Sampling & Sampling Distributions

Joseph J. Ferrare, Ph.D.
School of Interdisciplinary Arts & Sciences
University of Washington Bothell
www.josephferrare.com
Hypothetical Scenario

Imagine you are working for a candidate running for U.S. Senate to represent Washington State. To build a platform, you decide to learn what issues Washington voters are most concerned about.

How would you learn about these concerns in a way that is feasible?
What and How to Collect?

Primary Data
- Involves collecting original data
- Often collected through surveys, interviews, and observations
- Advantages: adaptable to specific research questions
- Disadvantages: resource intensive, participant burden, often relies on recall

Secondary Data
- Makes use of pre-existing data
- Archives, databases, publications, media, etc.
- Advantages: often does not rely on actor recall, can be less burdensome
- Disadvantages: constrains research questions, data entry woes
Sampling Terminology

**Population**: the entire set of observations about which researchers are interested in learning

**Sample**: subset of observations used to make inferences about the population

Image from Quara.com
“A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well.”

- What is the population?
- What is the sample?
- Can you identify any problems with choosing the sample in the way that the teacher did?
Example #4 from Course Text

“A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.”

■ What is the population?
■ What is the sample?
■ Can you identify any problems with choosing the sample in the way that the coach did?
Be Error Aware

**Sampling error**: the difference between our statistic and the population parameter

**Coverage error**: difference between sampling frame and population

**Nonresponse error**: differences between those who do and do not respond to your data collection efforts (e.g. survey)

**Measurement error**: difference between a measured value and its true value
Probability v. Non-Probability Samples

**Probability Sample:** each unit of a population has a known and non-zero probability of selection

- Should be representative (i.e. approximate population characteristics of interest)
- A sample will be representative if all members have an equal probability of selection (EPSEM)

**Non-Probability Sample:** sample is selected based on a non-random procedure
Probability Samples

Simple Random Sampling

- All members of the population have an equal probability of selection
- Usually requires a list of all population members
  - E.g., all full time undergraduate students at UWB
- Each member of the population is assigned a number and random numbers (computers) are used to select units for the sample
Probability Samples

**Systematic Sampling**

- Use every $k^{th}$ unit from the population for inclusion in the sample
- Can be useful if no list exists (e.g., sampling transit riders)
- Nearly identical to SRS (empirically & in form)
Probability Samples

Stratified Sampling
• Based on a random selection of appropriate numbers of units taken from within homogenous subsets (stratum) of the population
• Requires that you have information to define strata
• Important tool for low incidence populations
Non-Probability Samples

- **Convenience Sampling:** relies on who is available (e.g., students in large lectures)
- **Purposive or judgmental:** selection based on information about the population & purpose of the study
- **Snowball:** collect data from alters of respondents
- **Quota:** sampling based on characteristics but without random selection
Sampling Distributions

Recall that our sample statistics are estimates of population parameters
• It is likely that our statistics will not be identical to the parameter
• It is also likely that our statistics across samples will vary

Figure 1. The pool balls.

All possible outcomes are shown below in Table 1.

Table 1. All possible outcomes when two balls are sampled with replacement.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Ball 1</th>
<th>Ball 2</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2.0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Example from course text
Sampling Distributions of a Mean

![Bar chart and table]

### Table

<table>
<thead>
<tr>
<th>Mean</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1</td>
<td>0.111</td>
</tr>
<tr>
<td>1.5</td>
<td>2</td>
<td>0.222</td>
</tr>
<tr>
<td>2.0</td>
<td>3</td>
<td>0.333</td>
</tr>
<tr>
<td>2.5</td>
<td>2</td>
<td>0.222</td>
</tr>
<tr>
<td>3.0</td>
<td>1</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Example from course text
Sampling Distributions of a Range

Figure 3. Distribution of ranges for N = 2.

Table 3. All possible outcomes when two balls are sampled with replacement.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Ball 1</th>
<th>Ball 2</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4. Frequencies of ranges for N = 2.

<table>
<thead>
<tr>
<th>Range</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>0.333</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.444</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.222</td>
</tr>
</tbody>
</table>

Example from course text
Sampling Distributions

A sampling distribution of the mean has its own properties:

• Mean of the sampling distribution of the mean

\[ \mu_M = \mu \]

- \( \mu = \) pop. mean

• Variance of the sampling distribution of the mean

\[ \sigma^2_M = \frac{\sigma^2}{n} \]

- \( \sigma^2 = \) pop. variance
- \( n = \) sample size

• Standard error of the mean (i.e., SD of the sampling distribution of the mean)

\[ \sigma_M = \frac{\sigma}{\sqrt{n}} \]

- \( \sigma = \) pop. SD
- \( n = \) sample size
Central Limit Theorem

For random samples, as the sample size increases, the sampling distribution of the sample mean approaches normality

• This is true regardless of the shape of the population distribution
• The sampling distribution takes a more normal shape as the sample size increases
• Thus, we can know the probabilities associated with our sample mean
Central Limit Theorem

Note here from the simulation that even for a variable that has a wildly non-normal distribution, the sampling distribution of the mean will approach normality as the sample size increases.

Example from course text
Central Limit Theorem in Action

Mean = 49073.96
Stdev. = 42437.211
N = 2.352
Skew=1.247