HANDLING MISSING DATA

David Haziza and Karla Nobrega

SSC -2002, Hamilton

Outline

- Introduction
  - COFFEE
- Types of Non-response
- Solutions:
  - Follow-up
  - Change the Context
    - LUNCH
  - Re Weighting
  - Analytical Techniques
    - COFFEE
  - Single Imputation
  - Multiple Imputation
- Evaluation

Data Imperfections / Errors / Missing Data

- Poor study methodology/design
- Human error
- Refusal
  - to answer the survey (questionnaire design)
    - partial nonresponse
    - total nonresponse
    - inconsistent response
  - to undergo tests/treatment (non-compliance)
- Lack of protocol adherence (partial compliance)
- Lost to follow-up
- Chance

Does the missing data affect my ability to answer the analytical questions I have?

If so, how?
and
How can I address it?

Assessing the Impact of Missing Data

What is the final output of the study?

How much data are missing?

How are the data missing?

- **Missing Completely at Random (MCAR)**
- **Missing at Random (MAR)**
- **Non Ignorable (NI)**

Context Continued

Where are the data missing?

- Total Non-response
- Partial Non-response

What type of study is it?

- Cross-sectional
- Longitudinal

Context Continued

Why are the data missing?

- **Design (MCAR because of design)**
- **Subject Matter**
- **Restrictions on Data Access / Confidentiality**

Impact of Missing Data

- Biased Estimate(s)
- Incorrect Variance(s)
- Changing Sample Population

Bias

A systematic distortion of the outcome we want to measure

- The nature of the bias depends upon the missing mechanisms
  - No bias when data is MCAR (Random)
  - Possible bias with MAR and NMAR (systematic)
- The direction of the bias is related to the variables under study
  - Income, Drug Use, Weight, HIV

Variance

In general the variance of estimates with missing data is underestimated

Changing Sample Size

Using available cases will result in different samples depending upon which variables are used in the analysis

MI (Y/N) = f(AGE, GENDER, SMOKING, BMI, ACTIVITY, HYPERTENSION, LDL, HDL)

MI (Y/N) = f(AGE, GENDER, SMOKING, BMI, ACTIVITY, HYPERTENSION)

Solutions

The only really good solution to the missing data problem is not to have any

– Paul Allison, Missing Data

Solutions for Missing Data

- Imputation
- Re-weighting
- Follow-up
- Change the Context of the Analysis
- Analytical Techniques

Imputation

Filling in the missing data

- Mean
- Ratio
- Hot Deck
- Nearest-Neighbor
- Regression
- Historical
- Multiple
- Allocate

Imputation

- Creates a ‘complete’ data base for analytical use
- Can be taken into account when calculating the variance
- Best for partial non-response
- Requires a good knowledge of the variables in the data set
- May require complex computing programs
- May require auxiliary or partial information
- Can destroy the multivariate structure in the data

Re-Weighting

Changing the relative distribution of observations in the data set to reflect the relative distribution in the population

Re-weighting

- More convenient for total non-response
- Can be taken into account when calculating variances
- Can maintain the multivariate structure of the data
- Requires calculation of the weights
- May require more complex variance calculations
- Cross-sectional and longitudinal weights are different
- Weights for item non-response will be different

Follow-Up

Try to re-contact non-respondents to minimize the non-response

Try to follow non-respondents with other data sources

- **Record Linkage**

Follow-Up

- Completes the data set with the actual respondents data
- Maintains the multivariate structure of the data
- Costly (could be reduced by just getting key information)
- Time consuming
- Often not possible unless you were collecting the data

Change the Context of the Analysis

Frame your analytical questions to reflect only the complete data that you have

- **Listwise Deletion**
- **Pairwise Deletion**
- **Intent to Treat vs. Complete Case Analysis**

Change the Context of the Analysis

- Simple
- Coverage of the sub-population is generally good

- Conclusions are not generalizable
- Conclusions may not be comparable across studies
- May not be the sub-population we wanted to study

Analytical Techniques

Analyze the incomplete data with techniques that compensate/handle missing data

- Missing is a category
- Maximum Likelihood
  - Missing data patterns
  - EM algorithm
- Censoring
- Monte Carlo Methods
  - Gibbs Sampling

Analytical Techniques

- Can compensate for some types of missing data directly
- Can maintain the multivariate structure of the data
- Can be taken into account when calculating variances
- Can be complicated
- May require specialized software
- May need baseline information

Which Solution??

THE SOLUTION FOR YOUR MISSING DATA MUST FIT YOUR MISSING DATA PROBLEM

Longitudinal Studies

Longitudinal Studies

Observational

- Case-control
- Cohort
- Survey

Experimental

- Clinical Trials

(Laird, 1988 - Statistics in Medicine)

Case-Control

- Cases are selected on the basis of their disease status
- Controls are selected from a reference population

★ Present

☆ Absent

? To be determined

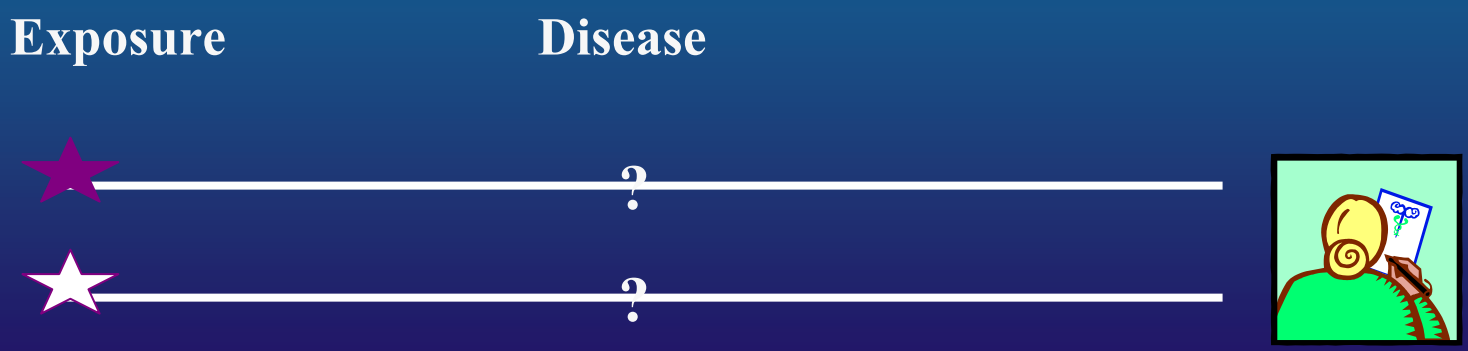
Investigator at the beginning of the study

Cohort

- **Population is chosen based on exposure status**

Cohort is followed prospectively

or retrospectively



Clinical Trial

- **Patients are exposed to an intervention and followed**

Bias from Data Imperfections

- **Post-entry exclusion**
  - Results no longer correspond to the target population
- **Selective loss of data**
  - Various effects
- **Assessment bias**
  - Strength of effect not correctly measured
- **Retroactive definitions**
  - Strong anti-conservative bias

Survey

- **Questionnaire is administered at more than one time point to the same individuals**

Cross-sectional Studies

SURVEYS

Survey steps

- Planning: objective, concepts, definitions, target population,...
- Survey frame
- Sampling design and sample selection
- Questionnaire design
- Data collection
- Data processing: editing, imputation,…
- Point and variance estimation

SURVEYS

Survey steps (cont.):

- Evaluation of data quality
- Confidentiality treatment
- Data analysis
- Documentation
- Data release and publication of analysis results

DEFINITIONS

LEVELS OF NONRESPONSE

1. Total or unit nonresponse:

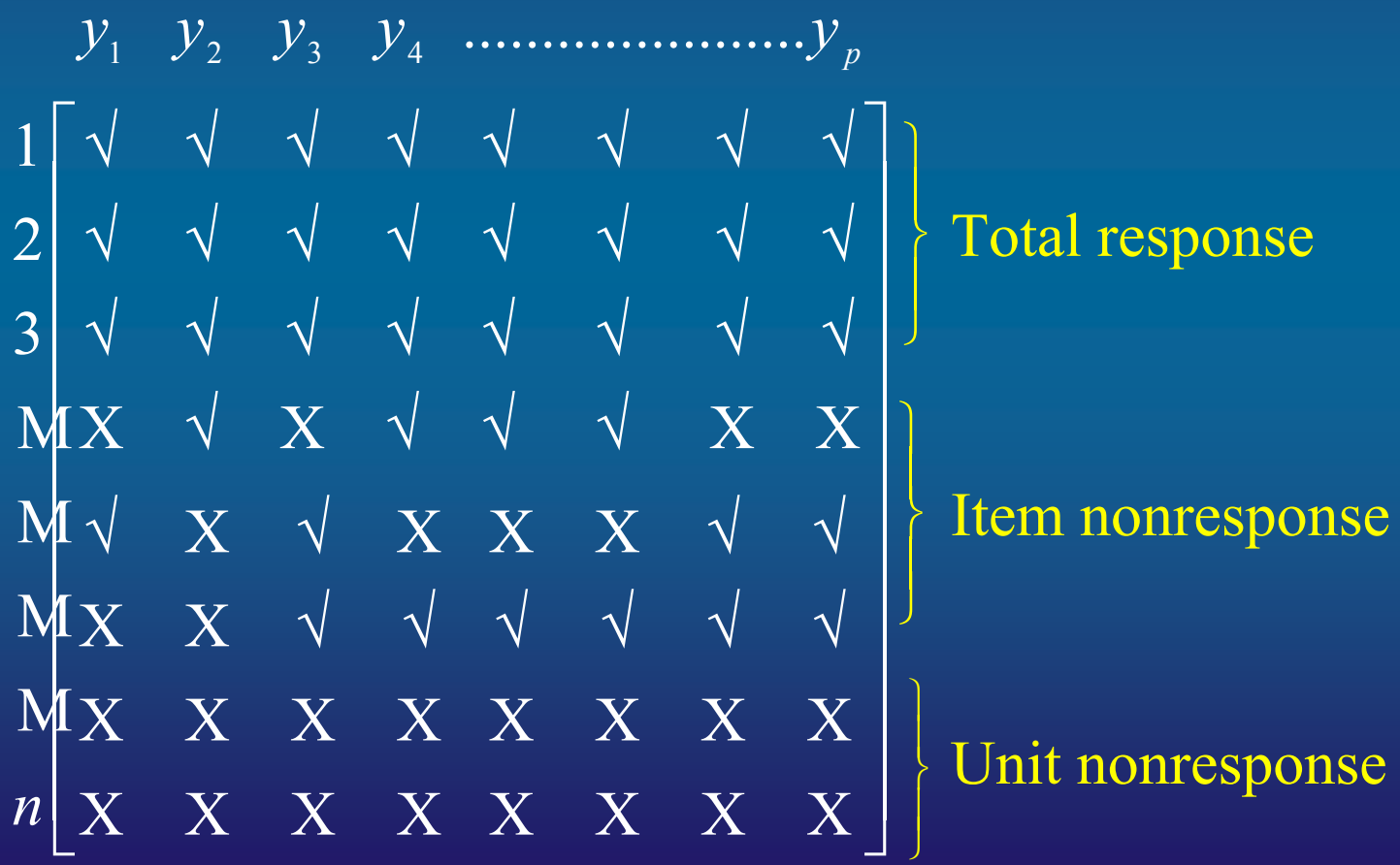
- No usable information was collected on a sampled unit

2. Partial or item nonresponse:

- Some (but not all) variables were collected

DEFINITIONS

LEVELS OF NONRESPONSE



CAUSES OF NONRESPONSE

1. Unit nonresponse:

- **Wrong contact information**
- **Respondent is absent**
- **Refusal**
- **Move**
- **Language problem**
- **Lost questionnaire**
- **Response burden too high**
- **Survey perceived not to be important**
- **Tight budget**
- **Timeliness**
- **Mandatory vs voluntary**

