Soc 60

SAMPLING
1936 Presidential Election

- Literary Digest Poll
  - 2 million people
  - 55% Alf Landon
  - 41% ???,
  - 4 Third party candidate

Alf Landon, Republican governor of Kansas
1936 Presidential Election

- Actual Outcome
  - 39% Alf Landon,
  - 61% Franklin D. Roosevelt

- George Gallup’s Poll
  - 3,000 people
  - 46% Landon
  - 54% FDR

- FDR
  - 4-term President of the US
  - 1932-1945
1948 Presidential Election

• Gallup Poll
  – 49.5% Thomas E. Dewey
  – 44.5% ???

Thomas E. Dewey
Republican Governor of New York
1948 Presidential Election

• Actual Outcome
  – 45.1% Dewey
  – 49.5% Harry Truman
Advantages of Sampling

- Cheaper
- Often the only possible way
- Better quality control over measurement
- Investigation can be “destructive”
Sample Planning

- **Step 1. Define the Sample components and the population**
  - Population and Sample
    - Define population very clearly
    - Parameter vs. Statistic
  
  - Elements (the ultimate or smallest units we are interested in)

  - Sampling Units (the units that we sample)
Sample Planning (cont.)

- Step 2. Evaluate Generalizability

  - Can findings be generalized from sample → populations?
    - Does the sample represent the population?
  - Questions:
    - From what population were the cases selected?
    - How were the cases selected (by what method)?

  - Can findings be generalized from a study of one population → another population

- Step 3. Assess the Diversity of the Population

- Step 4. Consider Census
Non-Probability Sampling Methods

• Availability Sampling
• Purposive Sampling
  – What is a typical case
  – Taking extreme cases
    • Getting a large enough spread
• Snowball Sampling
• Quota Sampling
# Distribution of people in the population

<table>
<thead>
<tr>
<th>Hypothetical Census Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
</tr>
<tr>
<td>Urban</td>
</tr>
<tr>
<td>Rural</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
Distribution of the quota sample

$N=400$

- The quotas:
  - 100 urban men   (25% of 400)
  - 100 rural men
  - 120 urban women (30% of 400)
  - 80 rural women  (20% of 400)
- Total 400 respondents
Probability
Sampling Methods

- **Simple Random Sampling**
  - Sampling with or without replacement

- **Systematic Random Sampling**
  - Total number of cases (M) divided by the sample (N), this is your sampling interval K. \( M/N=K \)
  - Use random start. Select each Kth case

**Stratified Random Sampling**
- Create homogenous groups (strata)
- Sample randomly from each separately

**Cluster Sampling**
- Pick groups (clusters) randomly (weight groups by size)
- Interview/observe every member in the group
General Rules of Probability Sampling

• The larger the sample the more confidence we have in the representativeness of our sample

• The more homogenous our population is the more confidence we have in the representativeness of our sample

• The fraction of the population that a sample contains does not affect the sample representativeness unless the fraction is large.(less than 2%)
Rules of Probability

Addition Rule

The probability of either of two incompatible (mutually exclusive) events happening is the probability of the first plus the probability of the second.

\[ P(A \text{ or } B) = P(A) + P(B) \]

The sum of all the probability of all incompatible events is 1.

- Multiplication Rule
  - The probability of two independent events happening together is the probability of the first times the probability of the second.
    \[ P(A \text{ and } B) = P(A) \times P(B) \]
Probability Distributions

• Imagine we flip a coin four times

Fair coin  P(H)=P(T)=.5
HH
• P(H and H)=.5*.5 ← Multiplication Rule
• P(H and H and H and H)=P(HHHH)=.5*.5*.5*.5=.0625
• P(HHHT)=5*.5*.5*.5=.0625

• P(3H, 1T in any order)=
• P(HHHT)+P(HHTH)+P(HTHH)+P(THHH)=4*.0625=.25
• 0H 0 TTTT
• 1H .25 TTTT, TTHT, TTTH, HTTT
• 2H .5 HHTT, HTHT, HTTH, THHT, THTH, TTHH, THHT
• 3H .75 HHHH, HHTH, HTHH, THHH
• 4H 1.0 HHHH
Imagine we take a sample of 4 students from UCSD where half of the students are Male and half Female

\[
P(M) = P(F) = .5
\]

• \( P(M \text{ and } M) = .5 \times .5 \quad \leftarrow \text{Multiplication Rule} \)

• \( P(M \text{ and } M \text{ and } M \text{ and } M) = P(MMMM) = .5 \times .5 \times .5 \times .5 = .0625 \)

• \( P(MMMF) = 5 \times .5 \times .5 \times .5 = .0625 \)

• \( P(3M, 1F \text{ in any order}) = \)

• \( P(MMMF) + P(MMFM) + P(MFMM) + P(FMMM) = 4 \times .0625 = .25 \)

