Sampling: Nonresponse

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

Learning Outcomes

Why do some individuals not respond to surveys? How can we encourage people to respond consistently to surveys when sampled? What can be done when non-response is unavoidable?

- Identify factors in survey or study design that may increase or decrease response rates.
- Use strategies to reduce nonresponse bias resulting from missing data
- Identify appropriate use cases as well as strengths and weaknesses for methods of dealing with nonresponse

Non-response

Types of Nonresponse

- **Unit nonresponse** occurs when an entire observational unit is missing
  - For example, a unit in the sample cannot be reached or is unable to respond
- **Item nonresponse** occurs when a specific measurement or measurements for a given observational unit are missing
  - For example, an individual might respond to a survey but refuse to state their income
- For non-human "respondents" (objects, land, animals, etc.), nonresponse is sometimes referred to plainly as **missing data**

Effects of Ignoring Nonresponse

- Inaccurate estimates
- Non representative samples
- Misallocation of resources
    - Increasing sample size without addressing nonresponse may result in even more biased estimates

Nonresponse Bias

- **Nonresponse bias** occurs when observational units that are measured differ systematically from observational units that are not measured. **This is usually the case**.
- When nonresponse bias is presents, results of a study can only represent the types of units who tended to respond to the survey, not the entire target population.
- Bias will be low if:
    - The mean for respondents is similar to the mean for non-respondents
    - The nonresponse rate is low

What is an acceptable response rate?

- Depends on the nature of nonresponse
  - If nonresponse bias is low, a lower response rate is fine
  - If nonresponse bias is high, any rate/amount of respondents may still produce invalid results
- Regardless of its value, it is important that the response rate is calculated consistently and reported with survey results.

Reducing Nonresponse

Response Rate Factors

- Factors that may influence response rate include:
  - **Survey content**: sensitive topics (particularly during interview-based surveys) may prevent some people from responding.
  - **Time of survey**: respondents may be harder to reach during certain times of the day (i.e. early in the morning), week (i.e. during business hours), or year (i.e. around major holidays).
  - **Surveyors**: some surveyors may be more skilled at observation and measurement than others. This may play a more significant role for non-human respondents.

Response Rate Factors

- **Data-collection method**: Mail, fax, and internet surveys tend to have higher unit nonresponse rates. Interview-based surveys tend to have lower item nonresponse rates
- **Questionnaire design**: Question wording and form design can influence individuals to respond or not respond.
- **Respondent burden**: Long or very detailed questionnaires may prompt individuals not to respond.

Response Rate Factors

- **Survey introduction**: A good introduction that provides context and motivation as well as ensuring respondent confidentiality can promote higher response rates.
- **Incentives or disincentives**: Giving benefits to respondents or revoking privileges from nonrespondents can help increase response rate.
- **Follow-up**: Following up, either using the same survey medium or a new one, can prompt people to respond even when they did not with the first attempt(s). This may get expensive, so it is important to keep track of the marginal cost and benefits for subsequent attempts.

Types of Missing Data

Propensity Scores

- Define random variable R_i such that,

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds} \\ 0 & \text{if unit } i \text{ does not respond} \end{cases}$$

- The **propensity score** for the i^{th} unit is the probability that the i^{th} unit will respond if sampled,

$$\phi = P(R_i = 1)$$

Types of Missing Data

- Let y_i represent the i^{th} response of interest, and x_i represent a vector of information known about unit *i* in the sample.
- We consider three types of missing data:
    i. Missing completely at random (MCAR)
    ii. Missing at random (MAR) given covariates
    iii. Not missing at random (NMAR)

Missing Completely at Random

- Data is **missing completely at random (MCAR)** if ϕ_i does not depend on y_i or x_i.
- Other features of data MCAR:
  - ϕ_i are equal for all values of *i*
  - All occurrences $\{R_i = 1\}$ are conditionally independent of one another and of the sampling procedure
  - Respondents are representative of the sample, so analysis of respondents produces approximately unbiased estimates for population parameters

Missing at Random Given Covariates

- Data is **missing at random (MAR) given covariates** if ϕ_i depends on x_i but not on y_i, meaning nonresponse is related to some features of a respondent, but not the specific variable of interest.
- We can account for nonresponse using a model since x_i values are known for all respondents regardless of their response status for y_i.

Not Missing at Random

- Data is not missing at random (NMAR) if ϕ_i depends on y_i and cannot be completely explained by x_i.
- Modelling can help adjust for some nonresponse if the nonresponse is partially dependent on x_i , but will not completely remove bias.

