Stat472/572 Sampling: Theory and Practice

Instructor: Yan Lu

University of New Mexico

Chapter 8: Nonresponse

Nonresponse

- Unit nonresponse: the entire observation unit is missing
- Item nonresponse: some measurements are present for the observation unit but at least one item is missing

Example: select 10,000 households for sample,
1000 ineligible for the survey
700 no contact
200 unresolved
1400 refusals

$$\begin{aligned}\text{Nonresponse rate} &= \frac{\text{nonrespondents} + \text{unresolved}}{N - \text{ineligible}} \\ &= \frac{1400 + 700 + 200}{9000} \\ &= .25\end{aligned}$$

Four approaches to dealing with nonresponse:

1. Prevent it. Design the survey so that nonresponse is low. This is by far the best method.
2. Take a representative subsample of the nonrespondents; use that subsample to make inferences about the other nonrespondents.
   —-Call backs and follow ups
   Take a subsample of nonrespondents for interviewing
   —-American community survey
   subsample about 1/3 of households that did not respond to mail survey

3. Use a model to predict values for the nonrespondents.
   —-Weighting class adjustment methods implicitly use a model to adjust for unit nonresponse.
   —- Imputation often adjusts for item nonresponse,
   —-Parametric models may be used for either type of nonresponse.
4. Ignore the nonresponse (not recommended, but unfortunately common in practice).

Main problem caused by nonresponse: **potential bias**

Stratum	Size	Total	Mean	Variance
Respondents	N_R	t_R	$\bar{y}_{RU}$	S_R^2
Nonrespondents	N_M	t_M	$\bar{y}_{MU}$	S_M^2
Entire Popn	N	t	$\bar{y}_U$	S^2

- A probability sample from the population will likely contain some respondents and some nonrespondents.
- If the population mean in the nonrespondent stratum differs from that in the respondent stratum, estimating the population mean using only the respondents will produce bias.

- Let $\bar{y}_R$ be an approximately unbiased estimator of the mean in the respondent stratum, using only the respondents.

$$\bar{y}_U = \frac{N_R}{N}\bar{y}_{RU} + \frac{N_M}{N}\bar{y}_{MU}$$

The bias is approximately

$$E(\bar{y}_R) - \bar{y}_U \approx \frac{N_M}{N}(\bar{y}_{RU} - \bar{y}_{MU})$$

- The bias is small if
—- either (1) the mean for the nonrespondents is close to the mean for the respondents, or
—- (2) N_M/N is small, there is little nonresponse.
— But we can never be assured of (1), as we generally have no data for the nonrespondents. Minimizing the nonresponse rate is the only sure way to control nonresponse bias.

