ statistic and is given by

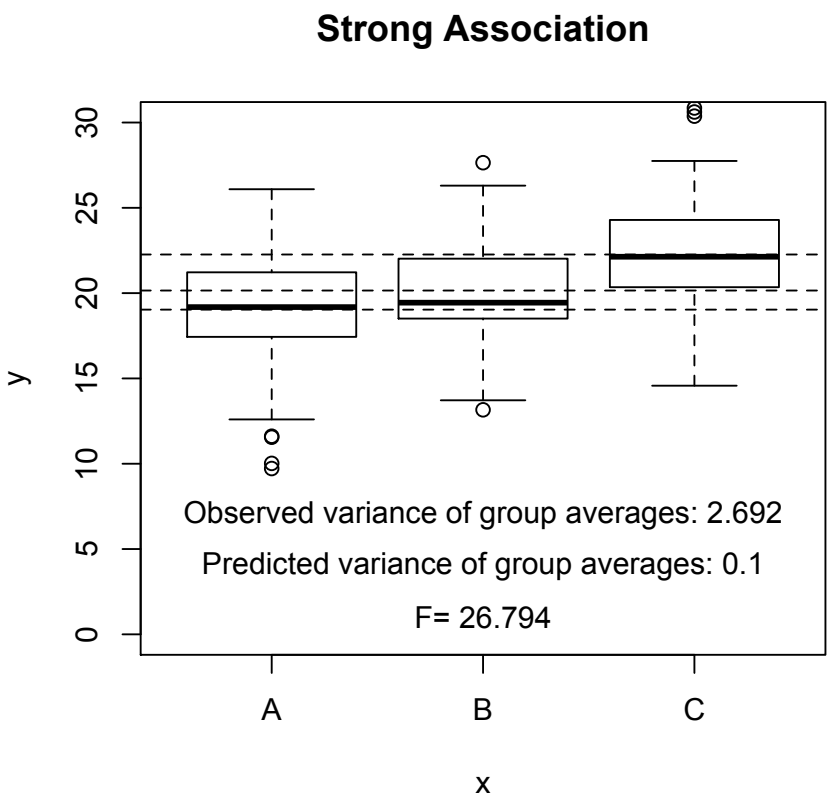
$$F = \frac{\text{Observed variance of group averages}}{\text{Expected variance of group averages if there was no association}}$$

The formula for the "predicted" variance assumes that, in reality, each group comes from identical Normal distributions. You can look up the exact formula for how $F$ is computed online, but software will always output $F$ so that you don't have to.

The formulas are a bit involved, but if there is no association then $F \approx 1$. When there is an association, then the value of $F$ will be "large".

# Illustration of ANOVA concepts



# Interpreting the value of F

The value of $F$ tells us something interesting about the association.

- If $F = 1$, then the observed variability in the average values of $y$ between levels of $x$ is exactly what you would expect had all levels had the same fundamental underlying average value of $y$.
- If $F = 5$, then the observed variability in the average values of $y$ between levels of $x$ is *five times higher* than what you would expect had all levels had the same fundamental underlying average value of $y$.

The question is: how large does $F$ need to be in order to convince that an association exists between $y$ and $x$, i.e., not all levels of $x$ have the same fundamental underlying average?

# Permutation procedure for finding "chance" values of $F$

Sampling variability alone will give values of $F > 1$ even when all levels have the same fundamental underlying average. We can find how large $F$ can get by chance using the permutation procedure.

- Create an artificial **permutation** dataset where the observed values of $y$ and $x$ are *randomly* paired together. By design, there is no association between these variables and all levels of $x$ fundamentally have the same underlying average.
- Calculate $F$ for this permutation sample and record it.
- Repeat this process a "lot" of times and make a histogram of the values of $F$ to see what happens "by chance" when $x$ and $y$ have no association.