Dealing with Nonresponse

Two-Phase Sampling

Two-Phase Sampling

- **Two-phase sampling** is a sort of stratified sampling that attempts to produce estimates that account for nonresponse bias.
- Steps:
  - Take an SRS of *n* units from the population of *N* units.
  - Within this sample, n_R units will respond and n_M units will not respond, with $n_R + n_M = n$.
  - Resample some fraction of the n_M nonrespondents, with ν representing the sampling fraction.

Two-Phase Sampling: Estimates

- If all n_M nonrespondents respond in the second phase, we can use the following estimators for the population mean and total:

$$\hat{\bar{y}} = \frac{n_R}{n}\bar{y}_R + \frac{n_M}{n}\bar{y}_M$$

$$\hat{t} = N\hat{\bar{y}} = \frac{N}{n}\sum_{i=1}^{n_R} y_i + \frac{N}{n\nu}\sum_{i=1}^{n_M} y_i$$

- with $\bar{y}_R$ and $\bar{y}_M$ representing the sample means for the initial respondents and secondary respondents respectively.

Dealing with Nonresponse

Weighting Methods

Weighting Class Adjustment

- Variables known for all observational units in the sample are used to divide respondents into different **weighting adjustment classes** under the assumption that respondents and nonrespondents in the same class share similar characteristics.
- Weights of respondents in each class are **increased** according to the number of nonrespondents in the class.
- In final calculations, respondents in each class represent the nonrespondents in their class as well as themselves.

Weighting Class Adjustment

- Assumptions:
  - Data is MAR (ϕ_i does not depend on y_i)
  - ϕ_i is the same for all elements in each class
  - Nonrespondents in a given weighting class share similar responses to respondents in the same writing class
- Weighting classes should be constructed such that units in each class are as similar as possible with respect to the main variable(s) of interest (similar to stratification)

Weighting Class Adjustment

- As usual, $w_i = 1/\pi_i$ be the weight of unit *i* in the sample. The estimated response probability within weight adjustment class *c* is,

$$\hat{\phi}_c = \frac{\text{sum of weights for respondents in class c}}{\text{sum of weights for selected sample in class c}}$$

- The weight for unit *i* in class *c* then becomes,

$$\tilde{w}_i = \begin{cases} \frac{1}{\pi_i \hat{\phi}_c} & \text{if unit } i \text{ is a respondent in class c} \\ 0 & \text{otherwise} \end{cases}$$

Weighting Class Adjustment

- Weights $\tilde{w}_i$ can then be used to estimate the sample total and mean,

$$\hat{t}_{wc} = \sum_{i=1}^{n} \tilde{w}_i y_i$$

$$\hat{\bar{y}}_{wc} = \frac{\hat{t}_{wc}}{\sum_{i=1}^{n} \tilde{w}_i}$$

- where *wc* indicates that these sample estimates have been weighting class adjusted.

Poststratification

- Respondents are stratified and weights are modified so they match **population counts**.
- Procedure:
    i. An SRS is taken from the population.
    ii. Sampled units are grouped into *H* distinct poststrata (usually based on demographic variables).
    iii. Population units are grouped into same strata and counted.
    iv. The weight of each respondent in a given stratum is increased according to how many units in the corresponding population stratum they represent.

Poststratification: Weights

- Let N_h represent the number of population units in stratum *h* . Let n_h represent the number of sampled units and let n_{hR} represent the number of respondents in poststratum *h* .
- Let $x_{hi} = 1$ if unit *i* is a respondent in poststratum *h* and 0 otherwise. Let w_i represent the weight of unit *i* in the initial probability sample. Then its modified weight is,

$$w_i^* = w_i \sum_{h=1}^{H} \frac{N_h}{\sum_{j=1}^{n_h} w_j x_{hj}}$$

Weighting Methods Considerations

- Weight adjustment can **improve but not eliminate nonresponse bias**
- Always need to consider the plausibility of assumptions involved
- Always need to state and justify any adjustments or models used
- Weight adjustments are usually used for unit nonresponse (not item nonresponse)

Dealing with Nonresponse

Imputation

Imputation

- **Imputation** is the process of assigning values to missing items in a data set
- Used to reduce nonresponse bias and produce cleaner data for analysis
- Imputed values are often taken from other respondents with similar non-missing responses as the unit with the missing item

Imputation Considerations

- Imputation allows data to be analyzed using standard processes and software
- If data is MAR, imputation can greatly reduce item nonresponse bias
- Any imputation needs to be **well documented**
    - This may include: indicating which responses are imputed, which donor was used for a specific value, how many times a record is used as a donor
- Variance of estimates computed using imputed data will be smaller than the true variance

Next

Ethics