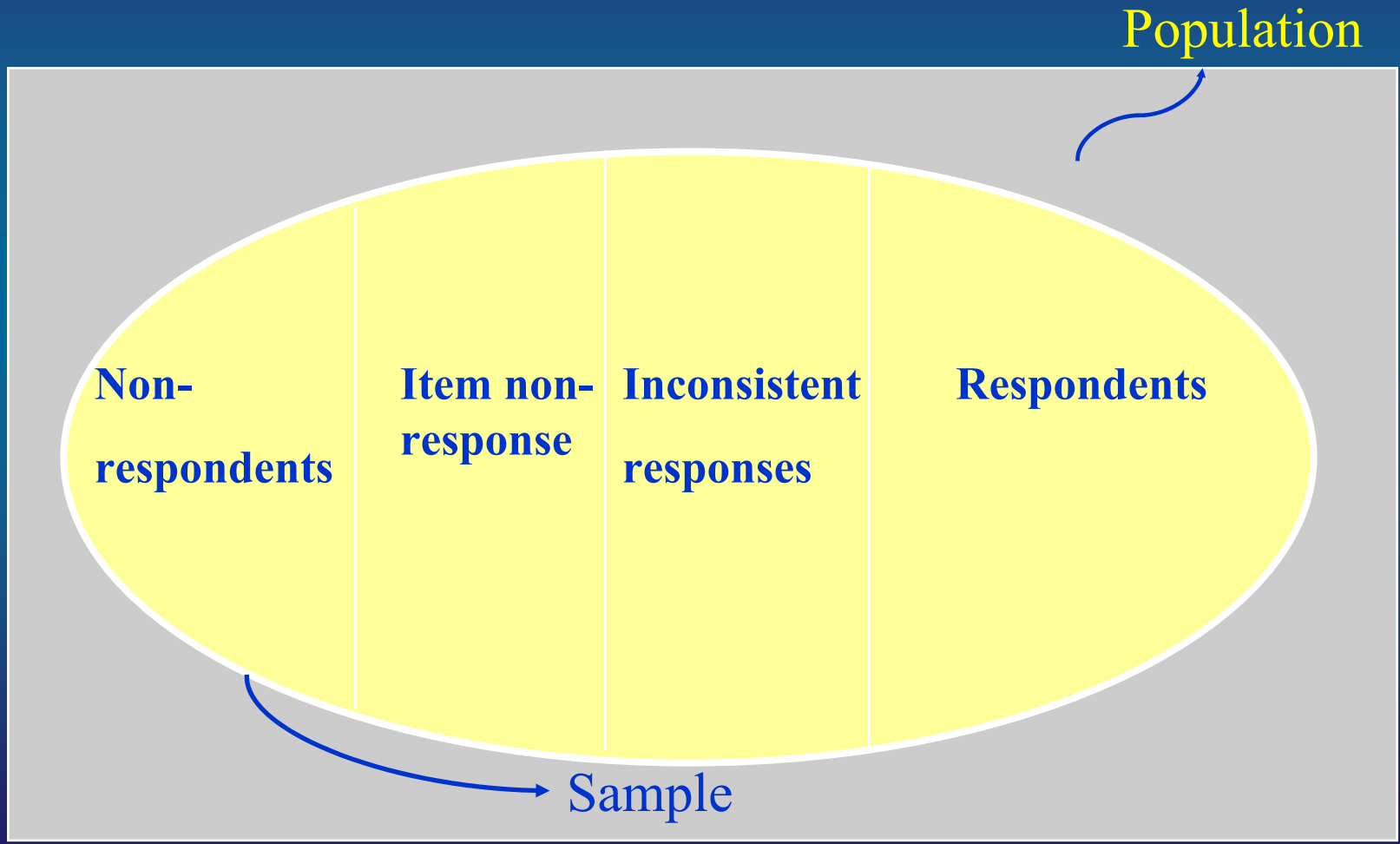
CAUSES OF NONRESPONSE

2. Item nonresponse:

- Question not understood
- Refusal
- Don't know
- Question forgotten by the interviewer
- Data not available

- Inconsistent or unusable response → (impossible response, question wrongly understood, question wrongly asked, response cannot be read, edits not satisfied, lost data…)

CAUSES OF NONRESPONSE



EXAMPLES OF NONRESPONSE

Example: Canadian Labour Force Survey (LFS)

- Not at home
- Respondent refuses to answer questions on income
- Unit in the sample for too long

Example: Survey of Labour and Income Dynamics

- Sampled units refuse to paticipate to the survey after they have participated to the LFS
- Questions relative to income are sensitive
- Tracing of sampled units may be complex

FULL RESPONSE THEORY

- Let P be a finite population of size N (N may be unknown)

$$\{1,2,...,i,...,N\}$$

- The goal is to estimate some parameters of the finite population such as
  - total or means
  - domain means and totals
  - ratio of two variables of interest
  - median
  - variance, regression coefficients and coefficients of correlation, etc.

FULL RESPONSE THEORY

• We select a random sample from P according to a sampling design $p(.)$ such that

1. $p(s) \geq 0 \quad \forall s \in S$
2. $\sum_{s \in S} p(s) = 1$

where S is the set of all possible samples and $p(s)$ is the probability of selecting s.

Examples of sampling designs: simple random sample, Poisson sampling, systematic sampling, stratified multistage sampling, etc.

FULL RESPONSE THEORY

Typical variables in a survey

		Auxiliary variables $\boldsymbol{z}$			Variables of interest $\boldsymbol{y}$			Sample indicator	Response indicators		
		1	…	q	1	…	p		1	…	p
Units In the population	1	z_{11}	…	z_{1q}	y_{11}	…	y_{1p}	I_1	a_{11}	…	a_{1p}
	.										
	.										
	.										
	i	z_{i1}	…	z_{iq}	y_{i1}	…	y_{ip}	I_i	a_{i1}	…	a_{ip}
	.										
	.										
	.										
	N	z_{N1}	…	z_{Nq}	y_{i1}	…	y_{Np}	I_N	a_{N1}	…	a_{Np}

FULL RESPONSE THEORY

- In practice, we usually consider two types of estimator
  - The Horvitz-Thompson (HT) estimator
  - The Generalized regression (GREG) estimator

FULL RESPONSE THEORY

The Horvitz-Thompson estimator

- Let $Y = \sum_{i \in P} y_i$ be a population total for variable of interest y.
- In order to estimate Y, we select a random sample, s, of size n, according to a sampling design $p(.)$ and we observe $y_1,...,y_n$.
- An unbiased estimator of Y, denoted by $\hat{Y}_{HT}$, is given by

$$\hat{Y}_{HT} = \sum_{i \in s} w_i y_i$$

where w_i denotes the sampling weight attached to sampled unit i,

and $w_i = 1/\pi_i$ where π_i denotes the first-order probability of inclusion for sampled unit i, $i = 1,...,n$.

FULL RESPONSE THEORY

The Horvitz-Thompson estimator

- We have $\pi_i = P(i \in s) = P(I_i = 1) = E_p(I_i)$

  where $I_i = \begin{cases} 1 & \text{if unit } i \text{ is selected in the sample} \\ 0 & \text{otherwise} \end{cases}$

  $= \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i \end{cases}$

Example: In the case of a simple random sample without replacement, we have $\pi_i = \frac{n}{N} \ \forall i \in s$. Hence, $\hat{Y}_{HT} = N/n \sum_{i \in s} y_i$.

Result: $E_p(\hat{Y}_{HT}) = Y$; that is, the HT estimator is design-unbiased for Y, where $E_p(.)$ denotes the expectation with respect to the sampling design.

FULL RESPONSE THEORY

The GREG estimator

- In pratice, auxiliary iformation is often available at the estimation stage. This information may be wisely used in order to construct more efficient estimators.
- We assume that a vector of q auxiliary variables $\mathbf{z}_i = (z_{1i}, z_{2i}, \ldots, z_{qi})$ is available for all units in the sample.
- Further, we assume that the population totals $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_q)$ for the auxiliary variables are available

where $$Z_k = \sum_{i \in P} z_{ki}, \quad k = 1, \ldots, q.$$

FULL RESPONSE THEORY

The GREG estimator

- We generally assume that the relation between the variable of interest y and the auxiliary variables is of the form

$$m: y_i = \mathbf{z_i'\beta} + \varepsilon_i$$

$$E_m(\varepsilon_i) = 0, V_m(\varepsilon_i) = \sigma_i^2, E_m(\varepsilon_i \varepsilon_j) = 0 \text{ if } i \neq j$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_q)$ is a vector a unknown parameters and σ_i^2 is also an unknown parameter.

- Then,

$$Y = \sum_{i \in P} y_i = \sum_{i \in P} (\mathbf{z_i'\beta} + \varepsilon_i) = \sum_{i \in P} \mathbf{z_i'\beta} + \sum_{i \in P} \varepsilon_i, \text{ where } \varepsilon_i = y_i - \mathbf{z_i'\beta}.$$

FULL RESPONSE THEORY

The GREG estimator

- The generalized regression estimator (GREG) of Y, denoted $\hat{Y}_{GREG}$, is given by

$$\hat{Y}_{GREG} = \sum_{i \in P} \mathbf{z}_i' \hat{\boldsymbol{\beta}} + \sum_{i \in s} w_i e_i = \hat{Y}_{HT} + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{HT}\right)' \hat{\boldsymbol{\beta}} = \sum_{i \in s} w_i g_i y_i$$

where $e_i = y_i - \mathbf{z}_i' \hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / \sigma_i^2 \right)^{-1} \sum_{i \in s} w_i \mathbf{z}_i y_i / \sigma_i^2$,

$\hat{\mathbf{Z}}_{HT} = \sum_{i \in s} w_i \mathbf{z}_i$

and $g_i = 1 + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{HT}\right)' \hat{\mathbf{T}}^{-1} z_i / \sigma_i^2$ with $\hat{\mathbf{T}}^{-1} = \sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / \sigma_i^2$.

FULL RESPONSE THEORY

The GREG estimator

- The GREG estimator is approximately design-unbiased for Y; that is,

$$E_p\left(\hat{Y}_{GREG}\right) \approx Y.$$

- The model simply justifies the form of the estimator and the properties of the estimator are still evaluated with respect to the sampling design (model-assisted approach)
- The GREG estimator remains approximately design-unbiased, even if the model does not fit the data correctly. In this case however, we expect the variance of the GREG estimator to be large.

FULL RESPONSE THEORY

Two-phase sampling

- Sometimes, the survey frame contains very little or no information at all
- In this case, two-phase sampling allows for the use of auxiliary information which lead to more efficient estimators.

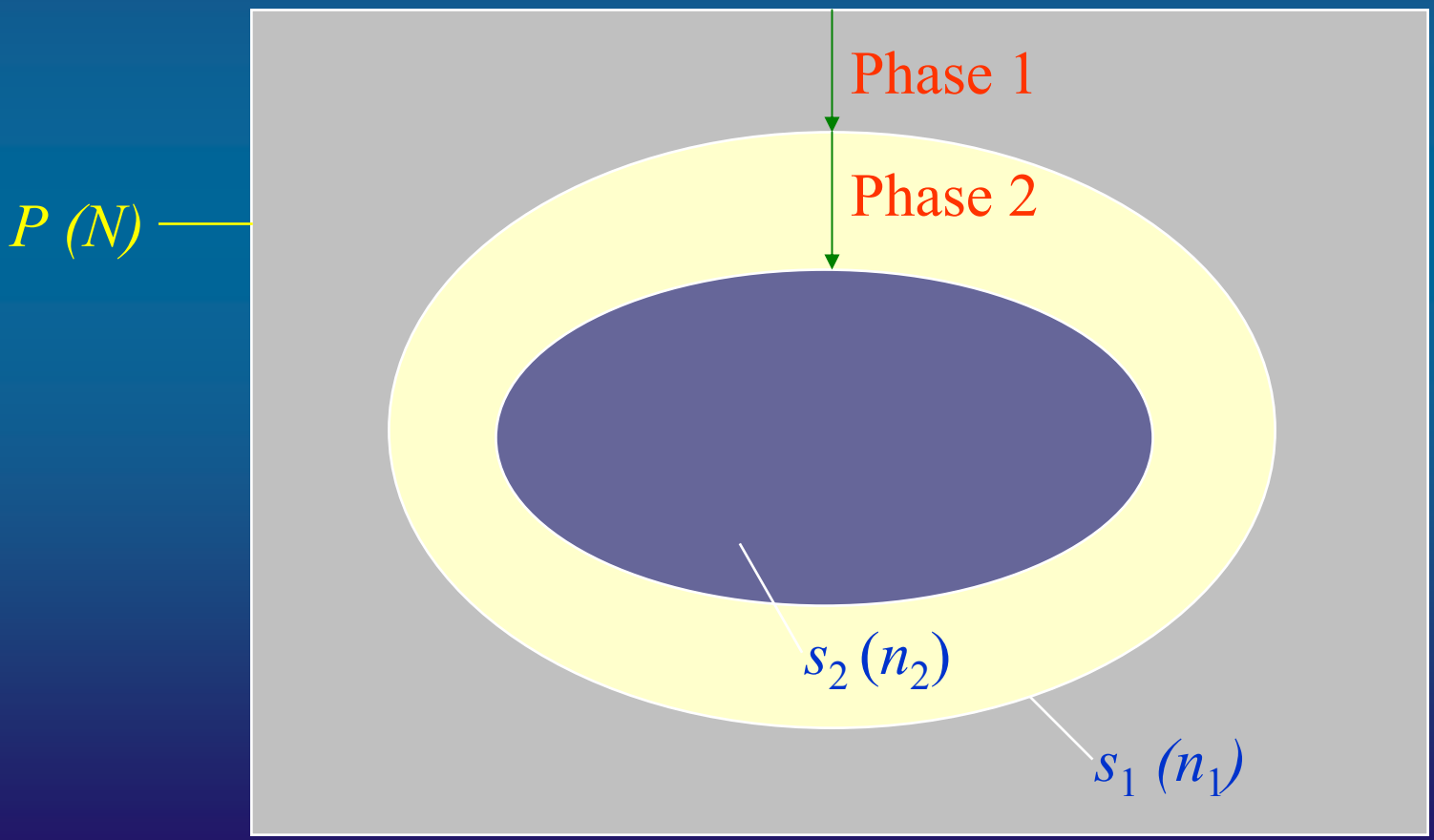
– Phase 1: A first sample s_1 is selected according to some sampling design $p_1(s_1)$.