Designing Surveys to Reduce Nonsampling Errors

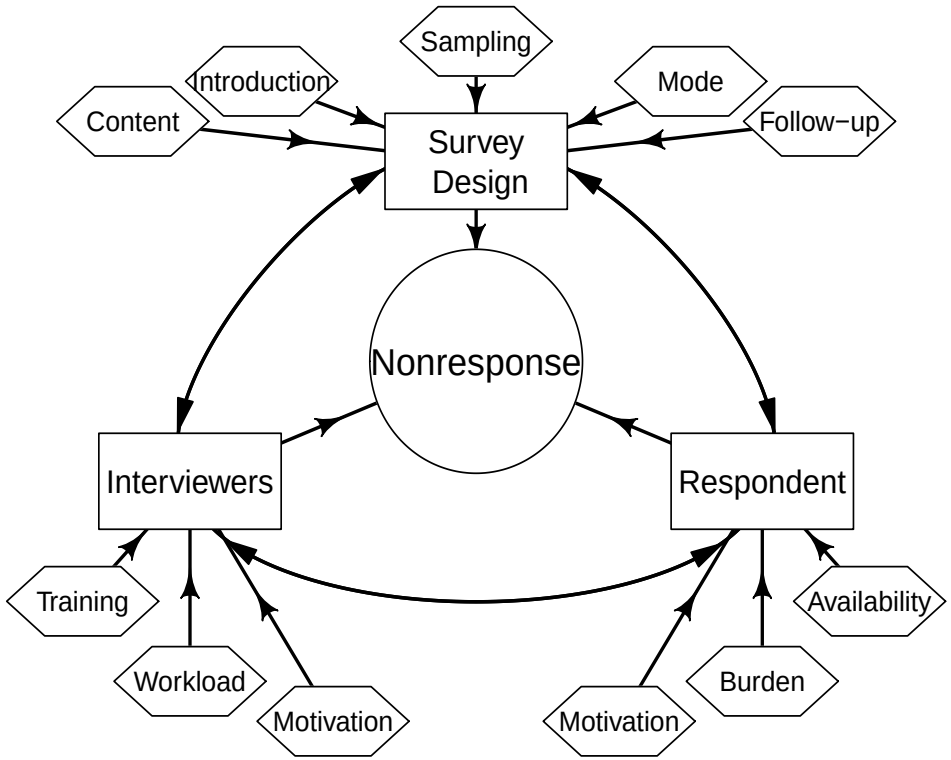

FIGURE 8.1
Some factors affecting nonresponse.

numbers for some of the addresses; these, if accurate, could be used for following up with nonrespondents. A sampling frame for a business survey may contain contact information for persons within the businesses who typically fill out the surveys. Having detailed, accurate information about units in the frame may make it easier to select

Two phase sampling (double sampling)

- Population divided into two strata, respondents and initial nonrespondents (persons who did not respond to the first call)
- Sample size n
- n_R respondents
- n_M nonrespondents (missing)
- Subsample $100v\%$ of the n_M nonrespondents in the sample
  —-the subsampling fraction v does not depend on the data collected
  —-suppose that through some superhuman effort all of the targeted nonrespondents are reached

The two phase sampling estimators of the population mean and total are

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

and

$$\hat{t} = N\hat{\bar{y}} = \frac{N}{n}\sum_{i\in S_R} y_i + \frac{N}{n}\cdot\frac{1}{v}\sum_{i\in S_M} y_i$$

$$E(\hat{t}) = t$$

$$\begin{aligned}\hat{V}(\hat{\bar{y}}) &= \frac{n_R-1}{n-1}\cdot\frac{s_R^2}{n} + \frac{n_M-1}{n-1}\cdot\frac{s_M^2}{vn} \\ &+ \frac{1}{n-1}\left[\frac{n_R}{n}(\bar{y}_R-\hat{\bar{y}})^2 + \frac{n_M}{n}(\bar{y}_M-\hat{\bar{y}})^2\right]\end{aligned}$$

Notes on Nonresponse and Two-Phase Sampling

- It is rare to achieve a 100% response rate in the subsample of nonrespondents.
- A common approach is to compare the subsample of nonrespondents to the original respondents to assess potential bias.
- If everyone in the subsample responds, two-phase sampling can:
  - Remove the bias introduced by the original nonresponse.
  - Properly account for nonresponse in the estimation of variance.

Mechanisms for Nonresponse

Define:

$$Z_i = \begin{cases} 1 & \text{if unit } i \text{ is selected in the sample (included in } S) \\ 0 & \text{otherwise} \end{cases}$$

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds} \\ 0 & \text{otherwise} \end{cases}$$

Probabilities:

- $P(Z_i = 1) = \pi_i$, the inclusion probability.
- $P(R_i = 1) = \phi_i$, the response propensity (propensity score).

Components:

- y_i: Response of interest.
- $\mathbf{x}_i$: Vector of auxiliary information known about unit i in the sample.
- If Z_i (selection) and R_i (response) are independent, then:

$$P(\text{unit } i \text{ is selected and responds}) = \pi_i \phi_i.$$

- Estimate the population total $t_y = \sum_{i=1}^{N} y_i$ using an adjusted Horvitz–Thompson estimator:

$$\hat{t}_\phi = \sum_{i=1}^{N} Z_i R_i \frac{y_i}{\pi_i \phi_i}.$$

Three Types of Missing Data (Little and Rubin, 1987)

1. **Missing Completely at Random (MCAR)**: The probability of responding, ϕ_i, does not depend on $\mathbf{x}_i$, y_i, or the survey design.

- Examples:
  - Laboratory error: Someone drops a test tube containing a participant's blood sample.
  - Survey example: A survey of student satisfaction at a university:
    - **X-variables**: Known for all students (e.g., race, gender, age, courses, GPA, etc.).
    - **Y-variable**: Students' satisfaction with the university.
    - **MCAR Condition**: Nonresponse is completely unrelated to any X-variable or the Y-variable.

