

# Sampling: Simple probability samples

```
$ echo "Data Science Institute"
```

# Learning Outcomes

*How might we select and study random individuals from a population? How do we effectively analyze a sample selected in this manner?*

1. Distinguish between different types of probability samples
2. Compute sample statistics for simple random samples
3. Identify scenarios in which certain probability sampling methods should be used or avoided

# Simple Random Samples

# Recall

- **Probability sampling** is a sampling method in which every population unit has a selection probability that is known to those conducting the sampling, and units are selected at random.
- Small probability samples can be used to make inferences about relatively large populations
- A **simple random sample (SRS)** is taken when every possible combination of population units has an equal chance of being sampled

# Sampling With versus Without Replacement

- Consider a population of size  $N$
- Simple random sampling **with replacement**:
  - i. Select one unit for measurement, with probability  $1/N$
  - ii. Sampled unit is return to the population
  - iii. Select second unit for measurement, with probability  $1/N$
  - iv. Repeat until desired sample size is obtained
- Simple random sampling **without replacement**:
  - i. Select one unit for measurement, with probability  $1/N$
  - ii. Select second unit for measurement, with probability  $1/(N-1)$
  - iii. Repeat until desired sample size is obtained. The final unit in the sample will be selected with probability  $1/(N-n+1)$ , where  $n$  is the total sample size.

# Sampling With versus Without Replacement

Basically:

- When we sample with replacement, an observation can be selected more than once
- When we sample without replacement, once an observation is selected, it cannot be selected again

# Sample Estimates and Variability for SRS

# Sample Mean and Variance

- Consider a sample of size  $n$  from a population of size  $N$ .
- The population mean,  $\mu$ , can be estimated using the sample mean,  $\bar{y}$ , calculated,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

(the sum of all observations in the sample, divided by sample size  $n$ ).

- The sample variance,  $s^2$ , can be calculated,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

and tells us how 'spread out' our sample data are.



# Estimator Variance

- The variance of  $\bar{y}$  is,

$$\hat{V}(\bar{y}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right)$$

- This is a measure of the variability of values of  $\bar{y}$  computed from different samples.
- **Estimator variance tells us how much our sample mean would differ across different samples  $n$**

# Estimator Variance

- $n/N$  is called the **sampling fraction** and represents the proportion of individuals from the population captured in the sample.
- $(1 - n/N)$  is called the **finite population correction**.
- As sample size increases, the value of the sampling fraction increases and the value of the finite population correction (and thus, estimate variance) decreases

**As sample size increases, we gain more information about the population, and the variance of our estimate decreases**

# Standard Error

- The **standard error (SE)** is the square root of the estimated variance of:  $\bar{y}$

$$SE(\bar{y}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)}$$

- SE tells us how much our sample mean would differ from our population mean, AKA: as sample size increases, we gain more information about the population and the variance of our estimate decreases

# Coefficient of Variation


- The SE can be used to calculate the **coefficient of variation (CV)** for:  $\bar{y} \neq 0$

$$\hat{CV}(\bar{y}) = \frac{SE(\bar{y})}{\bar{y}}$$

- The CV is a unitless measure of relative variability. The estimated CV is the SE expressed as a percentage of the sample mean.

# Sampling Weights

- Let  $\pi_i$  be the probability that population unit  $i$  is included in the sample.
- Then the **sampling weight** of unit  $i$  is defined as,


$$w_i = \frac{1}{\pi_i}$$

- The sampling weight can be interpreted as the number of population units that a given unit in the sample represents.

# SRS Sampling Weights

- All units have the same inclusion probability (by definition) –  $\pi_i = \frac{n}{N}$
- All sampling weights are the same –  $w_i = \frac{1}{\pi_i}$
- Each unit in the sample represents the same number of population units
  - This is called a self-weighting sample

# Systematic Sampling

# Systematic Sampling

1. Obtain a randomized list of the population (size  $N$  ).
2. Choose a sample size  $n$ .
3. Calculate  $N/n$  . If  $N/n$  is an integer, let  $k = N/n$  . If it is not an integer, let  $k$  be the next integer after  $N/n$  .  $k$  is the selection interval.
4. Select a random integer  $R$  between 1 and  $k$ .  $R$  is the starting point for the sampling procedure.
5. Working through the list, sample units  $R$  ,  $R+k$  ,  $R+2k$  , etc. until the end of the list is reached.



# Systematic Sampling Considerations

- Systematic sampling is a type of cluster sampling, not SRS, *but...*
- If the original population is truly randomized, systematic samples behave similarly to SRS and the same analysis methods can be used
- If the original population is in increasing or decreasing order, or periodic in some way (i.e. alternating male and female names), the sample will not be representative and will not behave like an SRS
- Usual considerations apply when defining a target population and frame
  - From what source are you obtaining your list?
  - If sampling in person, where are you sampling?

**When should an SRS be used?**

# When should an SRS be used?

- When you have a complete list of and access to all possible observation units
  - If your original list is not comprehensive, an SRS will not be representative and the usual analysis methods will not be possible
  - If the geographic area covered by the original list is too large, an SRS may contain units that cannot be observed and the usual analysis methods will not be possible
- When there is little supplementary information about the population that can be used to design the survey/study
  - This eliminates the possibility of stratified or cluster sampling
- When the main focus of the survey/study is multivariate relationships
  - Analysis can be much simpler for an SRS

# Next

Stratified sampling