# Permutation Dataset

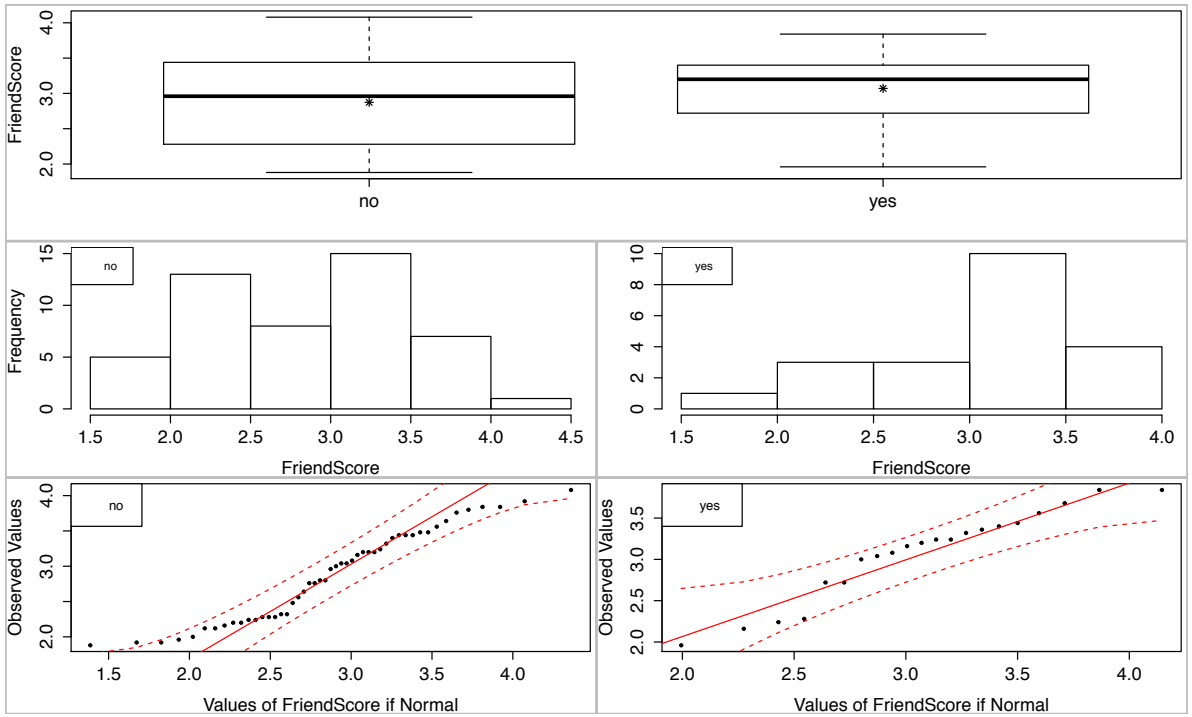
The **permutation procedure** is most easily illustrated when $x$ and $y$ are both numerical as we have seen. Below is a table of the original observed values and of three permutation datasets where the values of $x$ and $y$ have been randomly paired together (more specifically, the values of $y$ are shuffled up and assigned at random to each individual).

| Individual | Observed Data | | Permutation 1 | | Permutation 2 | | Permutation 3 | |
|---|---|---|---|---|---|---|---|---|
| | *Rent* | *Spend* | *Rent* | *Spend* | *Rent* | *Spend* | *Rent* | *Spend* |
| 1 | Yes | 0.2 | Yes | 3.4 | Yes | 5.6 | Yes | 3.4 |
| 2 | Yes | 3.4 | Yes | 6.0 | Yes | 3.4 | Yes | 0.2 |
| 3 | No | 3.4 | No | 5.6 | No | 6.0 | No | 5.6 |
| 4 | No | 5.6 | No | 3.4 | Yes | 3.4 | Yes | 3.4 |
| 5 | Yes | 6.0 | Yes | 0.2 | No | 0.2 | No | 6.0 |

# Comparing means example

Is there an association between smiling and friendship potential?

```
associate(FriendScore~Smile,data=FFRIEND)
```



