Review: The Theory of Causation

- In order to identify the causal effect of X on Y, treatment and control groups must be comparable (~ internal validity)
  - in experiments: this is true if randomization of treatment assignment works
  - in observational studies: this is true if we satisfactorily deal with all of the potential confounders, either through subclassification or through statistical control
- In order to be able to generalize our findings to the population outside of our sample, our sample must be representative of the population (~ external validity)
  - in experiments: this might be hard unless we can randomly select observations from the population
  - in observational studies: this is easier, since we usually observe the entire population

Review: The Usage of Data for Causation

So far we have seen:
1. how to calculate the difference-in-means estimator, which is our estimate of the effect of X on Y if treatment and control groups are comparable
2. how to determine whether the effect of X on Y is small or large in substantive terms (e.g., put it in terms of the sd of the outcome variable)
3. how to check whether treatment and control groups are comparable based on pre-treatment variables (also a good way to check whether randomization worked)

Later in the semester, we will also see:
4. how to determine whether the effect is statistically significant (i.e. distinguishable from zero)

Data Analysis

Data and statistical models can be used for:

1. causation – i.e., answer ‘what is the effect of X on Y?’
   - correlation between X and Y is NOT good enough
   - need to make sure no confounding bias exists; that treatment and control groups are comparable
   - e.g., what is the effect of class size on test scores?
2. prediction – i.e., answer ‘given the value of X, what is our best guess about the value of Y?’
   - correlation between X and Y is good enough
   - usually has a temporal dimension: want to predict future, knowing the past
   - e.g., who will win the presidential election?

Example: Using Polls to Predict 2008 Election

- We will use the latest polls conducted before the election in every state to predict the election outcomes in each state
- First we load the data and look at it (as always)

```
# load data and show first observations
polls <- read.csv("polls.csv")
head(polls)
```

```
## state Obama McCain date latest
## 1 AK 39 58 10/29/08 1
## 2 AK 41 57 10/28/08 0
## 3 AK 42 53 10/18/08 0
## 4 AK 38 57 10/15/08 0
## 5 AK 40 55 10/6/08 0
## 6 AK 38 55 10/4/08 0
```

- Notice that we have several polls for each state. Let’s try to understand the data some more...
How many observations and variables are in the data?

```r
dim(polls)
## [1] 1332  5
```
- `dim()` is the function that provides the dimensions of the object inside the parens: rows, columns
- so in this case, we have how many observations?
- and how many variables?

Each observation represents a poll. How many polls are there in the data set per state?

```r
table(polls$state)
##          AK          AL          AR          AZ          CA          CO          CT          DC          DE          FL          GA          HI          IA          ID          IL          IN          KS          K
##       20         24          14         23          31         46          15          2          9         71         30          4         33         7        13         31        18         26         13         21         7         19         39         49         40
##          MS          MT          NC          ND          NE          NH          NJ          NM          NV          NY          OH          OK          OR          PA          RI          SC          SD          TN          TX          UT          VA          VT          WA          WI          WV          WY
##         16         18         49         12          9        41         34         28         36         38         38         80        13         31        76         9        13        10         13         18         11         61         8        34         43         15
```
- `table()` is the function that provides us with the list of possible values/categories of a variable as well as the number of observations that take on that value/fall under that category

What are the states included in the dataset?

```r
# extract unique state names
st.names <- unique(polls$state)
st.names
## Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS K
##          AK          AL          AR          AZ          CA          CO          CT          DC          DE          FL          GA          HI          IA          ID          IL          IN          KS          K
## 20         24          14         23          31         46          15          2          9         71         30          4         33         7        13         31        18         26         13         21         7         19         39         49         40
##          MS          MT          NC          ND          NE          NH          NJ          NM          NV          NY          OH          OK          OR          PA          RI          SC          SD          TN          TX          UT          VA          VT          WA          WI          WV          WY
##         16         18         49         12          9        41         34         28         36         38         38         80        13         31        76         9        13        10         13         18         11         61         8        34         43         15
```
- `unique()` is the function that returns a vector with all of the unique values in the variable inside the parens

How many states are included in the dataset?

```r
# provides the length of a vector
length(st.names)
## [1] 51
```
- `length()` is the function that returns the length of the vector inside the parens

Let’s take a quick look (visually) at the distribution of the variable Obama (which is numeric). What is the best function to visualize a single numeric variable?

```r
hist(polls$Obama)
```

Now, let’s visualize the relationship between Obama and McCain. What is the best function to visualize the relationship between two numeric variables?

```r
plot(x=polls$Obama, y=polls$McCain)
```
Looking at the plot, do you think the correlation between the variables is positive or negative?