Key Characteristic: Data are missing for reasons completely unrelated to the observed or unobserved variables.

MCAR: Key Implications

- Nonresponse is random: Nonrespondents are randomly distributed across the sampled units.
- Respondents are representative: The respondents provide an unbiased representation of the entire sample.
- Nonresponse can be ignored: Standard analyses can be performed without additional adjustments for nonresponse.
- **Limitations:**
  - True MCAR is rare in practice.
  - If MCAR does not hold, analyses assuming MCAR may lead to bias.

Practical Takeaway: Use MCAR assumptions only when there is strong evidence that missingness is unrelated to both observed and unobserved variables.

2. **Missing at Random (MAR):** The probability of responding, ϕ_i, depends on the observed variables $\mathbf{x}_i$, but not on the unobserved variable y_i.

- **Key Characteristics:**
  - Nonresponse is related to observed data but unrelated to the missing values themselves, conditional on $\mathbf{x}_i$.
  - Unlike MCAR, adjustments should be made to account for nonresponse using the observed data $\mathbf{x}_i$.
- **Example:**
  - In the National Crime Victimization Survey (NCVS), suppose:
    - The probability of responding to the survey depends on *observed factors* such as race, sex, and age (all known quantities).
    - However, within each race/sex/age group, the probability of response does not depend on whether the individual experienced victimization (unobserved).

MCAR and MAR: Ignorable Nonresponse

- Ignorable nonresponse means that the nonresponse mechanism can be explained and accounted for using a suitable model.
- Once the model accounts for the nonresponse, its effects can be "ignored" in the analysis.
- **Important:** "Ignorable" does not mean that the nonresponse can be completely disregarded; adjustments or modeling are still necessary to address potential bias.

3. **Nonignorable Nonresponse (Not Missing at Random - NMAR)**: nonresponse is considered **nonignorable** when the probability of responding (ϕ_i) is directly related to the outcome variable (y_i), even after accounting for known variables ($\mathbf{x}_i$).

- Example: In the NCVS (National Crime Victimization Survey), it is suspected that individuals who have been victims of crime are less likely to respond to the survey than nonvictims, regardless of shared characteristics such as race, age, or sex.
- **Key Challenge:** Nonresponse bias cannot be corrected using observed data alone; it requires additional assumptions or external information about the relationship between nonresponse and y_i.

Methods to Address Nonresponse

In survey sampling methodology, two primary methods are commonly used to address nonresponse:

- **Weighting:**
  - Adjusts the weights of responding units to compensate for nonrespondents.
  - Example: If younger participants are less likely to respond, their responses are given higher weights to represent their population segment adequately.
- **Imputation:**
  - Fills in missing values using observed data or predictive models.
  - Example: Estimate missing income values based on variables like age, education, and occupation.

Goal: Reduce bias introduced by nonresponse and improve the reliability of survey estimates.

Weighting class method

- Divide sample into classes based on variables known for everyone
- Assume MCAR within those classes i.e. assume that nonrespondents are similar to respondents within each weighting class

Example:

	Male	Female
R	10,000	12,000
NR	8000	5000
Sample	18,000	17,000

- Started one with $w_i = 100$ for everyone
- New weight w_i^* for each male respondent

$$w_i^* = w_i \times \frac{18,000}{10,000} = 180$$

each male respondent represents 180 men in population

- Females

$$w_i^* = w_i \times \frac{17,000}{12,000} = 141.7$$

—- Each female respondent now represents 142 females in the population.

—- Assumption: The distribution of y_i (the variable of interest) is similar for both responding and nonresponding women.

- **General Formula for Adjusted Weights:**

$$w_i^* = w_i \times \frac{\text{Sum of weights for the selected sample in class}}{\text{Sum of weights for respondents in class}}$$

—- Adjust weights within each response class (e.g., gender, age group) to account for nonresponse.