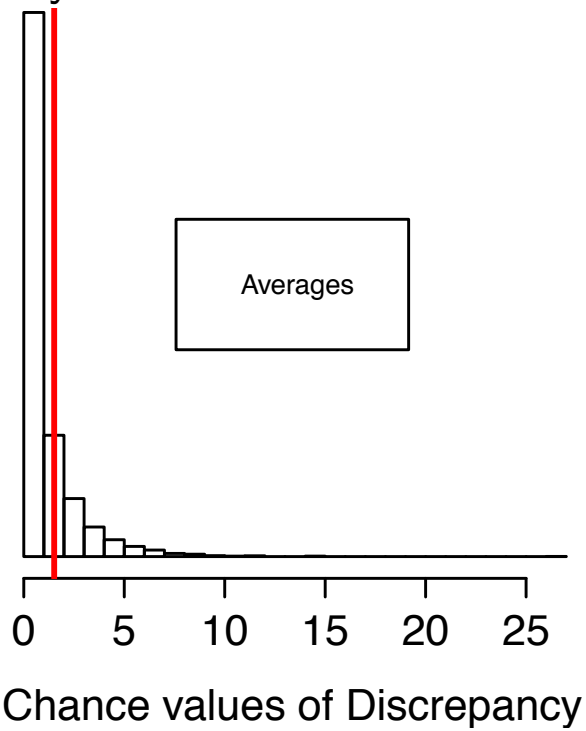
# Comparing means example

| | no | yes | Discrepancy | Estimated p-value |
|---|---|---|---|---|
| Averages (ANOVA) | 2.873 | 3.07 | 1.514 | 0.2172 |
| Mean Ranks (Kruskal) | 39.59 | 25.95 | 1.303 | 0.2526 |
| Medians | 2.96 | 3.2 | 1.572 | 0.1926 |

- The average is appropriate to summarize the distribution (some points outside the bands in the QQ plot at the edges, but this is ok).
- There is a noticeable difference in averages (2.87 vs. 3.07).
- The value of $F$, the discrepancy in averages, is 1.5. So the variability in averages in levels is 1.5 higher than what we'd expect by chance.
- What values of $F$ appear by chance?

# Comparing means example

Creating 500 permutations, we find that such a discrepancy in averages occurs fairly often by chance, so maybe there is not an association after all.



# Median test

When the distribution of $y$ is not well-described by a Normal distribution (there may be outliers or skewness), the average does not do a good at summarizing the distribution. Thus, the preceeding ANOVA test to compare averages is not a good idea.

> When at least one of the distributions has extreme outliers or is skewed, we will compare the **medians**.

```
x <- c(3,6,8,9,12)     #Nice symmetric distribution; mean summarizes data
mean(x)
## [1] 7.6
x <- c(3,6,8,9,120)    #Very skewed with an extreme outlier (typo); mean doesn't summarize c
mean(x)
## [1] 29.2
```

# Median test

The median test is the most general test and requires no assumptions about the underlying distributions!

The test comes up with a clever way to measure the variability in the median value of $y$ between each level of $x$, then it uses the permutation procedure to determine whether such variation is explainable by chance or if there is evidence of an association.

# Median test mechanics

- When there is no association between $y$ and $x$, the fraction of individual values above and below the overall median should be about the same.
- If individuals in one group tend to have more values above (or below) the median than expected, this implies that there is an association between $y$ and $x$ (know the group tells you something about the distribution of $y$).
- Median tests determines if there is an association between $x$ and $y =$ "is value above median?"

# Illustration of median test

Original data:

| | Values | | | | | | | Larger than median (of 3.1) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| White | 3.1 | 2.1 | 3.7 | 1.8 | 2.3 | 1.9 | | no | no | yes | no | no | no | |
| Black | 3.7 | 3.3 | 4.1 | 3.4 | | | | yes | yes | yes | yes | | | |
| Other | 4.1 | 1.9 | 3.0 | 1.7 | 4.2 | 3.6 | 1.6 | yes | no | no | no | yes | yes | no |

Converted into a contingency table:

| | Larger than median | |
|---|---|---|
| Race | Yes | No |
| White | 1 | 5 |
| Black | 4 | 0 |
| Other | 3 | 4 |

Here, 5/6 of values for "white" are below the median while 0/4 of values for "black" are below the median, indicating a possible association. However, the test comes back with a $p$-value greater than 5%, indicating no statistically significant association.

# *p*-values

When `associate()` is run, output that calculates and compares the means and medians (you can ignore the middle row about Mean Ranks) is displayed along with the *p*-value of the differences in means/medians (found via the permutation procedure).

> The *p*-value tells us the probability of finding at least as big a discrepancy in averages/medians "by chance" when each level of $x$ has the same fundamental underlying average/median, i.e., when $x$ and $y$ are independent.

The *p*-value is estimated to be the fraction of permutation datasets (where no association exists) that produced a discrepancy in averages/medians at least as large as the discrepancy in the original data.