We want to base our predictions on the latest polls, so let’s delete all the polls that were not the latest.

```
# subset the data: data[rows, columns]
# only keep observations that meet the test
# keep all columns
polls <- polls[polls$latest==1,]
```

What are the number of observations in the data now?

```
dim(polls)
## [1] 51 5
```

- 51 (number of rows); 5 variables
- makes sense since we have one poll for each state + DC

Now, let’s calculate the margin of Obama over McCain for the latest poll in each state.

```
# creates the variable margin inside polls
polls$margin <- polls$Obama - polls$McCain
```

Let’s look at a few observations to ensure that the new variable was created correctly.

```
# you could use head(polls) or subset the data using [,]
# show first four observations (rows) and all columns
polls[1:4,]
```

```
## state Obama McCain date latest margin
## 1 AK 39 58 10/29/08 1 -19
## 2 AL 36 61 10/27/08 1 -25
## 45 AR 44 51 10/29/08 1 -7
## 59 AZ 46 50 10/29/08 1 -4
```

```
polls$margin captures, then, our predictions of the margin of Obama over McCain in each of the states.
```

How many observations does this dataset have? Do you remember the function that we need?

```
dim(real)
## [1] 51 3
```

- 51 observations: one for each state + DC

Also here, let’s calculate the margin of Obama over McCain in each state.

```
# compute Obama's margin
real$margin <- real$Obama - real$McCain
```

```
real$margin captures the real margin of Obama over McCain in each of the states.
```

### Prediction Error: How good were the predictions?

```
# create vector called errors
errors <- real$margin - polls$margin
names(errors) <- st.names # add state names to vector
mean(errors) # calculate the mean of prediction errors
## [1] 1.078431
```

- average prediction error $\approx 1$ percentage point
  - this is rather small (very close to zero)
  - on average, across all states the poll predictions were *unbiased*
  - positive prediction errors are offset by negative prediction errors on average
What does the distribution of errors look like? Again, what is the function that will show us a visualization of the distribution of a single numeric variable?

\[ \text{hist}(\text{errors}) \]

### Poll Prediction Errors

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Error in Predicted Margin for Obama (in percentage points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-20</td>
</tr>
<tr>
<td>15</td>
<td>-10</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

However, as seen in histogram, some polls were really off by as much as 20 percentage points in either direction.

Standard deviation of the errors gives us a sense of the variation of the errors across polls:

\[ \text{sd}(\text{errors}) \]

## \[1\] 5.833843

### Correlation and Prediction

- High correlation >>> good prediction
- The closer to 1 the \( \text{cor} \) (actual, predicted) is the better the prediction
- What does the relationship between predicted margins and real margins look like? Again, what is the function that will show us the visualization between two numeric variables?

\[ \text{cor}(\text{polls}$\text{margin, real}$\text{margin}) \]

## \[1\] 0.9690625

However, as seen in histogram, some polls were really off by as much as 20 percentage points in either direction.

- [ADVANCED] Do the polls accurately predict the winner in each state?

\[ \text{real}$\text{state}[\text{sign} \left( \text{polls}$\text{margin}) != \text{sign} \left( \text{real}$\text{margin})] \]

## \[1\] IN MD NC

## 51 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI I.
- In all but three states the polls rightly predicted the winner.

- [ADVANCED] How close where the elections in the states where the polls predicted the wrong winner?

\[ \text{real}$\text{margin}[\text{sign} \left( \text{polls}$\text{margin}) != \text{sign} \left( \text{real}$\text{margin})] \]

## \[1\] 1 -1 1
- The margins of victory in all of these three states were of 1 percentage point (very very small)
Today’s Class and Next

Today

➤ Concepts:
   ➤ prediction error
   ➤ relationship between prediction and correlation

➤ Used data for a simple prediction example

➤ New R functions:
   ➤ `dim()`, `table()`, `unique()`, `length()`
   ➤ slicing of data using `[rows,columns]`

Next Class

➤ Regression Analysis!!!
   ➤ please pay close attention to the readings