Weighting-class adjustment

- assumption: respondents and nonrespondents in the same weighting-adjustment class are similar
- weights of respondents in the weighting-adjustment class are increased so that the respondents represent the nonrespondents share of the population as well as their own

Estimate response propensity score for class c by

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

$$w_i^* = \sum_c \frac{w_i x_{ci}}{\hat{\phi}_c},$$

where $x_{ci} = 1$ if unit i is in class c

$$\hat{t}_{wc} = \sum_{i \in S} w_i^* y_i$$

$$\hat{\bar{y}}_{wc} = \frac{\hat{t}_{wc}}{\sum_{i \in S} w_i^*}$$

In an SRS,

$$\hat{\phi}_c = \frac{n_{cR}}{n_c}$$

$$\begin{aligned}\hat{t}_{wc} &= \sum_{i \in S} \sum_c \frac{N}{n} \frac{n_c}{n_{cR}} x_{ci} y_i \\ &= N \sum_c \frac{n_c}{n} \bar{y}_{cR}\end{aligned}$$

$$\hat{\bar{y}}_{wc} = \sum_c \frac{n_c}{n} \bar{y}_{cR}$$

Poststratification

- Similar to weighting class adjustment, except that population counts are used to adjust the weights
- SRS sample is collected
- Units are grouped into H different post strata based on demographic variables such as race and sex
- N_H units in poststratum, n_h units were selected for the sample and n_{hR} respond

Assumptions:

- within each stratum each unit selected to be in the sample has the same probability of being a respondent
- the response or nonresponse of a unit is independent of the responses or nonresponses of all other units
- nonrespondents in a poststratum are like the respondents Data are MCAR within each poststratum

The poststratified estimator for $\bar{y}_U$ is

$$\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hR} \tag{1}$$

The weighting-class estimator for $\bar{y}_U$, if the weighting classes are the poststrata, is

$$\bar{y}_{wc} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{hR} \tag{2}$$

- Estimate (1) and estimate (2) are similar in form
- Difference is that in poststratification the N_h are known, whereas in weighting-class adjustments the N_h are unknown and estimated by $N * n_h/n$

For the poststratified estimator, often the conditional variance given n_{hR} is used. For an SRS,

$$V(\bar{y}_{post}|n_{hR}, h = 1, \cdots, H) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_{hR}}{N_h}\right) \left(\frac{S_h^2}{n_{hR}}\right)$$

Poststratification Using Weights

Let

$$x_{hi} = \begin{cases} 1 & \text{if unit } i \text{ is a respondent in poststratum } h \\ 0 & \text{otherwise} \end{cases}$$

- The weight for each respondent adjusts for unequal selection probabilities and nonresponse within the poststratum.
- The sum $\sum_{i \in S_h} w_i \cdot x_{hi} = \sum_{i \in S_h} w_i$ estimates the population count N_h for that subgroup of respondents.

Poststratification uses a **ratio estimator within each subgroup** to adjust the sample estimates according to the known true population count for each subgroup.

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

$$\begin{aligned}
\sum_{h=1}^{H} \sum_{i \in S} w_i^* x_{hi} &= \sum_{h=1}^{H} \sum_{i \in S} w_i \frac{N_h}{\sum_{h=1}^{H} \sum_{j \in R} w_j x_{hj}} x_{hi} \\
&= \sum_{h=1}^{H} \sum_{i \in S} w_i x_{hi} \frac{N_h}{\sum_{h=1}^{H} \sum_{j \in R} w_j x_{hj}} \\
&= N_h
\end{aligned}$$

$$\hat{t}_{post} = \sum_{h=1}^{H} \sum_{i \in S} w_i^* x_{hi} y_i$$

$$\hat{\bar{y}}_{post} = \sum_{h=1}^{H} \sum_{i \in S} w_i^* y_i / N = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hR}$$

Raking adjustment

A poststratification method that can be used when poststrata are formed using more than one variable, but only the marginal population totals are known

Assumption

- within each poststratum, each unit selected to be in the sample has the same probability of being a respondent
- the response or nonresponse of a unit is independent of the behavior of all other units
- nonrespondents in a poststratum are like the respondents
- response probabilities depend only on the row and column and not on the particular cell

Example: have population information on two variables, but don't have cell counts information

	M	F	
Assistant professor			80
Associate professor			180
Full professor			540
	600	200	800

Sum of weights from survey

	M	F	
Assistant professor	40	60	100
Associate professor	80	60	140
Full professor	320	40	360
	440	160	600

Iteration 1 (adjust columns)

Multiply weight of males by 600/440, females by 200/160

	M	F	
Assistant professor	54.5	75	129.5
Associate professor	109.1	75	184.1
Full professor	436.4	50	486.4
	600	200	800