# *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.

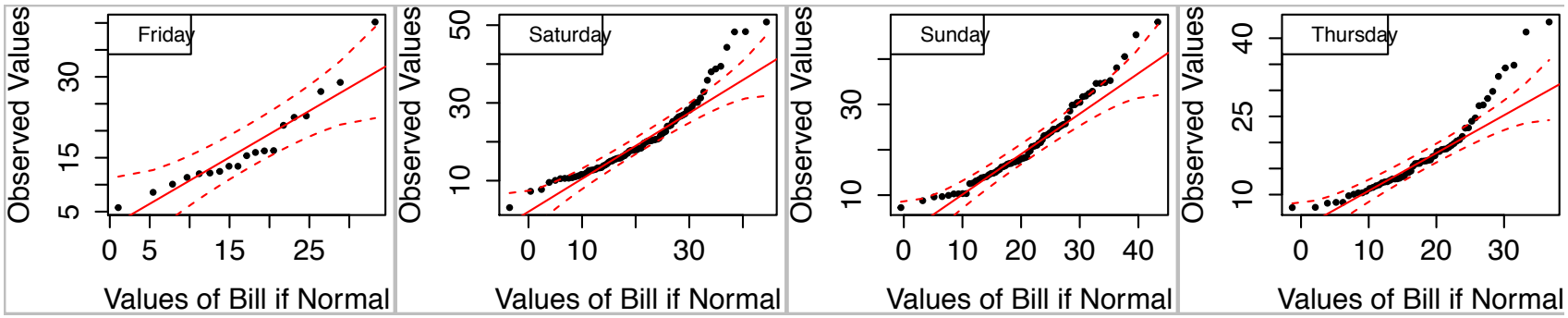
# Final example: Bill vs. Weekday

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.

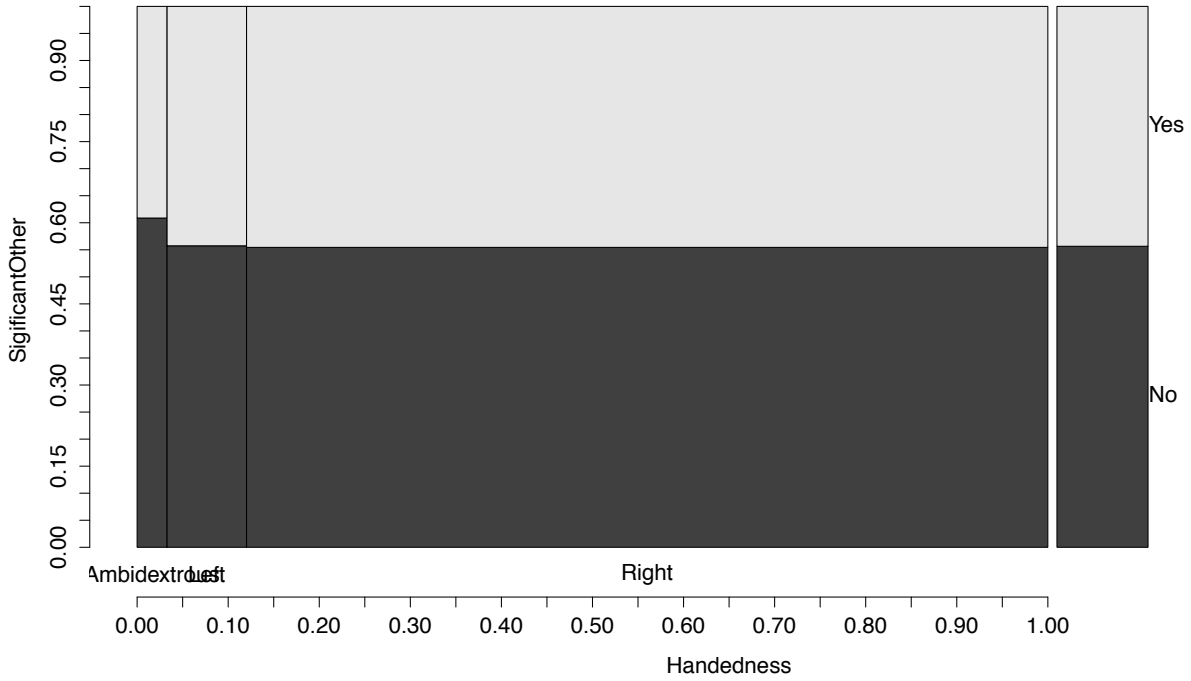
```
associate(y~x,data=...,permutations=500,seed=...)
```

- `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)

