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.

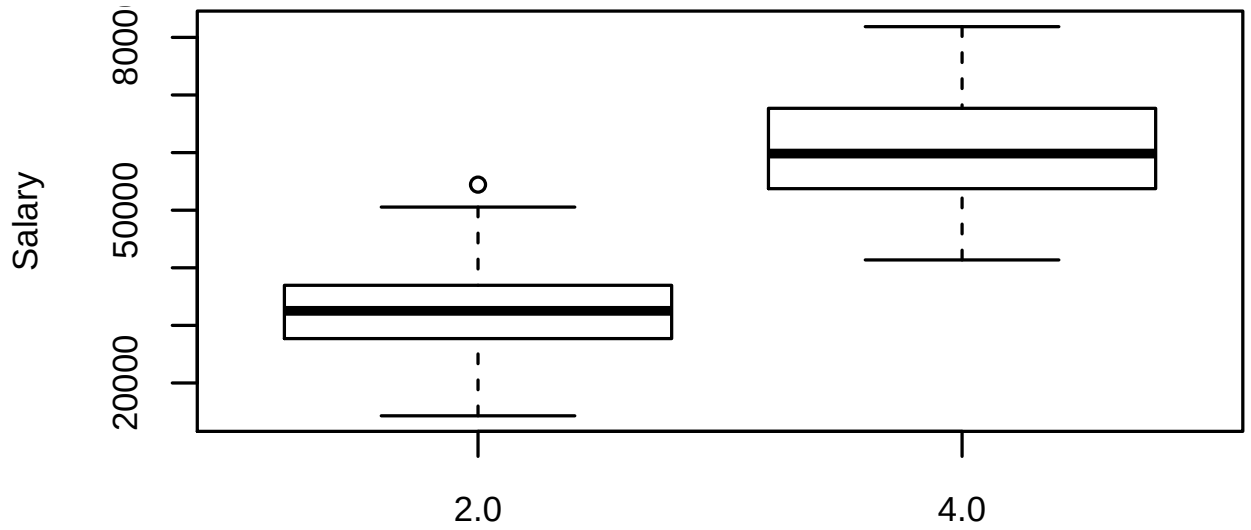
Amount of money spent at The Home Depot and at Kroger

- 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

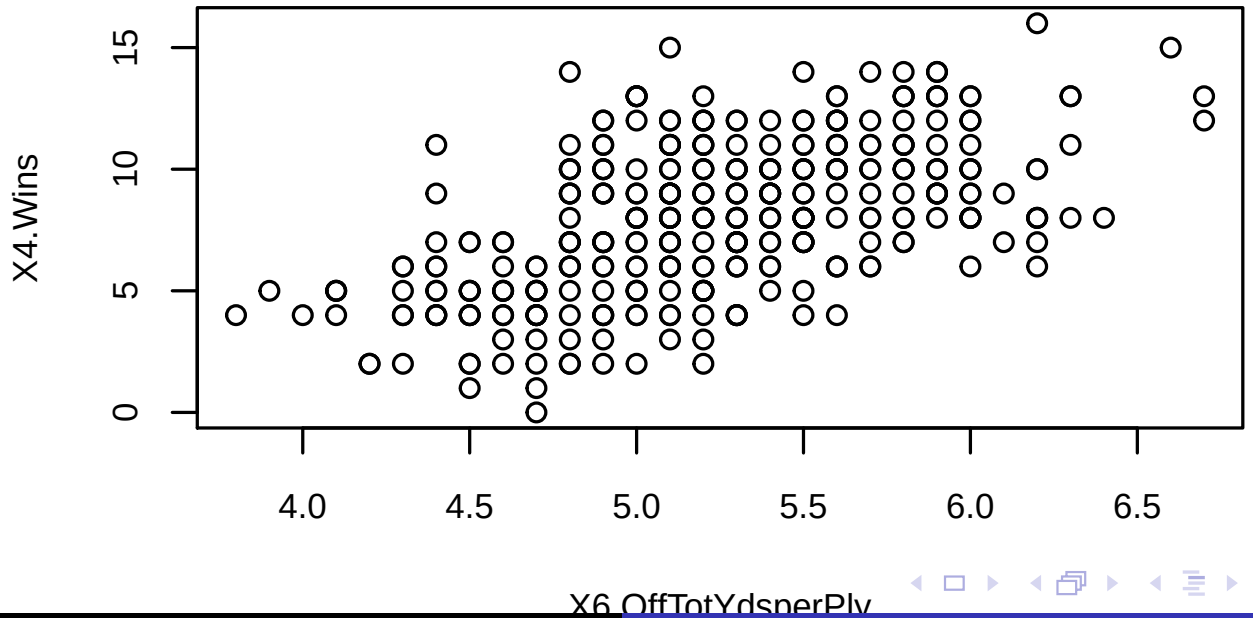
- Starting salaries of UT undergrads may range from \$15,000-\$80,000.
- Among students with 4.0s, salaries may typically range from \$40,000-\$80,000. Among students with 2.0s, salaries may range from \$15,000-\$50,000
- Knowing GPA narrows down the range of salaries so the two quantities have an association.



