Plan for Today

- Survey data
  - survey sampling and representation
  - the role of randomization
  - potential complications
- Visualizations:
  - of a single variable: histogram
  - of the relationship between two variables: scatterplot
- Descriptive statistics of relationship between two variables:
  - correlation coefficient

Survey Data

- Data has to come from somewhere
- Probably the most common data source are surveys
- In a survey, relatively few individuals — the sample — are selected to understand the features of a larger population

Sample Must Be Representative

- For the survey data to provide valid conclusions about the target population, the sample needs to be representative
  - a representative sample is a subset of a population that accurately reflects the members of the population

Sample Must Be Representative

- For the survey data to provide valid conclusions about the target population, the sample needs to be representative
  - a representative sample is a subset of a population that accurately reflects the members of the population (its characteristics are identical to those of the population, over repeated samples)
Sample Must Be Representative

- The mathematical laws of probability dictate that if a sufficient number of individuals are chosen at random, their views will tend to be representative of the population.

The Role of Randomization

- Remember: when estimating causal effects, randomization of treatment assignment
  - eliminates confounding bias
  - provides internal validity
  - ensures treatment and control groups are comparable

Randomization Gives Researchers Super Powers

Potential Complications of Survey Data

1. No comprehensive list of individuals in target population
2. Unit non-response: selected individuals refuse to participate in survey
3. Item non-response: respondents refuse to answer a particular question
4. Mis-reporting: respondents lie (e.g., in the case of social desirability bias, respondents choose an answer that is seen as socially desirable regardless of the truth)

If systematic differences between individuals in sampling frame and individuals not in it (complication 1) OR between individuals who agree to respond and individuals who do not (complications 2 and 3) then, survey results will not be representative (i.e., cannot be generalized to the target population).

Let’s Go Back to Playing with Data . . .
Visualizations

- Recall that one of the first things we will always do when analyzing data is understand the variables needed for the analysis (e.g. the treatment and outcome variables)
  - visualizations are a great way of accomplishing this
- In this class, we will only use **histograms** as a way to visualize one variable at a time
- And, we will use **scatterplots** to visualize the relationship between two variables (e.g., X and Y)

### Visualizations of a Single Variable: Histogram

- A histogram plots the frequency or density of the values of numeric variables (i.e., 1,2,3,4)
  - in R, the function `hist()` plots the frequency or the density of the values
    - by default: the function plots the frequency
    - if you want to plot the density so that you can more easily compare distributions with different number of observations, you will need to include as an argument in the function: `freq = FALSE`

To create the histogram of fourth grade reading test scores in the data from the STAR project we can simply run

```r
data <- read.csv("STAR_project.csv") # load data
hist(data$g4reading) # plot histogram with frequencies
```

Now if we want the histogram to display densities instead of frequencies (useful when comparing distributions with different number of observations), we simply add to the function `freq=FALSE` as an argument

```r
hist(data$g4reading, freq=FALSE) # hist with densities
```

### Visualization of the Relationship Between Variables

- To visualize the relationship between two variables, we will use **scatterplots**
  - the function `plot()` in R creates a scatterplot; plots one variable against another in order to visualize relationship
Imagine we have two variables $x$ and $y$

First, let's plot this point: $(x, y) = (4, 2)$

Now, let's plot this point: $(x, y) = (8, 5)$

Finally, let's plot this point: $(x, y) = (11, 7)$.
Scatterplot

Imagine we have two variables:

- $\begin{array}{c|c|c}
4 & 2 \\
8 & 5 \\
11 & 7 \\
\end{array}$

- First, let’s plot this point: $(x,y) = (4,2)$
- Now, let’s plot this point: $(x,y) = (8,5)$
- Finally, let’s plot this point: $(x,y) = (11,7)$

Let’s create the scatterplot, by plotting each point at a time.

Now let’s do the same with R:

```r
# input data
Y <- c(2,5,7)  # c() creates a vector of data
X <- c(4,8,11)
# create the scatterplot:
# notice that the first argument is the X variable
# the second argument is the Y variable
plot(X,Y)
```

Descriptive Statistics of Relationship Between Two Variables: Correlation Coefficient

To summarize the relationship between two variables, we will use the correlation coefficient:

- A way to measure the strength of the (linear) association between two variables (X and Y)
- It does NOT necessarily imply CAUSATION
- It simply represents how, on average, two variables move together relative to their respective means
- The function `cor()` in R calculates the correlation coefficient

Imagine we have the following two variables:

```
X
0 0.2 0.4 0.6 0.8 1.0
Y
0 0.2 0.4 0.6 0.8 1.0
```

```
Cor(X,Y)>0 or Cor(X,Y)<0? Cor(X,Y)<0
```

Cor(X,Y)

We can use the function `cor()` in R to calculate the correlation:

```
cor(x, y)
```

```
# [1] -0.7759617
```

- Is $\text{cor}(x,y) = \text{cor}(y,x)$?

```
cor(y,x)
```

```
# [1] -0.7759617
```

- What is $\text{cor}(x,-y) = ?$ $\text{cor}(x,-y) = -\text{cor}(x,y)$

```
cor(x,-y)
```

```
# [1] 0.7759617
```
\[ \text{Cor}(x, y) < 0 \] \quad \text{Cor}(x, -y) > 0 \]

\[ \text{Cor}(X, Y) \]

```
cor(x, y)
## [1] -0.7759617
```

- Is \( \text{cor}(x, y) = \text{cor}(x, 10y) \)?
```
cor(x, 10*y)
## [1] -0.7759617
```
- Yes, because \( \text{cor}() \) is not sensitive to the units used to measure the variables.

\[ \text{Cor}(X, Y) = 1; \text{perfect positive correlation} \]

\[ \text{Cor}(X, Y) = -1; \text{perfect negative correlation} \]

\[ \text{Cor}(X, Y) = 0; \text{no linear association} \]

Today’s Class and Next

Today
- Concepts: Survey data
  - what are representative samples?
  - how does randomization can help?
  - what can go wrong?
- Back to playing with data:
  - \text{hist}(): histogram of single variable
  - \text{plot}(): scatterplot of two variables
  - \text{cor}(): correlation coefficient between two variables

Next Class
- Bring your computers! Exercise in class