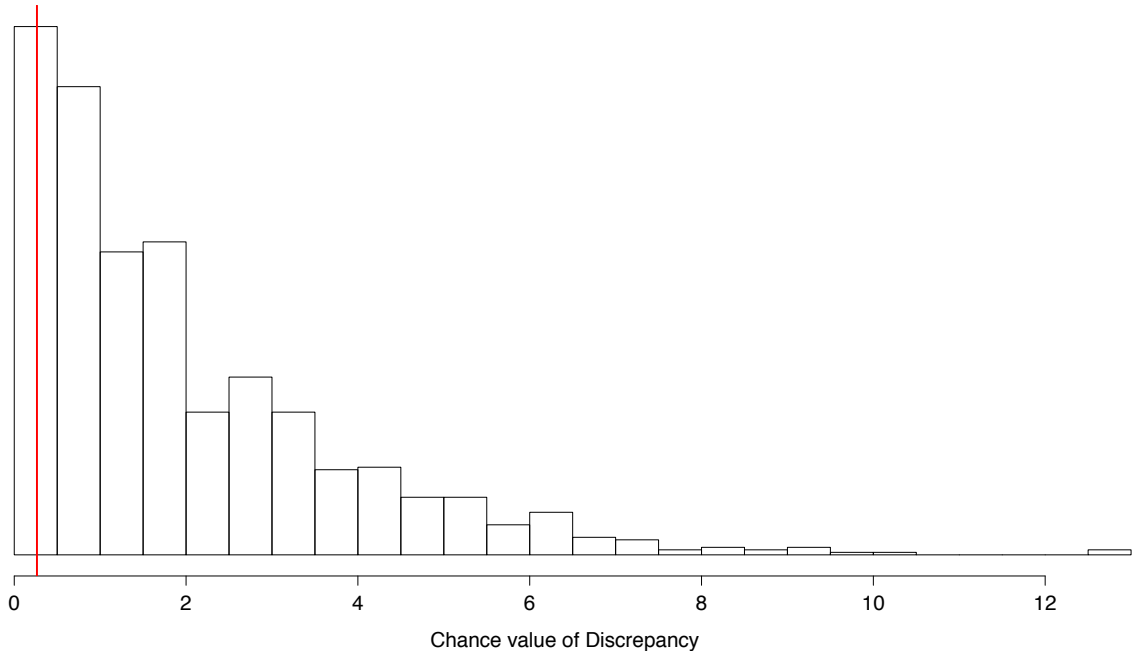
```
library(regclass) #need to load up regclass to use associate
data(SURVEY10) #loads up this dataset built-in to regclass
associate(SigificantOther~Handedness,data=SURVEY10,permutations=1000)
```

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 $\geq 0.05$ (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).