Iteration 2 (adjust rows)

multiply weight of assistant professor by 80/129.5

multiply weight of associate professor by 180/184.1

multiply weight of full professor by 540/486.4

	M	F	
Assistant professor	33.7	46.3	80
Associate professor	106.7	73.3	180
Full professor	484.5	55.5	540
	624.9	175.1	800

Iterate until the marginal counts are equal to the marginal counts of the population (the first table)

	M	F	
Assistant professor	29.6	50.4	80
Associate professor	97.3	82.7	180
Full professor	473	67	540
	600	200	800

Note:

- if some of the cell estimates are zero, may not converge
- overadjustment: if there is little relation between the extra dimension in raking and the cell means, raking may increase the variance

Some Comments on Weighting Adjustments

- Weighting class adjustments and poststratification can help reduce nonresponse bias.
- In each weighting cell or poststratum, respondents and nonrespondents are assumed to be similar.
  —- Specifically, each individual in a weighting class is assumed to have an equal likelihood of responding to the survey or a response propensity uncorrelated with y.
- Some practitioners treat the weighting adjustment as a complete remedy for nonresponse and proceed as though nonresponse no longer exists.
  —- This approach should be avoided, as it overlooks potential biases.

- Weighting can improve many estimates but rarely eliminates all nonresponse bias.
- When weighting adjustments are applied, practitioners should explicitly state the assumed response model and provide evidence to support it.
- Weighting adjustments are typically used for unit nonresponse and not for item nonresponse, which requires different weights for each item.
- Poststratification is a special case of calibration methods used in survey sampling.

Item Missing

Missing survey items can occur for various reasons:

- The interviewer may forget to ask a question.
- The respondent may refuse to answer or may not know the information.
- A data entry clerk might inadvertently skip the value.
- Sometimes, responses are marked as missing during data editing or cleaning:
  - For example, if discrepancies arise—such as a 3-year-old reported to have voted in the last election—both values (age and voting status) may be set to missing.

Imputation: Filling in Missing Items

Imputation is a common technique to address missing survey data:

- A replacement value is assigned to the missing item, often based on data from another respondent who is similar to the nonrespondent on other variables.
- To maintain transparency, an additional variable should be created in the data set to indicate whether a response was measured or imputed.
- Imputation helps:
  - Reduce nonresponse bias.
  - Create a "clean," rectangular data set without gaps for missing values.

Example

```
#data impute
      person age gender education crime violcrime
 1         1    47 M            16     0         0
 2         2    45 F            NA     1         1
 3         3    19 M            11     0         0
 4         4    21 F            NA     1         1
 5         5    24 M            12     1         1
 6         6    41 F            NA     0         0
 7         7    36 M            20     1        NA
 8         8    50 M            12     0         0
 9         9    53 F            13     0        NA
10        10    17 M            10    NA        NA
```

	person	age	gender	education	crime	violcrime
11	11	53	F	12	0	0
12	12	21	F	12	0	0
13	13	18	F	11	1	NA
14	14	34	M	16	1	0
15	15	44	M	14	0	0
16	16	45	M	11	0	0
17	17	54	F	14	0	0
18	18	55	F	10	0	0
19	19	29	F	12	NA	0
20	20	32	F	10	0	0

Deductive Imputation

Deductive imputation uses logical relationships between variables to fill in missing values:

- Values are imputed during data editing based on known information.
- Example 1: A respondent person 9 in the "impute" data is missing the response for being a victim of violent crime, but she reported not being a victim of any crime. The logical imputed value for violent crime would be 0.
- Example 2: In a longitudinal survey, a woman reports having two children in year 1 and year 3, but the value is missing for year 2. The logical imputed value for year 2 would be two.
- Deductive imputation is particularly useful in scenarios where strong logical or temporal relationships exist among variables.

Cell Mean Imputation

- Respondents are grouped into classes (cells) based on known variables.
- The mean value of the responding units in each cell, $\bar{y}_{cR}$, is used to replace missing values within that cell.
- Assumes missing items are missing completely at random (MCAR) within each cell.
  Example:
  - The mean annual income (y) for black males aged 19–24 is \$19,242.
  - Impute \$19,242 for every black male aged 19–24 who has a missing income value.
- **Limitations:** - Can create artificial spikes in the data (e.g., in histograms). - Distorts multivariate relationships by ignoring individual variability within cells.