– Phase 2: A subsample s_2 of s_1 is then selected according to the design $p_2(s_2|\, s_1)$.

– Idea: The first phase will allow to collect cheap auxiliary information that will be used to improve the quality of the estimators produced using the data in s_2.

FULL RESPONSE THEORY

Two-phase sampling



FULL RESPONSE THEORY

Two-phase sampling

- We have $\pi_{1i} = P(i \in s_1)$ and $\pi_{2i} = P(i \in s_2 \mid i \in s_1)$
- Let $w_{1i} = 1/\pi_{1i}$ and $w_{2i} = 1/\pi_{2i}$
- Let $Y = \sum_{i \in P} y_i$ be the parameter of interest
- In the two-phase sampling context, a design-unbiased estimator of Y (that does not use auxiliary information from the first-phase) is given by $\hat{Y}_{TP} = \sum_{i \in s} w_{1i} w_{2i} y_i$
- Indeed, $E(\hat{Y}_{TP}) = E_1 E_2 (\hat{Y}_{TP} \mid s_1) = Y$

FULL RESPONSE THEORY

Two-phase sampling

- The variance of $\hat{Y}_{TP}$ is given by

$$V\left(\hat{Y}_{TP}\right) = \underbrace{V_1 E_2\left(\hat{Y}_{TP} \mid s_1\right)}_{\text{Variance due to phase 1}} + \underbrace{E_1 V_2\left(\hat{Y}_{TP} \mid s_1\right)}_{\text{Variance due to phase 2}}$$

- Nonresponse is often viewed as a second phase of sampling.
- In the presence of nonresponse, we have

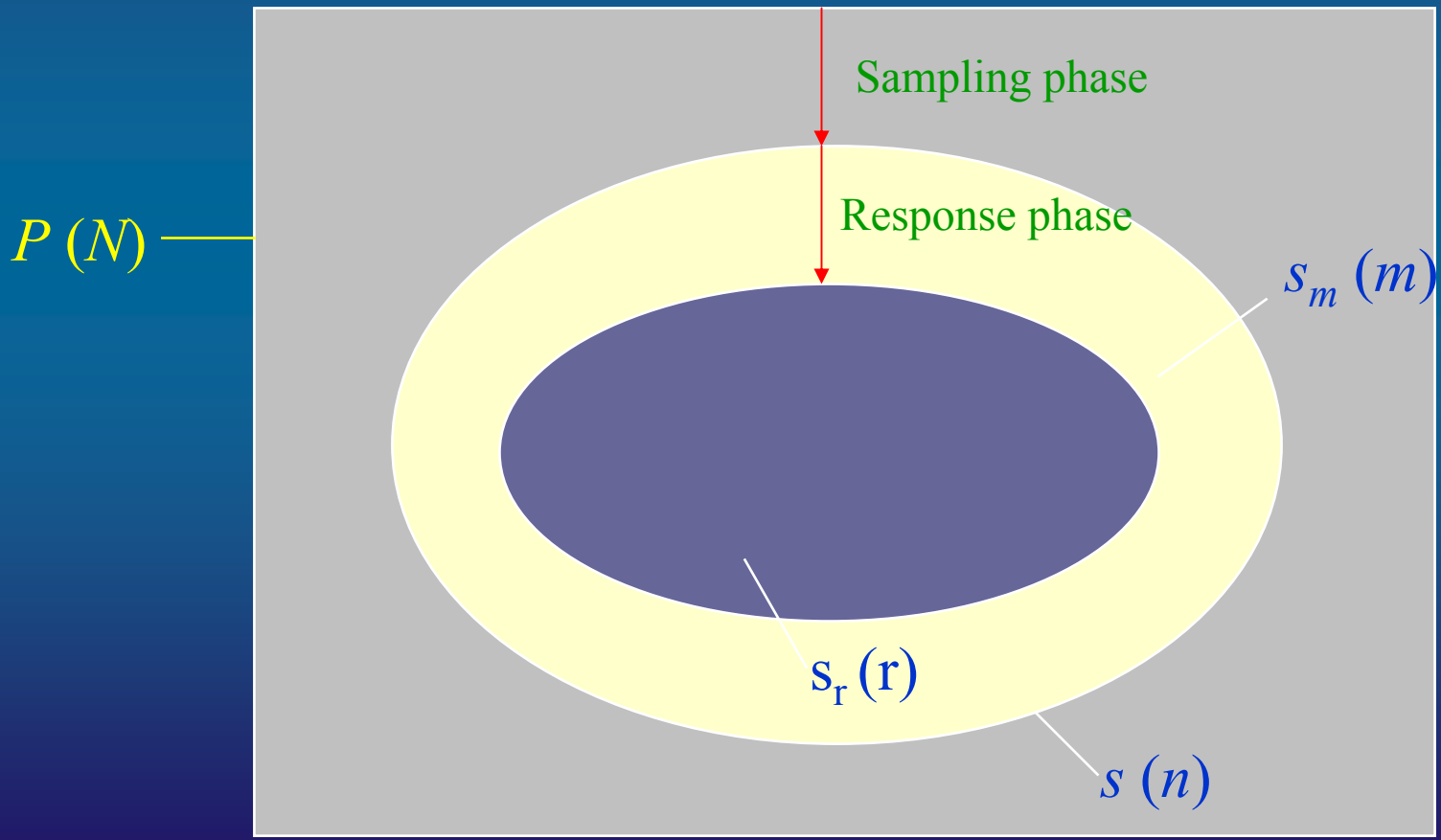
$$\pi_{2i} = P\left(i \in s_2 \mid i \in s_1\right) = p_i$$

where p_i denotes the probability of response for unit i.

- Note that the p_i's are typically unknown.

FULL RESPONSE THEORY

Nonresponse viewed as the second phase of sampling



COFFEE

TYPES OF NONRESPONSE

- As we have seen, nonresponse is often seen as the second phase of sampling.
- However, in practice the response mechanism is unknown.
- Then, we do not have the choice but to make assumptions on the response mechanism.

TYPES OF NONRESPONSE

• To better understand the notion of bias due to nonresponse, consider the following example.

Example: Suppose that the variable of interest is the variable *EARNING*. Suppose that high-income people have less tendency to respond than others. If we use the mean of the respondents, the estimates obtained are likely to underestimate the population mean for the variable *EARNING* .

• In the preceeding example, the bias comes from the fact that the response mechanism has not been taken into account to obtain the estimate.

TYPES OF NONRESPONSE

• Incorporating appropriate auxiliary variables may help to reduce significantly the nonresponse bias.

• Assume that $a_i \overset{ind}{\sim} B(1, p_i), i = 1,\ldots, N.$

• Statisticians generally distinguish between 3 types of response mechanisms:

(i) uniform (MCAR)

(ii) ignorable (MAR)

(iii) nonignorable (NMAR)

TYPES OF NONRESPONSE

Uniform response mechanism (MCAR)

1. A response mechanism is uniform if $p_i = p$ for all units in the population.
2. When the response mechanism is uniform, the data are said to be Missing Completely at Random (MCAR).
3. This mechanism is not realistic in practice. However, the uniform assumption can be relaxed by forming classes and assume uniform-within-classes.

TYPES OF NONRESPONSE

Uniform response mechanism

Résult: Let P be a population of size N, s be a simple random sample without replacement of size n selected from P and s_r be the subsample of respondents of size r. Suppose that the response mechanism is uniform; that is $P(i \in s_r) = p$. Then, conditionnally on s and r, s_r is a simple random sample without replacement selected from s; that is,

$$P(s_r \mid s, r) = \frac{1}{\binom{n}{r}}$$

TYPES OF NONRESPONSE

Ignorable response mechanism

1. On one hand, we have a model for the variable of interest y which describes the distribution of this variable conditional on a vector of auxiliary variables $\mathbf{z}_1$, $f(\mathbf{y} \mid \mathbf{z}_1; \boldsymbol{\beta})$. A model frequently used in pratice is given by

$$m_1 : y_i = \mathbf{z}'_{1i}\boldsymbol{\beta} + \varepsilon_i,$$
$$E_m(\varepsilon_i) = 0, V_m(\varepsilon_i) = \sigma_i^2, Cov_m(\varepsilon_i, \varepsilon_j) = 0 \quad \text{si } i \neq j$$

1. On the other hand, we have a response model that describes the distribution of the response indicators conditional on a vector of auxiliary variables $\mathbf{z}_2$, $f(a \mid \mathbf{z}_2; \boldsymbol{\gamma})$. A model frequently used in practice is given by

$$m_2 : \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{z}'_{2i}\boldsymbol{\beta}.$$

TYPES OF NONRESPONSE

Ignorable response mechanism

1. The response mechanism is ignorable with respect to the model m_1 if

   (i) the probability of response is independent of the errors of the model m_1

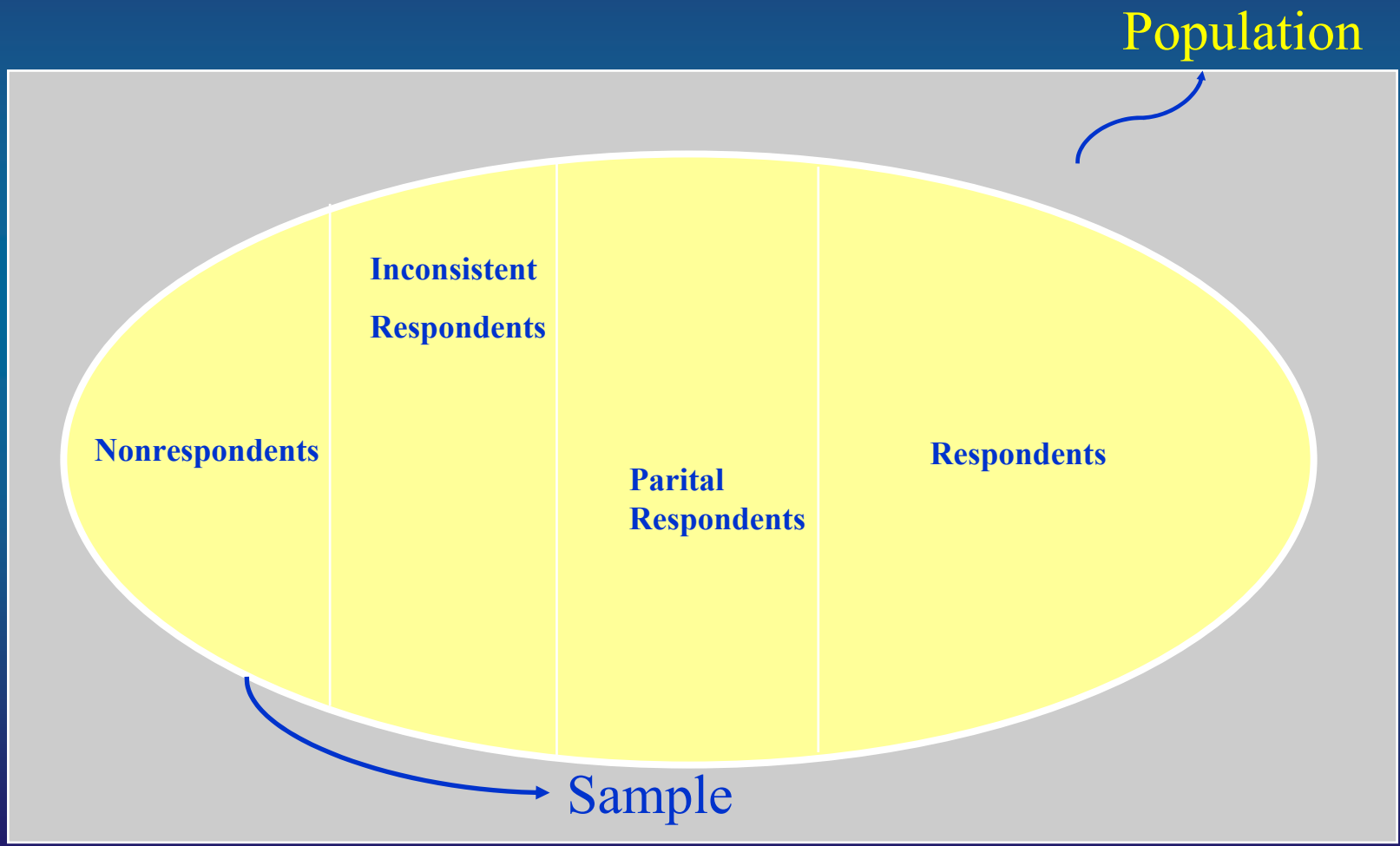
   (ii) The model parameters β et γ are distincts

