Variables and their measurement

- When we talk about political phenomena and their relationships we are talking about variables.
Claim: “more guns lead to more crime”.

Variables: gun ownership in states, crime rates in states & their relationship.
Variables and their measurement: example

- **Claim:** “more guns lead to more crime”.
- **Variables:** gun ownership in states, crime rates in states & their relationship.

Source: [http://en.wikipedia.org/wiki/Gun_violence_in_the_United_States_by_state](http://en.wikipedia.org/wiki/Gun_violence_in_the_United_States_by_state)
So what is a variable after all?
- It is something (usually a quantity) that can vary among subjects in a population.
Variables types

- **Quantitative and Categorical/Qualitative** – relates to the qualities of the variable.

- **Discrete and Continuous** – relates to qualities of the variable and the statistical methods used to analyze them.
Quantitative variables

- **Quantitative variables** are variables which contain a measure of the magnitude of something on a scale.
  - Hillary Clinton’s vote share in the 2016 election by state.
  - Age of each registered voter in Georgia.
  - Years of education.
Categorical variables

- **Categorical variables** are variables which contain information about a category. Numbers are usually used as placeholders.
  - Gender: Male = 0 or Female = 1.
  - Political party affiliation: Democrat = 0, Republican = 1, Independent = 2.
Discrete variables

- **Discrete variables** are those in which the possible values for a finite set of numbers.
- Gender: Male = 0 or Female = 1.
- Political party affiliation: Democrat = 0, Republican = 1, Independent = 2.
Continuous variables

- Variable can take on a (theoretically) infinite number of values on $\mathbb{R}$ (the real number line).
- Used mostly to treat variables that can take on many, but not necessarily and infinity number of values for statistical reasons.
- Eg) age, vote share, etc are all treated as continuous even though they may not be technically.
Dealing with variables in \( R \)

- Loading data.
- Accessing data.
- Basic numerical operations.
Data

Data: 2016 Presidential Election Data by precinct and state

Thanks to former UGA undergrad (now at Harvard) Stephen Pettigrew!
https://goo.gl/hHvD2q.

Variables

- **state** – abbreviated state.
- **jurisdiction** – election precinct.
- **candidate** – presidential candidate.
- **votes** – number of votes for the candidate.
- **totalvotes** – total number of votes in the precinct.
Loading the data in \texttt{R}

### Getting started

1. Download the data: `president-long.csv`
2. Download the code: `election-data.r`
Randomization

- Randomization is probably one of the *most* important concepts in statistical inference.
- It allows us to make causal inferences.
- It allows us to design studies which allow us to make inferences about populations from samples.
Randomization and causal inference

- Randomization is used in experiments to make causal inferences about an intervention.
- It can also be used to make causal inferences outside of experiments.
Randomization and causal inference: clinical trials

- Assessing the causal effect of a drug is done via *clinical trials* as stipulated by the FDA.
- Patients randomly assigned to a treatment (the actual drug) or control/placebo (a sugar pill)
- Outcomes measured to test effectiveness of drug.
Randomization and causal inference: clinical trials

- Who receives treatment or placebo is controlled by the experimenter – allows for causal inference.
- Who decides to participate in the clinical trial is not – voluntary participation.
- Potential issues?

Professor Jason Anastasopoulos  ljjanastas@uga.edu

[POLS 4150] R, Randomization and Sampling

University of Georgia
Randomization and design

- Randomization is also what allows us to design studies and polls to obtain accurate information about populations from samples.
- Allows us to estimate *parameters* using *statistics*.
Pew research polls

Among presidential candidates, Trump seen as least religious

% of adults who view each candidate as _____ religious

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Very</th>
<th>Somewhat</th>
<th>NET</th>
<th>Not too</th>
<th>Not at all</th>
<th>NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ben Carson</td>
<td>35%</td>
<td>33%</td>
<td>68%</td>
<td>7%</td>
<td>4%</td>
<td>11%</td>
</tr>
<tr>
<td>Ted Cruz</td>
<td>25%</td>
<td>40%</td>
<td>65%</td>
<td>7%</td>
<td>5%</td>
<td>12%</td>
</tr>
<tr>
<td>Marco Rubio</td>
<td>16%</td>
<td>45%</td>
<td>61%</td>
<td>8%</td>
<td>5%</td>
<td>13%</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>10%</td>
<td>38%</td>
<td>48%</td>
<td>22%</td>
<td>22%</td>
<td>43%</td>
</tr>
<tr>
<td>Bernie Sanders</td>
<td>5%</td>
<td>34%</td>
<td>40%</td>
<td>18%</td>
<td>17%</td>
<td>35%</td>
</tr>
<tr>
<td>Donald Trump</td>
<td>5%</td>
<td>25%</td>
<td>30%</td>
<td>22%</td>
<td>37%</td>
<td>60%</td>
</tr>
</tbody>
</table>

Source: Survey conducted Jan. 7-14, 2016. Don’t know/refused responses not included.
### Pew research polls

**One-third of GOP voters are white evangelicals; more than a quarter of Democratic voters are religious ‘nones’**

**Religious profile of registered voters**

<table>
<thead>
<tr>
<th>Religious Group</th>
<th>Among all RVs, %</th>
<th>Among Dem RVs, %</th>
<th>Among Rep RVs, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>White evangelical Protestant</td>
<td>20%</td>
<td>8%</td>
<td>35%</td>
</tr>
<tr>
<td>White mainline Protestant</td>
<td>14</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Black Protestant</td>
<td>9</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Other Protestant</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>White Catholic</td>
<td>13</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Hispanic Catholic</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Other Catholic</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Religiously unaffiliated</td>
<td>21</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>All others</td>
<td>11%</td>
<td>12%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Note: Based on U.S. registered voters. “Dem” includes Democratic and Democratic-leaning registered voters and “Rep” includes Republican and Republican-leaning voters. The “Other Protestant” category includes other minority Protestants as well as Protestants who declined to provide their race and a few white Protestants who were not asked the evangelical/born-again question. The “Other Catholic” category includes other minority Catholics as well as Catholics who declined to provide their race. The “All others” category includes Mormons, Orthodox Christians, Jews, Muslims, Buddhists, Hindus, members of other faiths, and those who decline to provide their religion; individually, each of these groups accounts for 2% or less of all registered voters.