• 0M 0 \ FFFF

• 1M .25 \ FFFM, FFMF, FMFF, MFFF

• 2M .5 \ MMFF, MFMF, MFFM, FMFM, FFMM, FMMF

• 3M .75 \ MMMF, MMFM, MFMM, FMFM

• 4M 1.0 \ MMMM
## Sampling Distributions

**Binomial Distribution**

- \( N = 4 \)  \( P = .5 \)

<table>
<thead>
<tr>
<th>( X )</th>
<th>( P(X) )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.06250</td>
<td>0.06250</td>
</tr>
<tr>
<td>.25</td>
<td>.25000</td>
<td>.31250</td>
</tr>
<tr>
<td>.5</td>
<td>.37500</td>
<td>.68750</td>
</tr>
<tr>
<td>.75</td>
<td>.25000</td>
<td>.93750</td>
</tr>
<tr>
<td>1.0</td>
<td>.06250</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

**Expected Value**

\[ E(X) = .50 \]
Sampling Distributions (cont.)

- \( N = 10 \quad P = .5 \)

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
<th>PROBABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.00098</td>
<td>.00098</td>
</tr>
<tr>
<td>.1</td>
<td>.00977</td>
<td>.01074</td>
</tr>
<tr>
<td>.2</td>
<td>.04395</td>
<td>.05469</td>
</tr>
<tr>
<td>.3</td>
<td>.11719</td>
<td>.17188</td>
</tr>
<tr>
<td>.4</td>
<td>.20508</td>
<td>.37695</td>
</tr>
<tr>
<td>.5</td>
<td>.24609</td>
<td>.62305</td>
</tr>
<tr>
<td>.6</td>
<td>.20508</td>
<td>.82812</td>
</tr>
<tr>
<td>.7</td>
<td>.11719</td>
<td>.94531</td>
</tr>
<tr>
<td>.8</td>
<td>.04395</td>
<td>.98926</td>
</tr>
<tr>
<td>.9</td>
<td>.00977</td>
<td>.99902</td>
</tr>
<tr>
<td>1.0</td>
<td>.00098</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

The probability of **no** men is \(.5^{10} = .00098\) or cc. .1%

\[ E(X) = .50 \]
Sampling Distributions (cont.)

- BINOMIAL DISTRIBUTION
  - N = 100  P = .5

- CUMULATIVE

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
<th>PROBABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>.45</td>
<td>.04847</td>
<td>---------</td>
</tr>
<tr>
<td>.46</td>
<td>.05796</td>
<td></td>
</tr>
<tr>
<td>.47</td>
<td>.06659</td>
<td></td>
</tr>
<tr>
<td>.48</td>
<td>.07353</td>
<td></td>
</tr>
<tr>
<td>.49</td>
<td>.07803</td>
<td></td>
</tr>
<tr>
<td>.50</td>
<td>.07959</td>
<td>.72875</td>
</tr>
<tr>
<td>.51</td>
<td>.07803</td>
<td></td>
</tr>
<tr>
<td>.52</td>
<td>.07353</td>
<td></td>
</tr>
<tr>
<td>.53</td>
<td>.06659</td>
<td></td>
</tr>
<tr>
<td>.54</td>
<td>.05796</td>
<td></td>
</tr>
<tr>
<td>.55</td>
<td>.04847</td>
<td>---------</td>
</tr>
</tbody>
</table>

- E(X) = .50
- The probability of no (or all) men is .5^{100} = .00000000000000000000000000000000000000000079
Central Limit Theorem:

- 1. The sampling distribution is a normal distribution.
- 2. The average of the sample averages will be the population parameter.
- 3. As you increase the sample size the samples will cluster closer and closer to the population parameter (less sampling error or smaller standard error).
Sampling Distribution Animated Demo

• http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/
Figure 6-3 Standard Normal Distribution

X bar

-3z  -2z  -1z  0  +1z  +2z  +3z

.34 + .34 = area

.47 + .47 = area

.49 - .49 = area
Sampling error

• **Confidence level**
  - 90%, 95%

• **Margin of error**
  - Estimate + or – Multiplier*Standard Error
    - Multiplier: depends on our confidence level,
      - the higher the confidence level the larger the multiplier
    - Standard Error: a measure of the spread of the sampling distribution
      - the greater the diversity in the population, the greater the spread of the sampling distribution

• Confidence level 95% the maximum sampling error of a proportion is

• N=400 ~+- 5%
• N=900 ~+- 3.3%
• N=1600 ~+- 2.5%
• N=6400 ~+- 1.2%
Determining Sample Size

- Less error $\rightarrow$ larger sample
- More homogenous population $\rightarrow$ smaller sample
- More variables cross cutting $\rightarrow$ larger sample
- When weak relationships are expected $\rightarrow$ large sample

- Usual sample size: over 400 (between 1000 and 1500)