1. A response mechanism is said to be nonignorable if it is not ignorable

TYPES OF NONRESPONSE

1. When a response mechanism is ignorable, the data are said to be Missing At Random (MAR)
2. When a response is nonignorable, the data are said to be Not Missing At Random (NMAR)
3. Of course, if the probability of response depends on the variable of interest, then the response mechanism is nonignorable. In this case, estimates are likely to be biased.
4. Hence, in order to reduce the nonresponse bias, it is important to include all the auxiliary information (if appropriate) in our model, especially the variables that are correlated with the probability of response.

Survey


Population
Inconsistent
Respondents
Nonrespondents
Parital
Respondents
Respondents
Sample

Solutions

Follow-up

Follow-up

- Prevention!
- Follow-up
- Linkages

Prevention !

- **Development / Planning**
  - realistic, clear, time-frame
- **Design (Study / Survey / Experiment)**
  - sample size, coverage, sub-samples
- **Questionnaire Design / Chart Abstraction / Data Collection**
  - simple, appropriate length, collection mode
- **Study Protocols**
  - adherence, adverse side effects, sub-population issues
- **Collection**
  - refusals, recruitment strategies, training, supervision

Re-contact / Follow-up

- Follow-up should be part of the initial budget and overall plan
- Allow for enough time in the collection period to have adequate follow-up
- Prioritize follow-up so that the most important information is obtained

Linkages

Probabilistic or Hierarchical Record Linkage (Exact Matching)

Assumptions - Hierarchical Record Linkage

- The two files contain the same individuals
- Error = incorrectly create a pair

Assumptions - Probabilistic Record Linkage

- Incomplete information and subject to error
- Randomly identical records
- Independent Errors (Fellegi-Sunter)

Changing the Context of the Analysis

Changing the Context of the Analysis

- Listwise Deletion / Complete Case Analysis
- Pairwise Deletion
- Intent To Treat (ITT) vs Treatment Received vs Adherers

Listwise Deletion

- Delete all observations that have any missing values of Y or X in the analysis
  - Can be used with any statistical analysis
  - No special computational methods are required
- Under MCAR
  - parameter estimates are unbiased
  - standard errors are larger (less power due to smaller sample size)
- Under MAR
  - can be biased

Pairwise Deletion

- Delete observations that are not there for each pair of variables, then calculate means or covariances and using these, estimate coefficients

- Under MCAR
  - estimates are consistent (approximately unbiased)
  - standard errors are biased
- Under MAR
  - estimates are biased
  - standard errors are biased
- Covariance matrix may not be positive definite

Clinical Trials - Missing Data

- **Protocol related**
  - **Evaluability criteria**
  - **Retroactive definitions**
  - **Ineligible patients**
- **Unrecorded data**
  - **Censoring**
- **Treatment non-adherence**
  - **ITT**
  - **Treatment Received**
  - **Adherers Only**

MISSING DATA SOLUTION MUST BE PART OF THE STUDY DESIGN (ANALYSIS PLAN)

ITT vs Treatment Received vs Adherers Only

- **Intent To Treat (ITT)**
  - Subjects in a randomized clinical trial should be analyzed according to the treatment group that they were assigned, even if they do not receive treatment or only received a portion of it
- **Treatment Received**
  - Subjects in a randomized clinical trial should be analyzed according to the treatment that they actually received
- **Adherers Only**
  - Discard all patients who did not comply with their treatment assignment

Hypothesis Testing

- **Type I error:**
  - **p-value is not a statement about the true effect, but about estimates the might be obtained under the null hypothesis**
  - **ITT yields the most desirable properties**

- **Type II error:**
  - **the power of the test is the chance of declaring a treatment effect of difference of a given size to be statistically significantly different from the null hypothesis value when the alternative hypothesis is true**
  - **Treatment Received yields the most desirable properties**

LUNCH

REWEIGHTING

- Reweighting methods are generally used to treat unit nonresponse. Reweighting consists in:
  - Delete units with missing values and
  - Adjust the sampling weights of respondent units to compensate for those that have been deleted.
- Let $p_i = P(a_i = 1 \mid s, i \in s) = P(i \in s_r \mid s, i \in s)$
- The p_i's are unknown. To cope with this problem, we model p_i and we estimate it by $\hat{p}_i$.
- We then adjust the sampling weights w_i as follows:

$$w_i^* = w_i (1/\hat{p}_i)$$

REWEIGHTING

- In the case of reweighting, an estimator of a population total Y is given by

$$\hat{Y}^* = \sum_{i \in s_r} w_i^* y_i = \sum_{i \in s_r} \frac{w_i}{\hat{p}_i} y_i$$

- The estimator $\hat{Y}^*$ is unbiased for Y if $\hat{p}_i = p_i$. That is,

$$E\left(\hat{Y}_{TP}\right) = E_p E_r \left(\hat{Y}^* \mid s_r, r\right) = Y$$

- Here $E_r(.)$ denotes the expectation with respect with the response mechanism
- If, $\hat{p}_i \approx p_i$, $\hat{Y}^*$ is approximately unbiased for Y.

REWEIGHTING

- Two models frequently used in practice are:

  – Logistic model: $p_i = \left(1 + e^{-\mathbf{z}_i'\boldsymbol{\beta}}\right)^{-1}$

  – Uniform-within-class (response homogeneous groups):

  $$p_i = p_c \text{ if unit } i \text{ belongs to class } c$$

  For this model, estimated probabilities $\hat{p}_c$ are obtained by calculating the response rate within each class.

REWEIGHTING

- In practice, we often form reweighting classes instead of using estimated probabilities from a model:
  - We assume the uniform-within-class model
  - It gives some robustness if the nonresponse model is misspecified
  - It is easier to understand and to explain to users

REWEIGHTING

Justification for the formation of the classes

- **Let *P* a population of size *N*;**
- **We want to estimate the population total** $Y = \sum_{P} y_i$
- **We select a random sample *s* according to a sampling design *p*(.)**
- **We assume that the sampled units respond independently of one another such that**

$$a_i \sim B(1, p_i), i = 1, \ldots, N.$$

REWEIGHTING

Justification for the formation of the classes

- **An estimator of Y (based on one reweighting class) is given by**

$$\hat{Y}_1^* = \sum_{s_r} w_i^* y_i = \sum_{s_r} (w_i / \hat{p}) y_i = \sum_{s_r} w_i y_i \left(\sum_{s_r} w_i \Big/ \sum_{s} w_i \right)$$

where $\hat{p} = \sum_{s_r} w_i \Big/ \sum_{s} w_i$ is the response rate

- **The estimator $\hat{Y}_1^*$ is biased;**

$$Bias(\hat{Y}_1^*) = E_p E_r (\hat{Y}_1^* - Y) = \frac{1}{\overline{P}} \sum_{P} (p_i - \overline{P})(y_i - \overline{Y})$$

where $\overline{P} = (1/N) \sum_{P} p_i .$

REWEIGHTING

Justification for the formation of the classes

- **The bias of $\hat{\bar{Y}}_1^*$ is zero when the covariance between the probability of response and the variable of interest, which is satisfied in at least two cases:**
  - $p_i = p \ \forall i \in P$ **(uniform response mecanism troughout the population)**
  - $y_i = y_0 \ \forall i \in P$

- **These conditions are clearly not realistic in practice.**
- **The goal will be to partition the population into classes**

REWEIGHTING

Justification for the formation of the classes

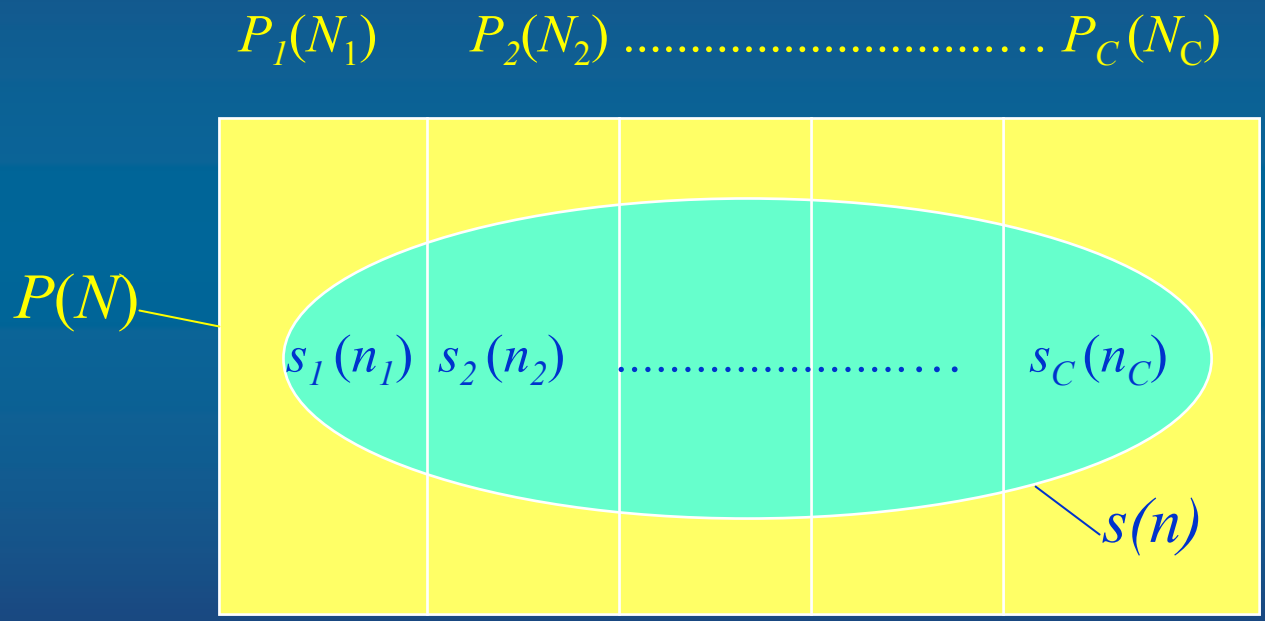
- **Suppose we partition the population P into C classes, P_1, $P_2,\ldots,P_C$ of size $N_1, N_2,\ldots,N_C$ such that**

$$P = \bigcup_{v=1}^{C} P_v \text{ et } P_v \cap P_l = \varnothing \ \forall v \neq l.$$

- **This partition leads to the corresponding partition in the sample $s_1, s_2,\ldots,s_C$ of size $n_1, n_2,\ldots, n_C$ such that**

$$s_v = s \cap P_v, v = 1,\ldots, C.$$

REWEIGHTING



REWEIGHTING

- **Within each class, we divide the sampling weight by the within-class observed probability of response** $\hat{p}_v, v = 1,\ldots,C.$
- **The estimator of *Y* based on *C* classes is given by**

$$\hat{Y}_C^* = \sum_{v=1}^{C} w_v' \hat{Y}_v^*$$

where $w_v' = \sum_{s_v} w_i / \sum_{s} w_i,$

$$\hat{Y}_v^* = \left[\sum_{s_{r_v}} (w_i / \hat{p}_v) y_i \right] \text{ and } \hat{p}_v = \sum_{s_{r_v}} w_i \Big/ \sum_{s_v} w_i .$$

REWEIGHTING

- **The estimator $\hat{Y}_C^*$ is also biased but the bias is now given by**

$$Bias(\hat{Y}_C^*) = \sum_{v=1}^{C} \frac{1}{\overline{P}_v} \sum_{P_v} (p_i - \overline{P}_v)(y_i - \overline{Y}_v)$$

where $\overline{P}_v = \frac{1}{N_v} \sum_{P_v} p_i$ and $\overline{Y}_v = \frac{1}{N_v} \sum_{P_v} y_i$

- **The bias of the estimator $\hat{Y}_C^*$ is zero if the covariance between the probabiltiy of response and the variable of interest is zero within each class.**

REWEIGHTING

- In practice, one attempts to satisfy this condition by forming classes such that, within each class, the sampled units have approximately the same probability of response (uniform-within-classes) AND/OR the sampled units have approximately the same value of the variable of interest.
- The classes are then homogeneous with respect to the response probabilities AND/OR the variable of interest.
- Hence, to obtain homogeneous classes, one has to carefully model the probability of response and/or the variable of interest.
- Eltinge and Yansaneh (1997) discuss the formation of reweigthing classes.