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 (≈ 4) the number of wins typically varies between 3-7. For teams with large values of yards (≈ 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.

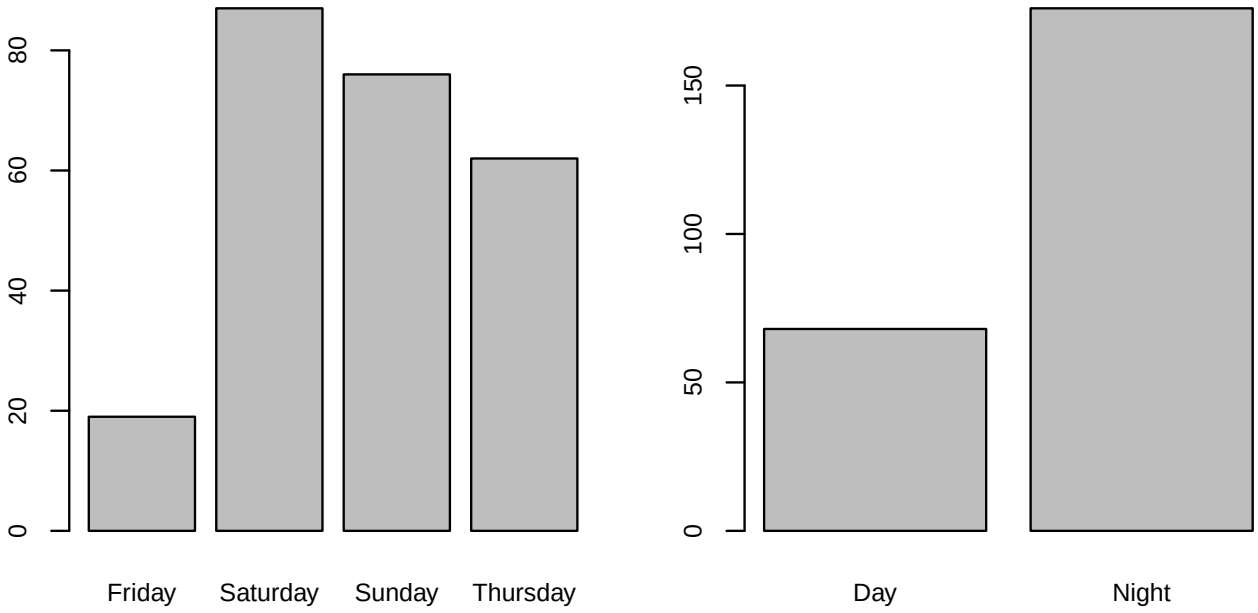
In R, you can make a relative frequency table by making a frequency table and dividing by the sample size (number of entries in the dataframe) by using the command `nrow`.

```
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)
```



