Sampling: Stratified sampling

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

Learning Outcomes

How might our study be impacted if we divide our population into groups by shared characteristics before sampling? How do we effectively study a sample selected in this manner?

- Identify benefits of using stratified random sampling
- Compute sample statistics for stratified random samples
- Design a study using stratified random sampling
- Distinguish between stratified random sampling and quota sampling

What is stratified sampling?

Stratified Sampling

1. Divide the whole population into non-overlapping subpopulations based on shared characteristics. These subpopulations are called **strata** .
2. Take independent probability samples (often SRS) from each stratum.
3. Pool individual samples together to calculate overall population estimates.

Stratified sampling often requires supplemental information about a population in order to divide it into separate groups.

- For example if you have a list of all student emails from a university and you want to stratify by gender, this list will need to be linked with a data source that includes each student’s gender

Why stratify?

- Preventing a non-representative sample
- Seeking estimates with known precisions for certain subpopulations
- Convenience and lower cost
- Higher precision (lower variance) estimates for population means and totals

Sample Estimates and Variability

Sample and Population Sizes

- Suppose we have a population of size *N* divided into *H* strata. Let N_h be the number of population units in stratum *h* . Then we must have,

  > $$N_1 + N_2 + \ldots + N_H = N$$

- Suppose we then take an SRS from each stratum. Let n_h represent the size of the sample selected from stratum *h* . The total sample size is,

  > $$n_1 + n_2 + \ldots + n_H = n$$

- Sample and population sizes do not have to be equal across all strata

Sample Mean

- The sample mean for stratum *h* can be calculated,

> $$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hj}$$

- To estimate the population mean, the sample mean for the whole sample (from all strata combined) can be calculated,

> $$\bar{y} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

- This is a weighted mean of all sample strata means.

Stratum Sample Variance

- The sample variance for the the sample from each stratum can be computed the same way as an SRS:

  > $$s_h^2 = \sum_{i=1}^{n_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}$$

Estimator Variance and Error

- The variance of the sample mean can be computed,

> $$\hat{V}(\bar{y}) = \sum_{h=1}^{H} \frac{s_h^2}{n_h}(1 - \frac{n_h}{N_h})(\frac{N^h}{N})^2$$

- The standard error (SE) and coefficient of variation (CV) remain the same as for an SRS:

> $$SE(\bar{y}) = \sqrt{\hat{V}(\bar{y})}$$

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

Weights

Weights

- When using stratified sampling, weights may differ by stratum.
- The inclusion probability for unit *i* of stratum *h* is,

> $$\pi_{hi} = \frac{n_h}{N_h}$$

- Where n_h is the size of SRS from stratum *h* and N_h is the total number of units in stratum *h* .
- As previous, the sampling weight for unit *i* of stratum *h* is then,

> $$w_{hi} = \frac{1}{\pi_{hi}}$$

Using Sample Weights

- The population mean can be estimated directly using a weighted mean of recorded observations:

> $$\bar{y} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{h} w_{hi} y_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{h} w_{hi}}$$

- In stratified sampling, we need to sum over the weights and units in each stratum, and then sum over all strata.

Defining Strata & Allocating Observations

Defining Strata

- How do you divide your population into strata?
    - Mean values should differ greatly between strata
        - Stratify by a variable that is closely related to the variable(s) you are trying to estimate
        - For example, if you wish to estimate average height, you might stratify by age or sex instead of geographic location
    - Data availability
        - Is there existing survey data to help you define appropriate strata? If not, are you able to collect preliminary data for this purpose?
        - More supplementary data often means more strata

Defining Strata (continued...)

- How do you divide your population into strata?
  - Difficulty and cost
    - More strata may mean a higher cost or effort involved
    - Is this additional cost worthwhile for the precision you wish to achieve or the type of analysis you wish to conduct?

Allocating Observations to Strata

- How many units should you sample from each stratum?
    - Proportional Allocation
        - Sample the same proportion of units from each stratum
        - Sample weights (πhi) are the same for each sampled unit regardless of stratum
    - Optimal Allocation
        - Variation among larger sampling units may be greater than variation among smaller sampling units, so a higher proportion of large units should be sampled
        - Useful for businesses, cities, and institutions like schools or hospitals

Allocating Observations to Strata (continued...)

- How many units should you sample from each stratum?
  - Allocation for Precision with Strata
    - Sample to reduce the variation in stratum-level estimates, not population-level estimates
    - Useful when the goal is comparing estimates between strata

Quota Sampling

Quota Sampling

- Population is divided into subpopulations like strata
- **Non-probability sampling** is conducted within each subpopulation
    - Often convenience sampling is used
- Specified amounts (**quotas**) of types of units are selected

Why use quota sampling?

- Probability sampling may be expensive or impractical
- May give better results than a pure convenience sample due to enforced quotas
- Cheaper than probability samples

Why *not* use quota sampling?

- Prone to selection bias
- Methods of analysis for probability samples do not apply

Next

Differential Privacy