REWEIGHTING

- **Choice of number of classes (Eltinge and Yansaneh, 1997)**
- **U. S. Consumer Expenditure Survey; *y* = income**

C	Estimate	s.e.
1	32967	569
3	32736	530
4	32779	518
5	32630	523
10	32640	514
20	32634	508

REWEIGHTING

Advantages:

- The reweighting methods used in practice are simple
- No artificial data is created
- Allows for the use of a complete data file
- Many software are available to obtain desired estimates (but not the variance estimates)

REWEIGHTING

Disadvantages:

- Reweighting methods are difficult to implement in the case of item nonreponse because a large number of adjusted weights is required if the number of variables is large
- For example, with 3 variables, we may need up to 7 sets of adjusted weights:
  - 3 sets of adjusted weights to perform univariate analyses
  - 3 sets of adjusted weights to perform bivariate analyses
  - 1 set of adjusted weights to perform trivariate analyses

Analytical Techniques

Analytical Techniques

- Missing as a Category
- Maximum Likelihood
  - Missing data patterns
- EM - Algorithm
- Censoring - Nonignorable Missing Data
- Monte Carlo Methods (data augmentation)
  - Gibbs Sampling

Missing as a Category - Dummy Variable

- **Create a variable D such that**

$$D = \begin{cases} 1 & \textit{when data are not missing} \\ 0 & \textit{when data are missing} \end{cases}$$

- **Create a variable Z^* such that**

$$Z^* = \begin{cases} Z & \textit{when data are not missing} \\ c & \textit{when data are missing} \end{cases}$$

  - **where c is any constant**
- **Regress the dependent variable Y on Z^*, D and any other variables**
- **Note: Under MCAR estimates are biased**

Missing Data Patterns

PATTERN:

- **Which values are missing?**

MECHANISM:

- **Why are the values missing?**

Missing Data Patterns

Y = Complete Data Matrix

- **$Y = (Y_{obs}, Y_{miss})$ where Y_{obs} are the observed cases of the variable of interest and Y_{miss} are the missing cases of the variable of interest**

R = Missing Data Indicator Matrix

$$A_{ij} = \begin{cases} 0, & y_{ij} \text{ observed} \\ 1, & y_{ij} \text{ missing} \end{cases}$$

Missing Data Patterns

Model the joint distribution of *Y* and *A*

PATTERN:

- **Concerns the distribution of *A***

MECHANISM:

- **Concerns the distribution of *A* given *Y***

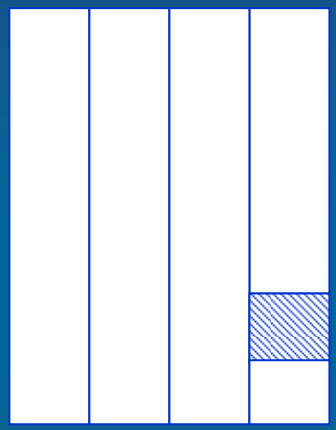
Missing Data Patterns

- Consider *n* variables, either a single variable *Y* measured at *n* time points or *n* different variables
- There will be 2^n possible missing data patterns
- For example with three variables we have the following possible patterns

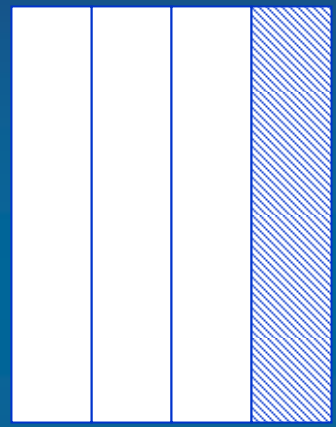
Missing Data Patterns

	y_{t1}	y_{t2}	y_{t3}	
1	√	√	√	Complete Response
2	√	√	X	Some observations are observed at each of the time points
3	√	X	√	
4	X	√	√	
5	X	X	√	
6	√	X	X	
7	X	√	X	
8	X	X	X	Total Nonresponse

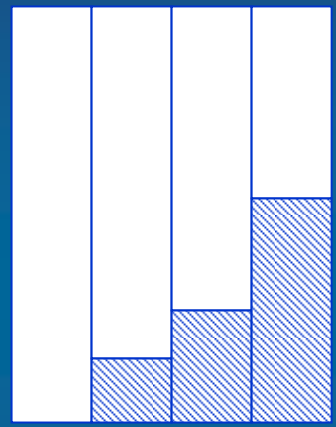
Missing Data Patterns



•Univariate Missing



•Unit Nonresponse



•Monotone Missing Data

•Little (1993, 1994, 1995), Rubin 1974, Li 1988, Hedeker and Gibbons (1997)

Maximum Likelihood - Monotone Pattern

- For a complete sample consider the likelihood for the entire sample is

$$L(\theta) = \prod_{i=1}^{n} P(z_i, y_i \mid \theta)$$

- Consider the case where we observe X and Y for a portion of the sample and the pattern is monotonic

$$L(\theta \mid Y) = \prod_{i=1}^{n} P(y_{i1}, \ldots, y_{ip} \mid \theta)$$

$$= \prod_{j=1}^{p} \prod_{i=1}^{n} P(y_{ij} \mid y_{i1}, \ldots, y_{i,j-1}, \theta)$$

Maximum Likelihood - Repeated Measures

- **Y = (Y_{obs} , Y_{miss}) where Y_{obs} are the observed cases of the variable of interest and Y_{miss} are the missing cases of the variable of interest**
- ***Z* = fixed covariates, possibly including time t**
- $$A = \begin{cases} 0, & observed \\ k, & drop-out \end{cases}$$
- **β = subject specific parameters**
- **θ = fixed model parameter**

Maximum Likelihood - Repeated Measures

Full likelihood

$$L_F(\theta \mid \mathrm{Y}_{\mathrm{obs}}, A) \propto \prod_{i=1}^{N} \int_{unobserved} P(Y_{obs}, Y_{miss}, A, \beta \mid Z, \theta)\, \mathrm{dY}_{\mathrm{miss}}\, d\beta$$

Likelihood without missing data mechanism

$$L_I(\theta \mid \mathrm{Y}_{\mathrm{obs}}) \propto \prod_{i=1}^{N} \int_{unobserved} P(Y_{obs}, Y_{miss}, \beta \mid Z, \theta)\, \mathrm{dY}_{\mathrm{miss}}\, d\beta$$

- **This is what is obtained with PROC MIXED, BMDP 5V, HLM, EGRET**

MLE - Models for Missing Mechanism

- **[Y, A, β | Z] = [Y | Z, β] [β, Z][RA| Z, Y, β]**
  (random-coefficient selection models)
- **MCAR**

$$[A \mid Z, Y, \beta] = [A]$$

- **MAR**

$$[A \mid Z, Y, \beta] = [A \mid Z, Y_{obs}]$$

(Source: Hedeker)

MLE - Models for Missing Mechanism

- **MAR - Co-variate Dependent Drop-out**

$$[A \mid Z, Y, \beta] = [A \mid Z]$$

- **Non-Ignorable - Outcome-dependent**

$$[A \mid Z, Y, \beta] = [A \mid Z, Y_{obs}, Y_{miss}]$$

- **Non-Ignorable - Random-coefficient Dependent**

$$[A \mid Z, Y, \beta] = [A \mid Z, \beta]$$

(Source: Hedeker)

The EM Algorithm

- A technique for finding maximum-likelihood estimates for parametric models when the data are not fully observed (Dempster, Laird and Rubin, 1977)

- 2 steps
  - Expectation: the function $Q(\theta \mid \theta^{(t)})$ is calculated by averaging the complete-data loglikelihood $l(\theta \mid Y)$ over $Q(\theta \mid \theta^{(t)})$
  - Maximization: in which $\theta^{(t+1)}$ is found by maximizing $Q(\theta \mid \theta^{(t)})$

Example-EM Multivariate Normal

- Suppose we have 3 variables with missing data in no particular pattern, X_1, X_2, X_3
- E:
  - estimate the covariance and means using listwise or pairwise deletion
  - based on these compute coefficients for the regression of any one of the *Xs* on any subset of the other two
  - then generate imputed values for the *X* of interest based on the other *Xs*

The EM Algorithm

- **M:**
  - **after imputation calculate new means and covariances using the imputed data and the nonmissing data**
  - **for means use regular formulas, for covariances take into account residual variances and residual covariances**
  - **with the new means and covariances start on E step again**

Non-ignorable Missing Data – Censoring

- Type I
- Type II
- Right
- Left

Non-ignorable Missing Data - Censoring

- Right Censoring

$$P(R_i = 1 \mid y_i) = P(y_i \quad \textit{observed} \mid y_i) = \begin{cases} 1 & y_i < 0 \\ 0 & y_i \geq 0 \end{cases}$$

Examples

- Incomplete Exponential Sample:

$$\hat{\theta}_{complete} = \sum_{1}^{m} \frac{y_i}{m}$$

$$\hat{\theta}_{missing} = (\sum_{1}^{m} y_i + (n+m)c) / m$$

Examples

- Duration until death
  - Censoring lost to follow-up / fall out from the study
- Study of heights (normal data)
  - Minimal height standards lead to a censored data set

Data Augmentation & Monte-Carlo Methods

- **Data is created by pseudo-random draws from probability distributions - a general method for calculating posterior distributions**

Example for mulitvariate normal

- **Choose starting values**
- **Use current values to calculate means and covariances**
- **Use regression estimates to generate predicted values BUT add a random draw from the residual normal distribution**
- **Recalculate means and covariances**
- **Make a random draw from the posterior distribution of the means and covariances**
- **Using the randomly chosen draws go back to step 2 and continue until convergence**

Gibbs Sampling

- **Suppose that for some p>1, the random variable $Z \in Z$, $Z=(Z_1,\ldots,Z_p)$ we have the corresponding conditional densities $f_1,\ldots,f_p$**

$$Z_i \mid z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_p \sim f_i(z_i \mid z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_p)$$

- **Algorithm**

$$\text{given } z^{(t)} = (z_1^{(t)}, \ldots, z_p^{(t)}), \text{ generate}$$

$$1.\ Z_1^{(t+1)} \sim f_1(z_1 \mid z_2^{(t)}, \ldots, z_p^{(t)});$$

$$2.\ Z_2^{(t+1)} \sim f_2(z_2 \mid z_1^{(t+1)}, z_3^{(t)} \ldots, z_p^{(t)});$$

M

$$p.\ Z_p^{(t+1)} \sim f_p(z_p \mid z_1^{(t+1)}, \ldots, z_{p-1}^{(t+1)}).$$

IMPUTATION

- Imputation is generally used to treat item nonresponse, although it may be used to treat unit nonresponse.