Example 8.9, page 336

Table 1: The four cells are constructed using variables age and sex.

	≤ 34	≥ 35
M	Persons 3,5,10,14	Persons 1,7,8,15,16
F	Persons 4,12,13,19,20	Persons 2,6,9,11,17,18

- Persons 2 and 6, who have missing values for years of education, would be assigned the average value of 12.25, calculated from the four women aged 35 or older who responded to this question.

Hot deck imputation : impute value from record in the data set

- **Sequential Hot-Deck Imputation:** impute the value in the same subgroup that was last read by the computer.

Table 2: The four cells for our example are constructed using the variables age and sex.

	≤ 34	≥ 35
M	Persons 3,5,10,14	Persons 1,7,8,15,16
F	Persons 4,12,13,19,20	Persons 2,6,9,11,17,18

In this example, person 19 is missing the response for crime victimization. Person 13 had the last response recorded in her subclass, so value 1 is imputed.

- **Random Hot-Deck Imputation:** A donor is randomly chosen from the persons in the subgroup with information on all the missing items. To preserve multivariate relationships, usually values from the same donor are used for all missing items of a person.
  Example: In "impute" data set, person 10 is missing both variables for victimization. Persons 3, 5, and 14 in his cell have responses for both crime questions, so one of the three is chosen randomly as the donor. In this case, person 14 is chosen, and his values are imputed for both missing variables.

- **Nearest neighborhood Hot-Deck Imputation:** Define a distance measure between observations, and impute the value of a respondent who is closest to the person with the missing item, where closeness is defined using the distance function.

Example: If age and sex are used for the distance function, so that the person of closest age with the same sex is selected to be the donor, the victimization responses of person 3 will be imputed for person 10.

Regression Imputation

- Build a regression model to predict missing values, e.g., victimization responses, using other variables.
- Impute the predicted values from the regression model.
- Cell mean imputation is a special case of regression imputation.
- May result in spikes in the data distribution.

Example: Logistic Regression Imputation

- We have only 18 complete observations for the response variable crime victimization (not sufficient to fit a robust model), but a logistic regression with age as the explanatory variable provides the following model for predicted probability of victimization:

$$\log \frac{\hat{p}}{1-\hat{p}} = 2.5643 - 0.0896 \times \text{age}$$

- For a 17-year-old, the predicted probability of being a crime victim is 0.74.
- Since 0.74 exceeds the predetermined cutoff of 0.5, the value 1 is imputed for Person 10.

Multiple imputation

In multiple imputation, each missing value is imputed m (≥ 2) different times.

—-Typically, the same stochastic model is used for each imputation.

—-This create *m* different "data" sets with no missing values.

—-Each of the *m* data sets is analyzed as if no imputation had been done;

—-The different results give the analyst a measure of the additional variance due to the imputation.

—- Multiple imputation with different models for nonresponse can give an idea of the sensitivity of the results to particular nonresponse models. (Rubin (1987, 1996, 2004))

Multiple Imputation

Multiple imputation is a statistical technique used to handle missing data by imputing each missing value m ($m \geq 2$) different times, generating multiple complete datasets.

- Typically, the same stochastic model is used for each imputation.
- This process creates m complete datasets with no missing values.
- Each of the m datasets is analyzed independently as if no imputation had been done.
- The variation across the m results provides a measure of the additional uncertainty (variance) introduced by the imputation process.
- Using multiple imputation with different nonresponse models allows researchers to assess the sensitivity of results to specific assumptions about the nonresponse mechanism.

- Multiple imputation is widely regarded as a robust approach for handling missing data, especially when data are missing at random (MAR).

(See Rubin, 1987, 1996, 2004 for foundational work on multiple imputation.)

Problems with All Imputation Methods

- Imputation creates synthetic (or "fake") data, which may not fully represent the true missing values.
- It is essential to flag which values are imputed to distinguish them from observed data.
- Variance estimates tend to be underestimated:
  - The effective sample size is artificially increased due to the inclusion of imputed values.
  - Imputed data are often treated as though they were collected during the actual data collection process, ignoring the uncertainty associated with the imputation process.

Disclaimer

Slides are intended for a course based on the book: *Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr
All examples, datasets, and pages directly scanned and included in this chapter are from the textbook.
References to papers or books cited in these slides can be found in the Bibliography section of the textbook.