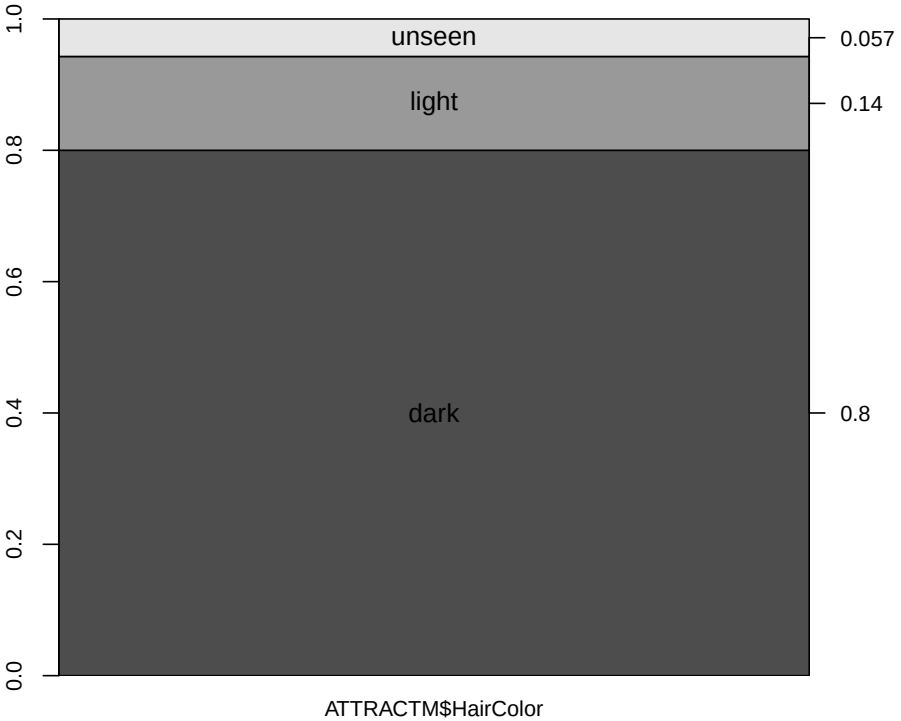
Segmented Bar Chart and Mosaic Plot

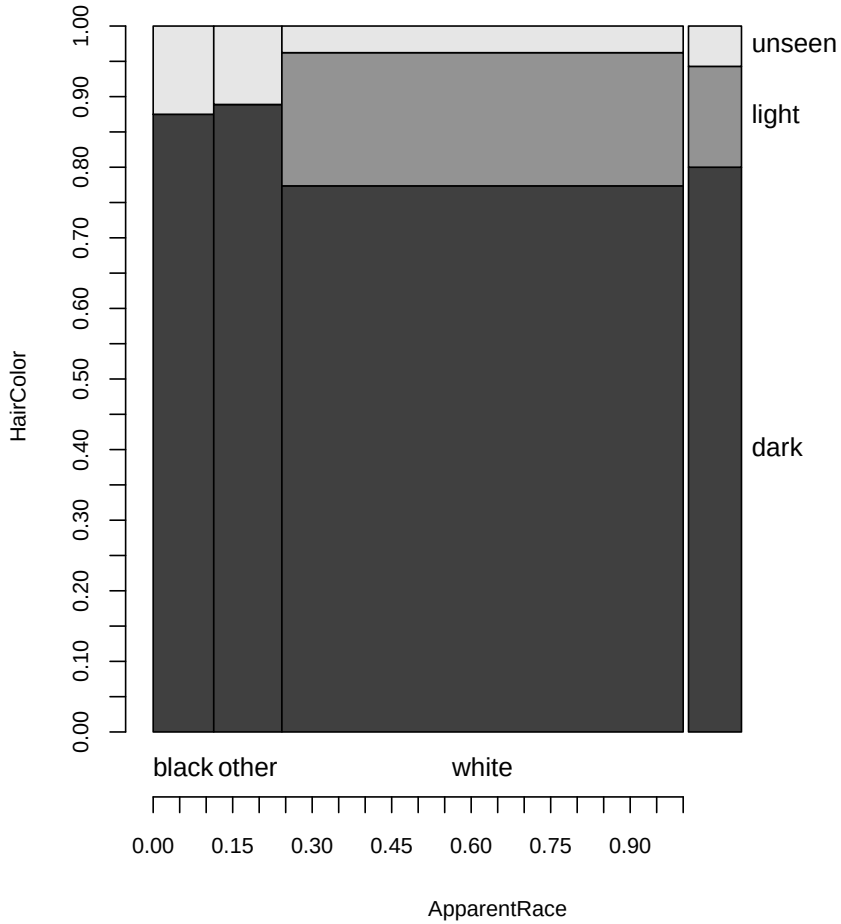
A **segmented bar chart** is another simple way to visualize the distribution of each level. The fraction of the bar devoted to a level is the percentage of individuals who have that level.

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.

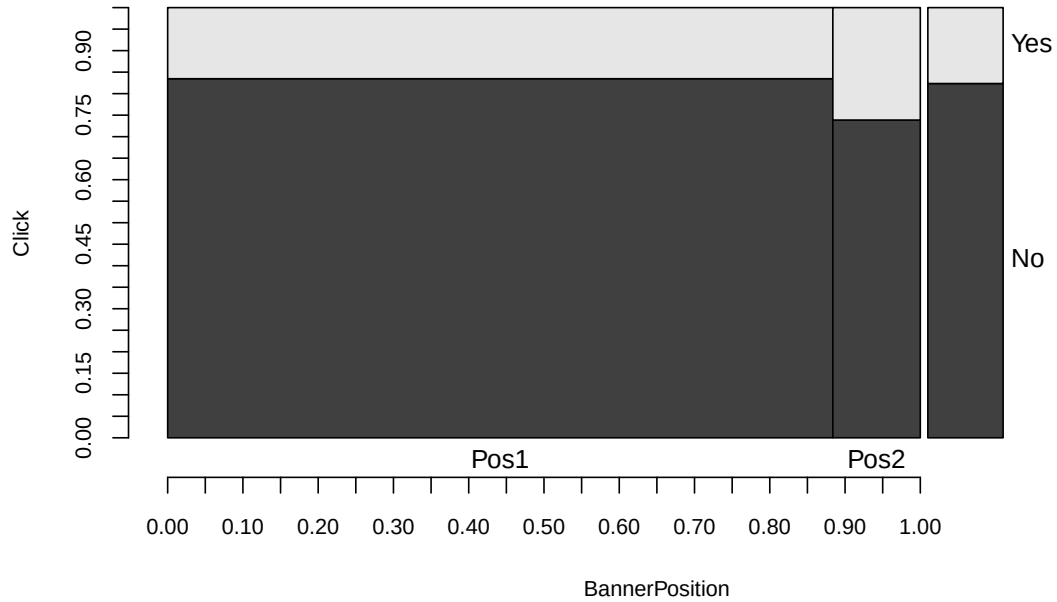
Segmented Bar Chart and Mosaic Plot





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.