- We distinguish between single and multiple imputation:
  - Single imputation consists in creating one single artificial value to fill the hole of a missing value
  - Multiple imputation consists in creating $M \geq 2$ artificial values to fill the hole of a missing value

IMPUTATION

- Imputation can be done by a computer or manually (we try to avoid manual imputation as much as possible since it is potentially subjective)
- Except for manual imputation, all imputation methods used in practice can be justified by a model
- Properties (bias, variance,…) of imputed estimators depend on the quality of the model and the quality of the auxiliary variables
- In practice, it is customary to first for imputation classes and then impute independantly within each class. The justification for the formation of the classes is similar to tthe one for reweighting classes.

SINGLE IMPUTATION

Advantages:

- Single imputation leads to the creation of a complete data file (simpler for the users)
- Unlike weighting adjustment methods, imputation permits the use of a single weight
- The results of different analyses are bound to be consistent with each other
- Imputation tries to use all observed data (no loss of information)
- Many software are available to obtain point estimate (but not variance estimates)

SINGLE IMPUTATION

Desadvantages:

- Imputed values are artificial and may give a false impression of precision
- Even though imputation leads to the creation of a complete data file, inference is valid only the underlying assumptions are satified.
- In general, single imputation does not preserve the relationships between the variables
- The fact that the imputed values are treated as observed values may lead to a substantial underestimation of the variance of the estimators, especially if the nonresponse rate is appreciable

SINGLE IMPUTATION

Imputation methods can be classified into two broad classes:

- Deterministic imputation: Given observed sample data, we would always obtain the same imputed values if we repeated the imputation process
- Stochastic or random imputation: Given observed sample data, we would not obtain the same imputed values if we repeated the imputation process

SINGLE IMPUTATION

Deterministic imputation: Regression imputation

• We assume that a vector of q auxiliary variables is available for all units in the sample. A missing value y_i is replaced by

$$y_i^* = \mathbf{z}_i'\hat{\boldsymbol{\beta}}_r$$

Particular cases:

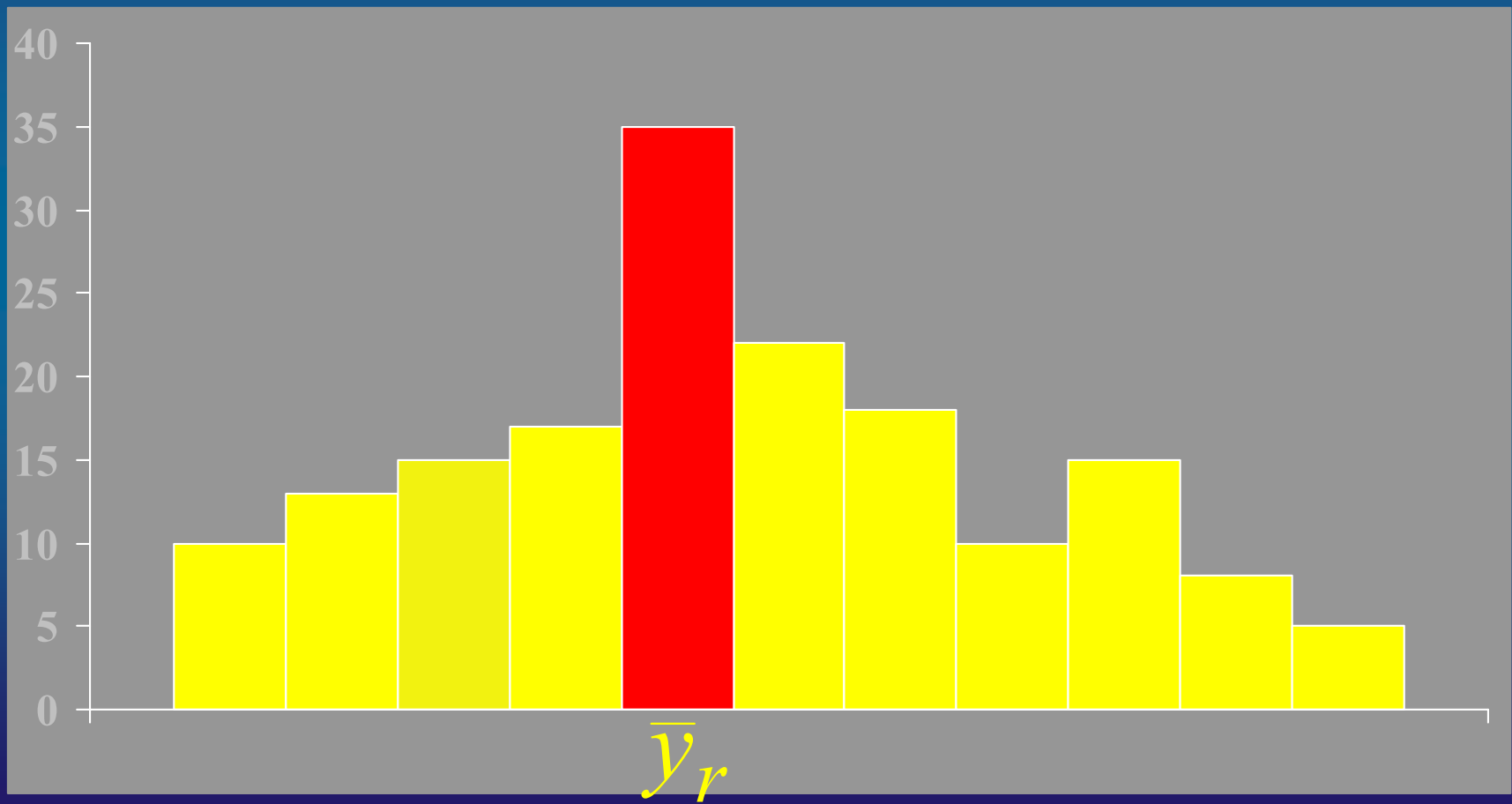
– Mean imputation: ⟶ $y_i^* = \bar{y}_r$ (mean of the respondents)

– Ratio imputation: ⟶ $y_i^* = (\bar{y}_r / \bar{z}_r) z_i$

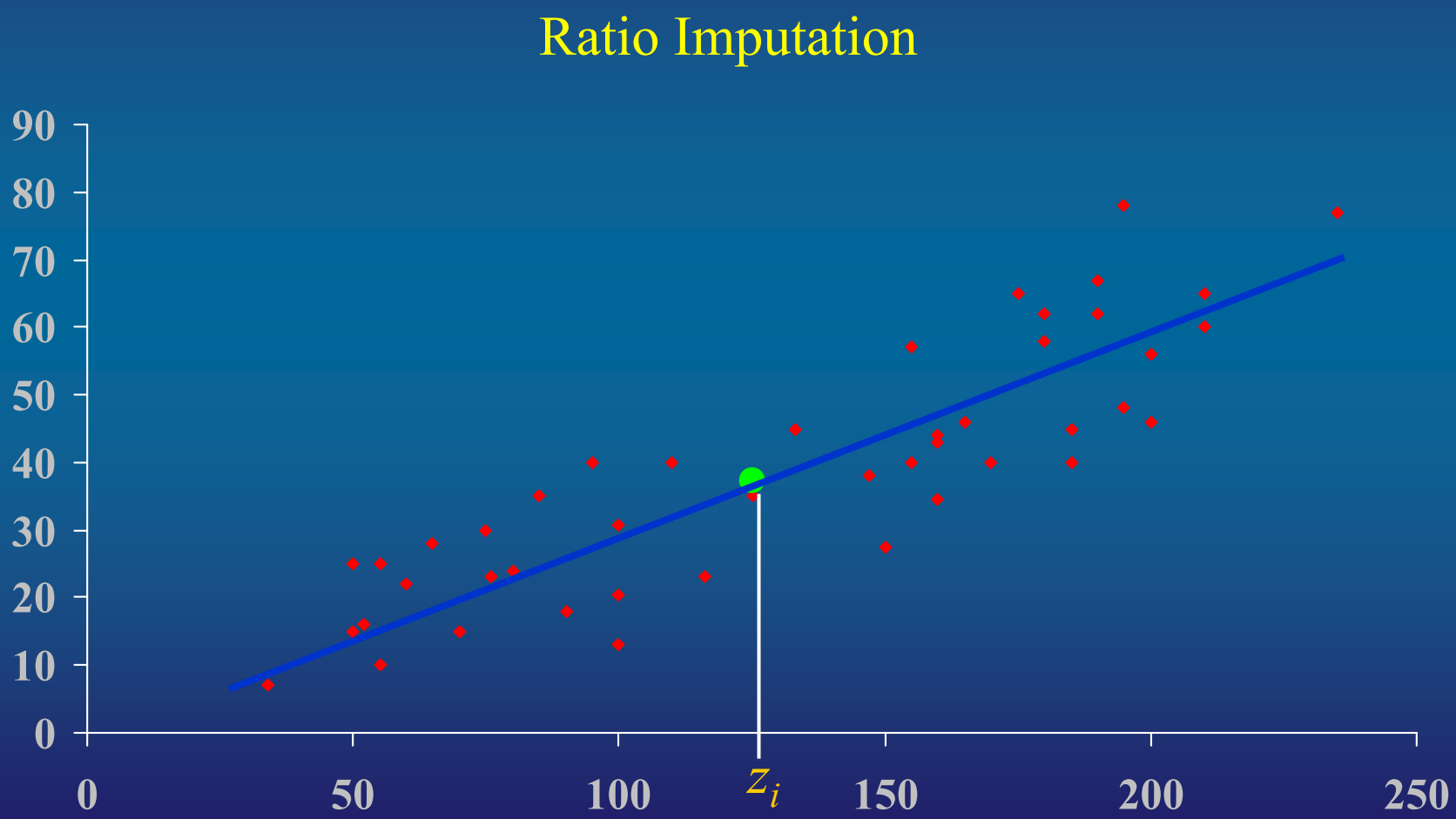
– Historical imputation: ⟶ $y_i^* = z_i$

SINGLE IMPUTATION

Distribution of the variable of interest after mean imputation



SINGLE IMPUTATION



SINGLE IMPUTATION

Deterministic imputation: Nearest-Neighbour imputation

The missing value y_i is replaced by the nearest-neighbour's value (according to a distance function based on one or more auxiliary variables),

$$y_i^* = y_j \text{ for } j \in s_r \text{ s.t. } dist(\mathbf{z_j}, \mathbf{z_i}) \text{ is minimum}$$

- Provides an existing value
- The distance function has to be determined (Euclidian,...)
- In the case of multiple auxiliary variables, it is better to first standardize these variables (for example, by substracting the mean and dividing by the standard deviation)

SINGLE IMPUTATION

Random imputation: Regression imputation with added residuals

A missing value y_i is replaced a predicted value to which we add a random residual