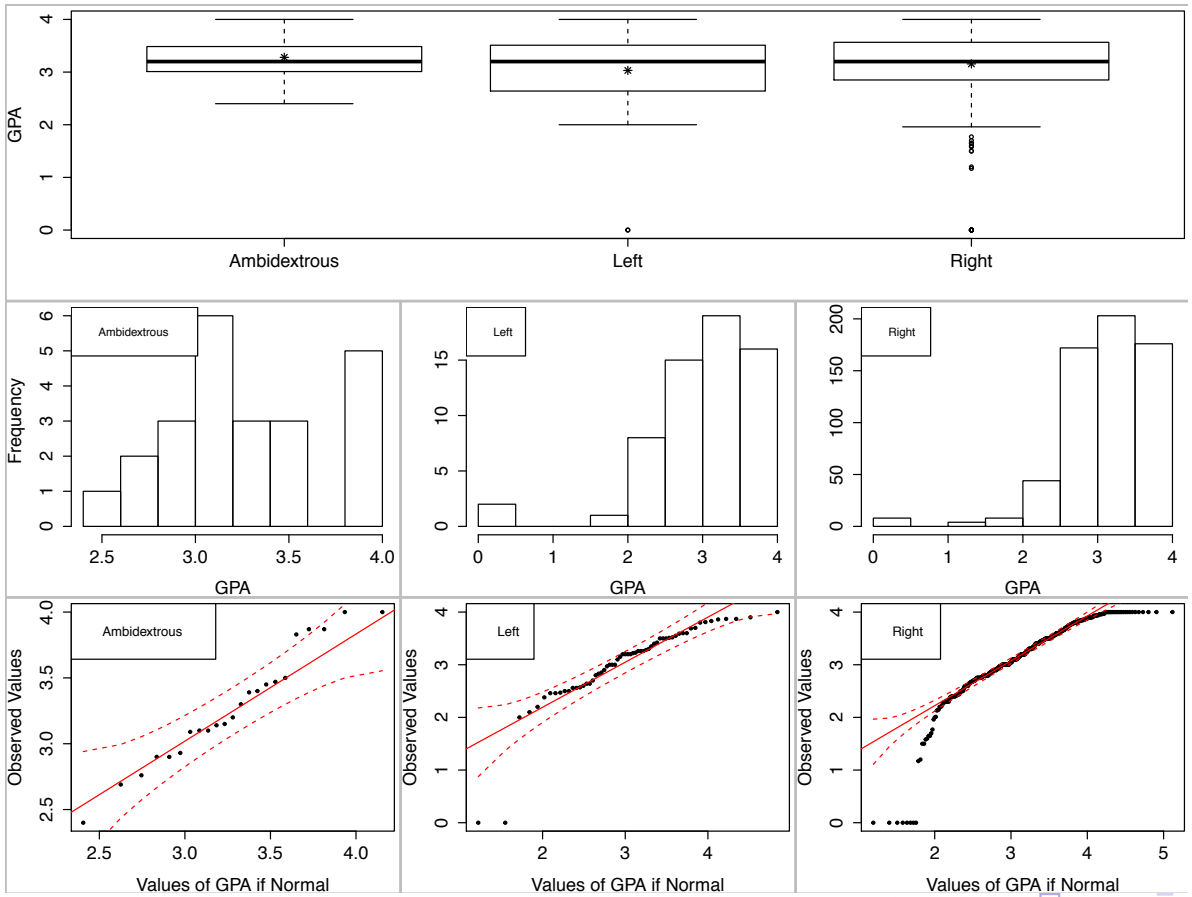
# Running associate (1 categorical and 1 quantitative variable)

```
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!

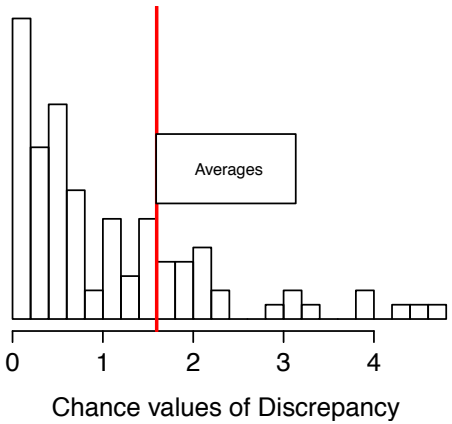
# Running associate (1 categorical and 1 quantitative variable)

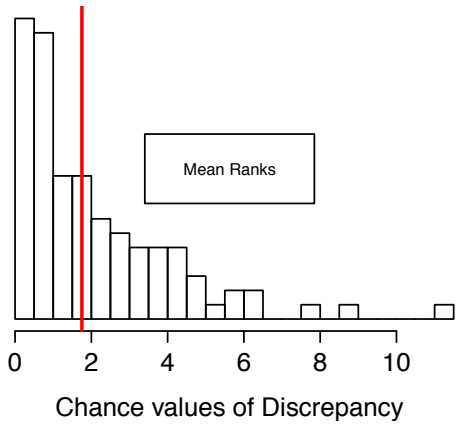
Visually gauge whether there is an association by comparing averages (*'s in the boxplots) if the distributions look approximately Normal in the QQ plots or medians (horizontal bars in boxplots) otherwise.

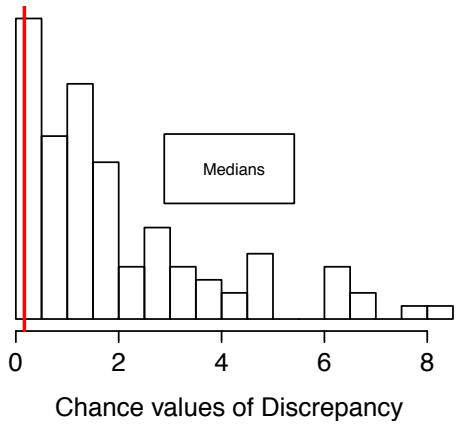


# Running associate (1 categorical and 1 quantitative variable)

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 $\geq 0.05$ (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).