Introduction to Inference

Population Parameter | Sample Statistic
---|---
Mean $\mu$ | Sample Mean $\bar{x}$
Proportion $p$ | Sample Proportion $\bar{p}$
Standard Deviation $\sigma$ | Sample Standard Deviation $s$

EX 1: Estimate $\mu =$ mean age of all DS 212 students
How? Take a random sample of, say, $n = 50$ students.
Suppose we find $\bar{x} = 22.3$ years $\leftarrow$ “point estimate”
Construct a 95% Confidence Interval (CI) for $\mu$:
CI = $[\text{Point Estimate} \pm \text{Margin of Error}] = [\bar{x} \pm MoE]$
= $[22.3 \pm 1.01] = [21.29, 23.31]$
We think $\mu$ is somewhere in the CI with 95% confidence.

EX 2: Do a Hypothesis (Significance) Test for, say, $p =$ % of all Californians who favor Proposition 75
“Null hypothesis” $H_0$: $p = 50$
“Alternative hypothesis” $H_a$: $p < 50$
How? Take a random sample of say, $n = 1,000$ likely voters.
Suppose we find $\bar{p} = 42$% favor Proposition 75
Calculate the probability of seeing such a low sample %
if the true population % really = 50%.
-- This probability is called the “p-value of the test”
The smaller the p-value, the more we want to reject $H_0$
Sampling Variability

In each case we take a random sample of \( n \) items (often people). Each time we do so, we get a different sample value. So, sample statistics are random variables:

- \( \bar{x} \) has a probability distribution called the “sampling distribution for the sample mean”
- \( \bar{p} \) has a probability distribution called the “sampling distribution for the sample proportion”

Once we understand these sampling distributions, we’ll be able to figure out the CI’s MoE and Significance Test’s p-value.

Sampling

Sample = A representative subset of the population

Examples:
- Auto Insurance Rates (from 100,000 drivers’ accidents)
- US Unemployment Rate (based on 60,000 households)
- Nielsen Ratings (only a few thousand households)

Advantages:
- Cheaper than looking at the whole population
- Faster …. and can be updated more often
- May be safer … e.g., new drug testing
- Possibly more accurate … e.g., U.S. Census
Sampling Design

- Key: Use *randomness* to select from the population
- Why? To avoid introducing any *bias* into the sample.
  - More on this later

**Randomization:**
- Gives everyone in the population the *same chance* of being selected for the sample
- Can be accomplished in several ways
  - We’ll look at 3 techniques using *random numbers*

What are Random Numbers (RNs)?

1. Unpredictable & independent
2. They all have an equal chance of occurring

How do I get random numbers?
- Dice
- Spinner
- Calculator
- Computer
- Table 7.1 in textbook, p. 275
Simple Random Sampling

- \( N \) items in the Population
- \( n \) items in the Sample Random Sample (SRS)

1. Label each unit in population with a unique ID

   **EX:**  
   \[ \begin{array}{ll}
   N & \text{Labels} \\
   1-100 & \text{2-digit RNs: 00, 01, 02, \ldots, 99} \\
   101-1000 & \text{3-digit RNs: 000, 001, 002, \ldots, 999} \\
   \end{array} \]

2. Use RNs to select \( n \) items
3. Ignore repeats and RNs larger than \( N \)

Systematic Random Sampling

1. Label each unit in population with a unique ID #
2. Use a RN to randomly select the \( 1^{st} \) item
3. Select every \( k^{th} \) item thereafter until \( n \) items have been selected, where \( k = \frac{N}{n} \)

   **EX:** Population size \( N = 100 \)  
   Desired sample size \( n = 5 \)  
   Then \( k = \frac{100}{5} = 20 \).  
   Suppose the first RN = 18.  
   Remaining picks would be __, __, __, and __
Systematic Random Sampling

Benefits:
• Don’t have to keep on generating RNs
• Avoids repeats, so may go faster than SRS
• It spreads selections out over the population and avoids clustering that may occur
• Still has properties of SRS

Usage: Last stage of a large sample survey, e.g., picking specific apartments with a large apartment building

Stratified Random Sampling

Tries to force sample to look like the population.
1. Divide population into groups of similar items (strata)
2. Choose a separate SRS within each group in proportion to its percentage of the whole population
3. Combine the samples

😊 May not need to sample as much as before.
😊 This is more difficult than other methods:
• Hard to determine each group’s % of population
• Must know population well before you sample from it
Sampling Biases (in surveys especially)

1. **Under-Coverage Bias:**
   Too few (or too many) of one type, *e.g.*, Geographic, Income, Age, Gender, Race, …

2. **Response Bias:**
   - Respondents may be intimidated by questioner;
   - When people don’t answer truthfully.
   - **EX:** People say they’re going to vote for X, then vote for Y (maybe they’re embarrassed to admit their true feelings).

3. **Non-Response Bias:**
   - When people don’t answer at all (*e.g.*, “Ring-no-answer”);
   - Some folks may be hard to reach or refuse to cooperate.

---

More Biases

4. **Self-Selection Bias:**
   - When people volunteer their responses they often have strong opinions (usually, in the negative)
   - **EX:** Dear Abby – Who would marry their spouse again?

5. **Leading Questions:**
   - Some questions lead to a particular answer. Avoid these!
   - Similarly, avoid poorly worded or confusing questions.

**In Summary:**
- It’s not easy to get an unbiased, random sample, but we’ll assume we have valid, unbiased data in this course.