$$y_i^* = \mathbf{z_i'}\hat{\boldsymbol{\beta}}_\mathbf{r} + e_i^*$$

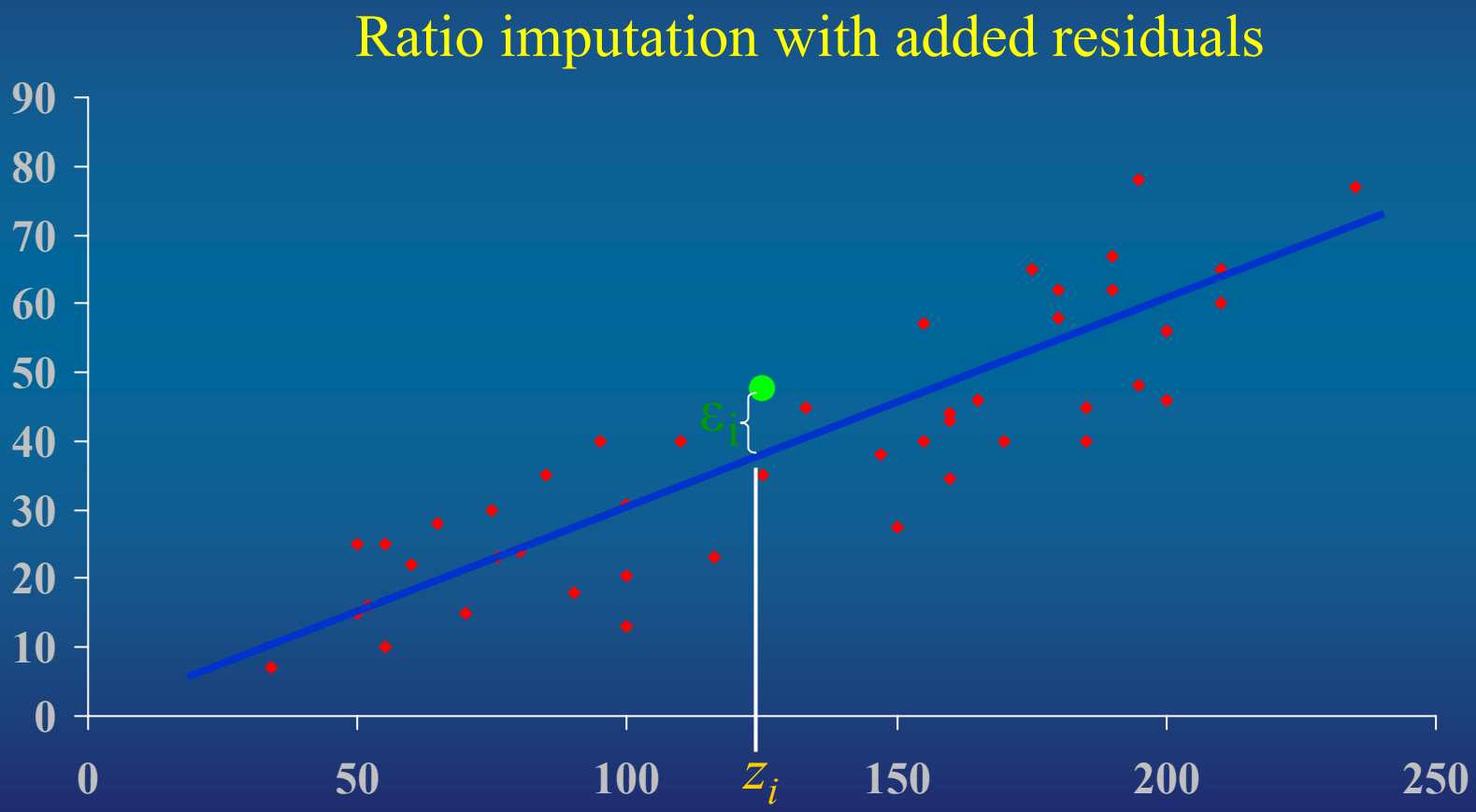
Particular cases:

Random Hot-deck imputation: $y_i^* = \bar{y}_r + e_i^*$

Ratio with added residuals: $y_i^* = (\bar{y}_r / \bar{z}_r) z_i + e_i^*$

- Residuals may be selected randomly:
  - from the respondent's set of residuals
  - from a given distribution (for example, $N(0,c)$)

SINGLE IMPUTATION



SINGLE IMPUTATION

Deterministic imputation

- **Tend to distort the distribution of the variable of interest**

Random imputation

- **Tend to preserve the distribution of the variable of interest**
- **Increase the variance of the point estimators**

SINGLE IMPUTATION

Which imputation method should one use ?

- Simplicity
- Auxiliary variable(s) ?
- Reproducability?
- In practice, a combination of several methods is used

SINGLE IMPUTATION

- In the presence of nonresponse to item y, we define an imputed estimator for the population mean $\bar{Y} = (1/N)\sum_{i \in P} y_i$ by

$$\bar{y}_I = \frac{1}{N}\left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^*\right]$$

where y_i^* denotes the imputed value for missing y_i

- Under mean imputation, $y_i^* = \bar{y}_r = \sum_{i \in s_r} w_i y_i \Big/ \sum_{i \in s_r} w_i \Rightarrow \bar{y}_I = \bar{y}_r$
- Under ratio imputation, $y_i^* = \bar{y}_r = \left(\sum_{i \in s_r} w_i y_i \Big/ \sum_{i \in s_r} w_i z_i\right) z_i$

  $\Rightarrow \bar{y}_I = (\bar{y}_r / \bar{z}_r)\bar{z}$ with $\bar{z} = \frac{1}{N}\sum_{i \in s} w_i z_i$

SINGLE IMPUTATION

1. Is the imputed estimator $\bar{y}_I$ unbiased for the population mean $\bar{Y}$
2. Design-based approach (DB): We assume that within each imputation class, the response mechanism is uniform.
3. Model-assisted approach (MA): We assume that, within each imputation class, the response mechanism is ignorable in the sense that the probability of response may depend on some auxiliary variables but not on the variable being imputed. We then need an imputation model usually of the form

$$m: y_i = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$
$$E_m(\varepsilon_i) = 0, V_m(\varepsilon_i) = \sigma_i^2, Cov_m(\varepsilon_i, \varepsilon_j) = 0 \quad \text{si } i \neq j$$

SINGLE IMPUTATION

• In the case of simple imputation methods (mean, ratio, regression,…), the imputed estimator is approximately unbiased.

(i) Design-based approach: $E(\bar{y}_I) = E_p E_r(\bar{y}_I \mid s, r) \approx \bar{Y}$

(ii) Model-assisted approach: $E(\bar{y}_I) = E_p E_m(\bar{y}_I \mid s, r) \approx \bar{Y}$

• The imputed estimator is therefore approximately unbiased if the underlying assumptions (response mechanism or imputation model) are verified. Otherwise, the imputed estimator may be biased.

SINGLE IMPUTATION

Example : Mean imputation under the design-based approach

Under this approach, we assume an uniform response mechanism; suppose that, in reality, the mechanism is not uniform but rather $P(i \in s) = p_i$. Then, the imputed estimator is biased;

$$Bias(\bar{y}_I) = E_p E_r(\bar{y}_I - \bar{Y} \mid s) = \frac{1}{N\bar{P}} \sum_{i \in P} (p_i - \bar{P})(y_i - \bar{Y})$$

• The bias is equal to zero if the covariance between the probability of response and the variable of interest is zero, which is satisfied, for example, in the case of uniform response.

SINGLE IMPUTATION

Example: Ratio imputation under the model-assisted approach

- Under ratio imputation, $\bar{y}_I = \frac{\bar{y}_r}{\bar{z}_r}\bar{z}$.
- The underlying imputation model is $m_1 : y_i = \beta z_i + \varepsilon_i$
- Suppose that the true model in the population is not m_1 but rather $m_2 : y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$
- Then the imputed estimator is biased;

$$Bias(\bar{y}_I) = E_p E_r E_{m_2}(\bar{y}_I - \bar{Y}) = \beta_0\left[\frac{\bar{Z}}{\bar{Z}_p} - 1\right]$$

where $\bar{Z}_p = \sum_{i=1}^{N} p_i z_i \Big/ \sum_{i=1}^{N} p_i$

SINGLE IMPUTATION

Variance estimation: Why estimate the variance?

- To measure of the quality of the estimates
- To helps drawing the right conclusions
- To contributes in correctly informing users
- In the presence of imputed values, to help to learn about the impact of imputation on variance estimates
- To improve future occasions of the survey by allocating resources more appropriately

SINGLE IMPUTATION

- Several variance estimation methods have been proposed in the litterature. They include the following:
  - Model-assisted approach (Särndal, 1992)
  - The adjusted jackknife (Rao and Shao, 1992)
  - The bootstrap (Shao and Sitter, 1996)
  - The reverse approach (Shao and Steel, 1999)
- We briefly describe these methods.

SINGLE IMPUTATION

- Model-assisted approach: This method has been developped under assumtion MA. The method is based on the following decomposition:

$$\bar{y}_I - \bar{Y} = \underbrace{(\bar{y} - \bar{Y})}_{\text{error due to sampling}} + \underbrace{(\bar{y}_I - \bar{y})}_{\text{error due to nonresponse}}$$

- The variance of the imputed estimator is given by

$$V_{tot} = V(\bar{y}_I - \bar{Y}) = E(\bar{y}_I - \bar{Y})^2 = E_r E_p E_m (\bar{y}_I - \bar{Y})^2$$

$$= V_{sam.} + V_{imp.} + V_{mix.}$$

- The component V_{mix} is zero for simple imputation methods or is small compared to the other components.

SINGLE IMPUTATION

Jackknife in the full response case

- **Let s be a sample, of size *n*, selected form a population *P* of size *N* according to unistage design *p*(.). Let *Y* be a parameter of interest and $\hat{Y}_{HT} = \sum_{i \in s} w_i y_i$ be an estimator of *Y*. The jackknife variance of is obtained as follows:**
- **(i) Remove an unit**

  (ii) Adjust the sampling weights

  (iii) Compute $\hat{Y}_{HT}$ using the adjusted weights

  (iv) Put back the unit removed at step (i); then, remove the next unit and re-compute $\hat{Y}_{HT}$

  (v) Repeat (i)-(iv) until all the units have been removed

SINGLE IMPUTATION

Jackknife in the full response case

- **The jackknife variance is then obtained as follows:**

$$v_J\left(\hat{Y}_{HT}\right) = \frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{Y}_{HT(j)} - \hat{Y}_{HT}\right)^2$$

where $\hat{Y}_{HT(j)} = \sum_{i \neq j \in s} w_{i(j)} y_i$ and $w_{i(j)} = \begin{cases} \dfrac{n}{n-1} w_i & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases}$

- **The jackknife is especially useful for complex parameters and/or complex designs**

SINGLE IMPUTATION

The adjusted jackknife: The Rao-Shao adjusted jackknife may be described as follows in the case of random hot-deck imputation:

- Works the same way as the jackknife under full response except that
  - If we remove a responding unit j, then the imputed values are adjusted by adding a quantity

$$y_i^* \rightarrow y_i^* + \underbrace{E_*^{(j)}\left(y_i^*\right) - E_*\left(y_i^*\right)}_{\text{adjustment}}$$

where $E_*^{(j)}$ is the expectation with respect to the random hot-deck procedure

  - If we remove a nonresponding unit j, then the imputed values remain unchanged.

SINGLE IMPUTATION

- The Rao-Shao jackknife variance estimator is then given

$$v_{JRS}\left(\hat{Y}_I\right) = \frac{n-1}{n}\sum_{j\in s}\left(\hat{Y}^a_{I(j)} - \hat{Y}_I\right)^2$$

where

$$\hat{Y}^a_{I(j)} = \begin{cases} \sum_{i\neq j\in s_r} w_{i(j)} y_i - y_j + \sum_{i\in s_m} w_{i(j)}\left[y^*_i + E^{(j)}_*\left(y^*_i\right) - E_*\left(y^*_i\right)\right] & \text{if } j\in s_r \\ \sum_{i\in s_r} w_{i(j)} y_i + \sum_{i\neq j\in s_m} w_{i(j)} y^*_i - y^*_j & \text{if } j\in s_m \end{cases}$$

SINGLE IMPUTATION

- **The bootstrap: The Shao-Sitter bootstrap may be described as follows:**

- **(1) Select a bootstrap sample s^* (simple random sample with replacement) of size $n^* = n-1$ from the sample $\tilde{s}$ after imputation where $\tilde{s} = \{y_i : i \in s_r\} \cup \{y_i^* : i \in s_m\}$ Note that $s^* = s_r^* \cup s_m^*$ (where s_r^* et s_m^* denote the samples of respondents and nonrespondents respectively in s^*)**
- **(2) Let a_i^* be the response indicator associated with $y_i \in s^*$, i.e., $s_m^* = \{i \in s^* : a_i^* = 0\}$. and $s_r^* = \{i \in s^* : a_i^* = 1\}$ Then, use the same imputation procedure/method than the one used in the original sample s to impute nonrespondents in s^*.**

SINGLE IMPUTATION

- (3) Compute the bootstrap version of the imputed estimator $\bar{y}_I, \bar{y}_I^*$ from the bootstrap sample s^*, i.e.,

$$\bar{y}_I^* = \frac{1}{N}\left[\sum_{i \in s_r^*} w_i^* y_i + \sum_{i \in s_m^*} w_i^* y_i^*\right]$$

where $w_i^* = w_i \times \frac{n}{n-1}$ denote the bootstrap weights

- Repeat steps (1)-(3) *B* times.
- The bootstrap variance estimator is given

$$v_B(\bar{y}_I) = \frac{1}{B-1}\left[\sum_{b=1}^{B}\left(\bar{y}_{I(b)}^* - \bar{\bar{y}}_I^*\right)^2\right] \text{ with } \bar{\bar{y}}_I^* = \frac{1}{B}\sum_{b=1}^{B}\bar{y}_{I(b)}^*$$

SINGLE IMPUTATION

Nonresponse viewed as two-phase sampling

Population → complete sample → sample of
respondents or P s (s_r, s_m)

