Class Notes

this time: sampling, probability part 1
next time: no class Mon, April 24

Wed, April 26: probability part 2
- discussions will occur as usual on Monday
- DD office hour Tuesday 4-5 cancelled, rescheduled to Friday, April 28 3-4 pm

\[ \text{potential subject} \xrightarrow{\text{randomize}} T \xrightarrow{C} \text{completely randomized design} \]

a different design: Pre-Post (or before and after):

ex: homework 1 #3 - drug for insomniacs.

Get 10 people who have trouble sleeping, measure their sleep while taking the drug and while not taking:

<table>
<thead>
<tr>
<th>Person #</th>
<th>(T) drug</th>
<th>(C) no drug</th>
<th>(T-C) difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.3</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>1.1</td>
<td>1.1</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6.1</td>
<td>6.2</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

randomize: \( \frac{1}{2} \) get and \( \frac{1}{2} \) don't get drug
**Sampling**

[p. 53 of reader] - Case study 6 - *Literary Digest* Poll

Sample Survey:

*Population:*

- all subjects (people, things) of interest

- subset → sample → observed

- unobserved (unsampled)

*How to find what kind of sample to take?*

**Pop**

- all voters in 1936 election

<table>
<thead>
<tr>
<th>favor Roosevelt?</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
</tr>
<tr>
<td>yes</td>
</tr>
<tr>
<td>no</td>
</tr>
</tbody>
</table>

\[ N = 80,000,000 \]

**Sample**

- the data

<table>
<thead>
<tr>
<th>favor Roosevelt?</th>
</tr>
</thead>
<tbody>
<tr>
<td>no turn no = 0</td>
</tr>
<tr>
<td>yes turn yes = 1</td>
</tr>
</tbody>
</table>

\[ \text{Mean (} \hat{g} \text{)} = \hat{p} \leftarrow \text{estimate for } \mu \]

- \( \mu \) and \( \hat{p} \) are "parameters" — unknown population summary of interest

**But how to choose?**

- Literary Digest thought \( p = 4\% \), error = 19%

- actual \( p = 60\% \)

- they did not choose the people properly
- basic principle: how sample and unsample should be related
- should be as similar as possible in all relevant ways except for the sample: unsample distinction by exactly the same as the randomized treatment and control groups - so we should do the same thing - CHOOSE AT RANDOM.

Ex: \[
\begin{bmatrix}
1 \\
2 \\
9
\end{bmatrix}
\begin{bmatrix}
y_1 = 2 \\
y_2 = \_ \\
y_9 = \_
\end{bmatrix}
\] - with replacement (independent identically distributed) - IID
- without replacement (simple random sampling)

Q: Did LD use IID or SRS?
A: No.
Q: Did their method of sampling produce a sample like that of IID or SRS?
A: No. They oversampled Republicans by sampling only wealthy people (those with phones or in clubs).

What did they do wrong?
"bias" - systematic tendency to over or under estimate the truth
- IID and SRS are unbiased

4. Biases
(1) selection bias - over or under represent relevant subsets of pop.
(2) nonresponse bias - people who chose to respond to survey may be systematically different in relevant ways (good response rate is 85% or higher)
2. response bias - the way the question is asked can influence the answer

4. Hawthorn effect - people tend to change their behavior when they know they're being tested