Source: Aggregated Pew Research Center surveys conducted January-June 2016.

“Evangelicals Rally to Trump, Religious ‘Nones’ Back Clinton”

Professor Jason Anastasopoulos  ljanastas@uga.edu

University of Georgia
Pew research polls

High Levels of Concern about Islamic Extremism in Most Countries

Are you ___ about Islamic extremism in our country?

<table>
<thead>
<tr>
<th>Country</th>
<th>Not concerned</th>
<th>Concerned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lebanon</td>
<td>8%</td>
<td>92%</td>
</tr>
<tr>
<td>Tunisia</td>
<td>19%</td>
<td>80%</td>
</tr>
<tr>
<td>Egypt</td>
<td>22%</td>
<td>75%</td>
</tr>
<tr>
<td>Palest ter.</td>
<td>31%</td>
<td>65%</td>
</tr>
<tr>
<td>Jordan</td>
<td>36%</td>
<td>62%</td>
</tr>
<tr>
<td>Turkey</td>
<td>37%</td>
<td>50%</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>25%</td>
<td>69%</td>
</tr>
<tr>
<td>Pakistan</td>
<td>14%</td>
<td>66%</td>
</tr>
<tr>
<td>Malaysia</td>
<td>24%</td>
<td>63%</td>
</tr>
<tr>
<td>Indonesia</td>
<td>55%</td>
<td>39%</td>
</tr>
<tr>
<td>Nigeria</td>
<td>21%</td>
<td>72%</td>
</tr>
<tr>
<td>Senegal</td>
<td>47%</td>
<td>46%</td>
</tr>
<tr>
<td>Israel</td>
<td>14%</td>
<td>84%</td>
</tr>
</tbody>
</table>

PEW RESEARCH CENTER
Randomization in the context of polling/survey research

\[ X = \{ x_1, x_2, \cdots, x_n \} \]

Select a set of individuals, \( N \), with \( p(N_1) = p(N_2) = \cdots \)

- Randomization, for this class, will involve a process known as simple random sampling.
- This process involves randomly choosing a sample of individuals \( N \) of size \( k \) from, for example a list.
- The list of subjects in the population is known as the sampling frame.
Simple random sampling

There are two ways to randomly sample:

1. With replacement – randomly choose one person from the sampling frame. Put it back. Randomly choose another.
2. Without replacement – randomly choose one person from the sampling frame, set it aside, randomly choose another person from the sample frame.
Simple random sampling with replacement

Possible Samples: $(A, B), (A, A), (B, E), \cdots$

$\# \text{ of Possible Samples: } N^k = 5^2 = 25$

$P(\text{Any Sample}): \frac{1}{N^k} = \frac{1}{25}$

- Imagine a sampling frame containing only 5 people: ABCDE.
- Let’s say we want to randomly sample 2 out of the 5 with replacement.
Simple random sampling without replacement

Possible Samples: \((A, B), (A, C), (B, C), \cdots\)

\[
\text{# of Possible Samples: } \binom{N}{k} = \frac{N!}{(N-k)!k!} = \frac{5!}{(5-2)!2!} = \frac{120}{12} = 10
\]

\[
P(\text{Any Sample}): \frac{1}{\binom{N}{k}} = \frac{1}{10}
\]

- Imagine a sampling frame containing only 5 people: ABCDE.
- Let’s say we want to randomly sample 2 out of the 5 with replacement.
Sampling with v. without replacement

- Sampling without replacement in the real world is almost always preferred.
- Why?
- Now algorithms perform sampling almost exclusively.
Randomization and sampling in action: polling Athens

- Polls rely on random sampling to make inferences about populations.
- We might be interested, for example, in knowing what % of people in Athens-Clarke County support the building of a border wall.
Randomization in action: polling Athens

**Step 1** Design a survey which asks respondents for information relating to your question.

- This can be done using *Qualtrics* or other survey software: https://ugeorgia.qualtrics.com/SE/?SID=SV_0HUOtBnB91o4iLOZ
- Or can be accomplished with a paper survey.
Randomization in action: polling Athens

Step 2  Find a list of registered voters in Athens-Clarke County:

- This list typically contains voters names and addresses but may also contain email contact information.
- The list of registered voters is now the population that you will be sampling from.
Randomization in action: polling Athens

**Step 3** Determine sample size – this will depend on a lot of factors including how much money you have and how accurate you would like the poll to be. We will discuss sample size calculation in greater detail later on.

For the sake of argument let’s say that we’ll sample $k = 2,000$ registered voters from the pool of about $N = 500,000$. 
Randomization in action: polling Athens

In R using the sample() function

"sample(x, size, replace = FALSE, prob = NULL)"

> sampling.frame = c("John", "Mary", "Joe", "Lefteris",..)
> random.sample = sample(x = sampling.frame, size = 2000)
> random.sample
[1] "Jack" "John" ...

Step 4  Randomly select $k = 2,000$ registered voters using an algorithm (with replacement).

In R we might use the sample() function.
Randomization in action: polling Athens

**Step 5** Collect data by sending out emails or mailing surveys to the people in your sample.

**Step 6** Analyze the data and write a paper/report or sell it to someone for $$$.