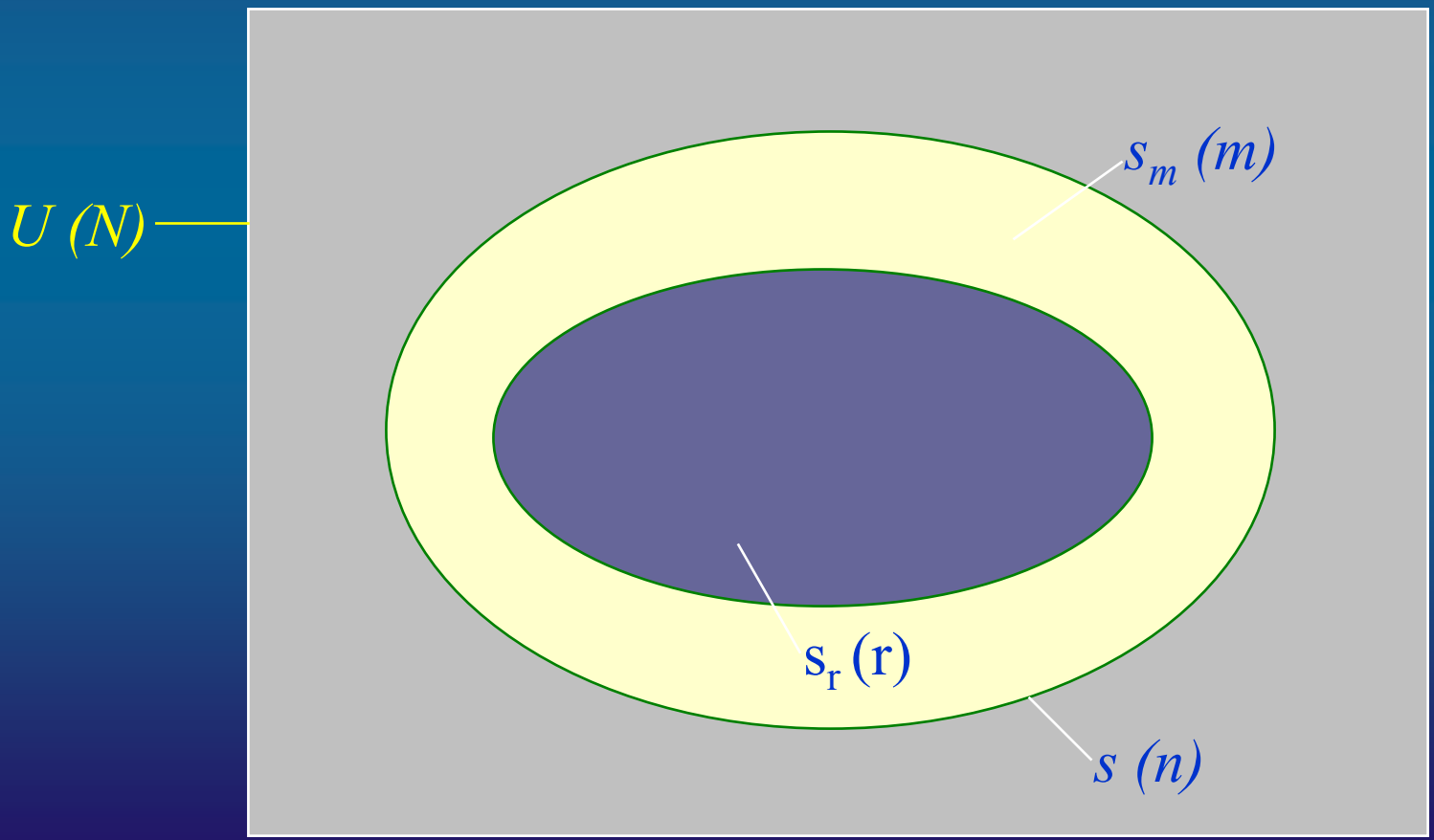
Reverse approach (Fay, 1991)

Population Population of respondents sample
and nonrespondents →

or P (P_r, P_m) (s_r, s_m)

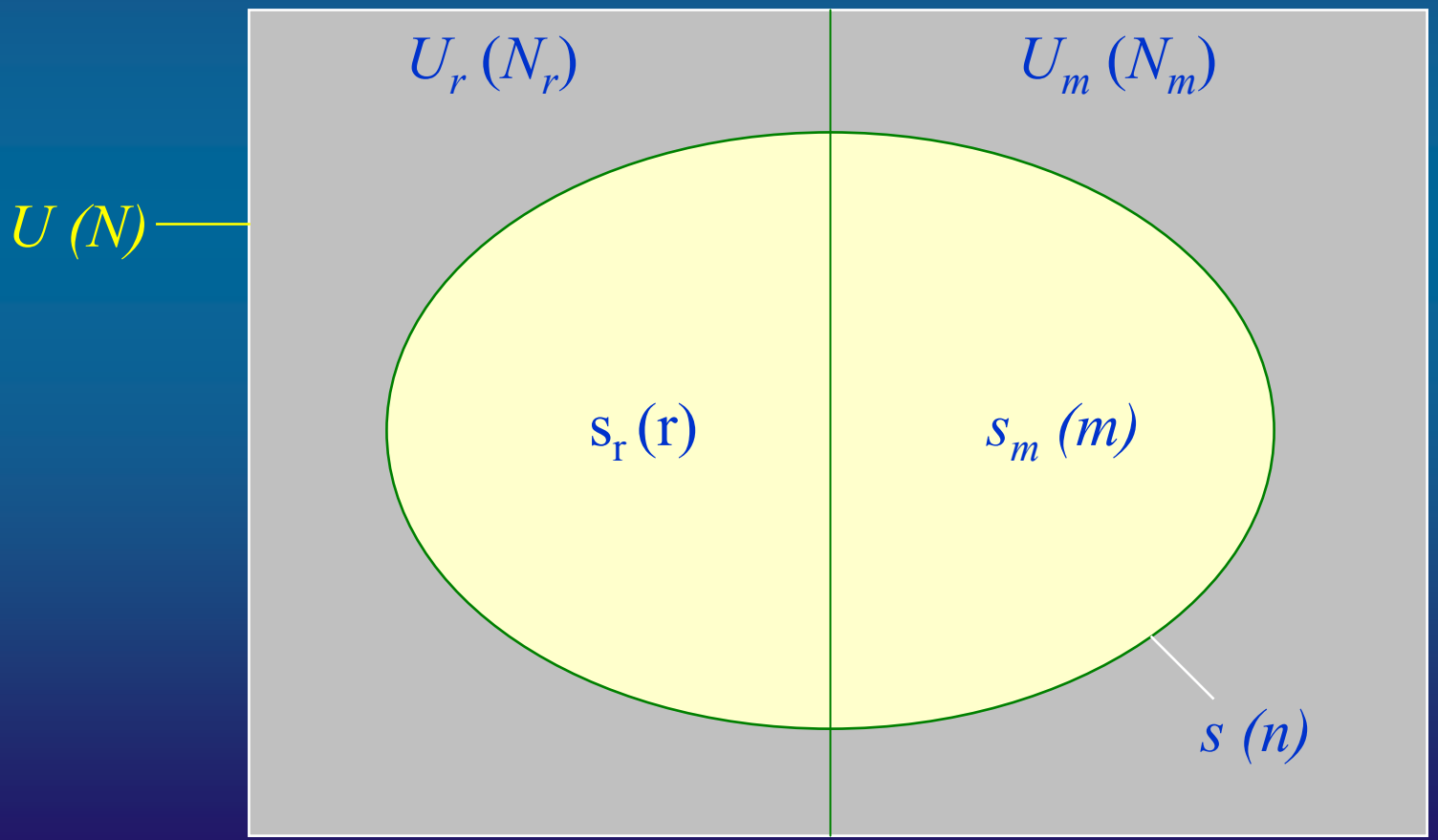
SINGLE IMPUTATION

The two-phase approach



SINGLE IMPUTATION

The reverse approach



SINGLE IMPUTATION

Two-phase approach:

$$E(\bar{y}_I) = E_p E_r(\bar{y}_I \mid s),$$

and

$$V(\bar{y}_I) = V_p E_r(\bar{y}_I \mid s) + E_p V_r(\bar{y}_I \mid s)$$

Reverse approach:

$$E(\bar{y}_I) = E_r E_p(\bar{y}_I \mid a_i),$$

and

$$V(\bar{y}_I) = E_r V_p(\bar{y}_I \mid a_i) + V_r E_p(\bar{y}_I \mid a_i)$$

SINGLE IMPUTATION

The reverse approach

$$V(\bar{y}_I - \bar{Y}) = E_r V_p(\bar{y}_I - \bar{Y} \mid a_i) + V_r E_p(\bar{y}_I - \bar{Y} \mid a_i)$$

Variance estimation (Shao and Steel, 1999):

We estimate each component separately:

i. Estimation of $E_r V_p(\bar{y}_I - \bar{Y} \mid a_i)$

   i. The estimation may be performed using Taylor linearization, jackknife, bootstrap,…

   ii. The estimate of this component does not depend on the response mechanism and/or imputation model

SINGLE IMPUTATION

The reverse approach

i. Estimation of $V_r E_p(\bar{y}_I - \bar{Y} \mid a_i)$

   i. the estimator depends on the response mechanism and/or imputation model

   ii. in some cases, the derivation may be tedious

   iii. if the sampling fraction is negligible (or small), the this component is negligible (small) relative to the first component

MULTIPLE IMPUTATION

Example : Consider a random sample of size $n = 10$ selected from an infinite population.

Unit	Age	Sex
1	55	M
2	60	F
3	?	M
4	58	F
5	70	M
6	?	F
7	60	F
8	73	F
9	?	M
10	?	F

$Age^{(1)}$	Sex
55	M
60	F
61	M
58	F
70	M
69	F
60	F
73	F
63	M
72	F

$Age^{(2)}$	Sex
55	M
60	F
50	M
58	F
70	M
56	F
60	F
73	F
59	M
65	F

$\bar{y}_I^{(1)} = 64.1$ $\bar{y}_I^{(2)} = 60.2$

$s^{2(1)}/n = 4.1$ $s^{2(2)}/n = 4.9$

MULTIPLE IMPUTATION

- The objective is to estimate the population mean $\bar{Y}$ of the variable *AGE*

Point estimation:

1. Compute the mean of the variable AGE $\bar{y}_I^{(i)}$ for each of the complete data file

2. Combine the point estimators $\bar{y}_I^{(i)}$ by computing the multiple imputed estimator

$$\bar{y}_{IM} = \frac{1}{M}\sum_{i=1}^{M}\bar{y}_I^{(i)}$$

- In the example, $\bar{y}_{I2} = \frac{1}{2}(64.1 + 60.6) = 62.4$.

MULTIPLE IMPUTATION

Variance estimation:

- Similar to variance estimation for ANOVA
- Let $\bar{U}_M = \frac{1}{M}\sum_{i=1}^{M}\frac{s^{2(i)}}{n}$ and $B_M = \frac{1}{M-1}\sum_{i=1}^{M}\left(\bar{y}_I^{(i)} - \bar{y}_{IM}\right)^2$
- $\bar{U}_M$ is the within variance and B_M is the between variance
- The total variance of the imputed estimator $\bar{y}_{I2}$ is given by

$$T_M = \bar{U}_M + \left(1 + \frac{1}{M}\right)B_M$$

- In the example,

$$\bar{U}_2 = (4.1 + 4.9)/2 = 4.5,\ B_2 = \left[(64.1 - 62.4)^2 + (60.6 - 62.4)^2\right]/1 = 6.1$$
$$\therefore T_2 = 4.5 + (1 + 0.5)6.1 = 13.6$$

MULTIPLE IMPUTATION

- **Let** $\mathbf{Y}_{sam} = \{Y_i : i \in s\} = \{\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\}$

  where $\mathbf{Y}_{obs} = \{Y_i, i \in s_r\}$ and $\mathbf{Y}_{mis} = \{Y_i, i \in s_m\}$

- **Let *Q* be an arbitrary parameter (scalar)**
- **Multiple imputation assumes that**

$$\frac{Q - \hat{Q}}{\sqrt{U}} \sim N(0,1)$$

- **where $\hat{Q}$ is an estimator of *Q* that we would use if no data were missing and *U* is its variance.**

MULTIPLE IMPUTATION

- Suppose we use *M* imputed values
- *M* complete data files are then created.
- The goal is to adequatly combine the point and variance estimates obtained from each complete data file

- Let $\hat{Q}_I^{(i)}$ and $U^{(i)}$ be the point and variance estimates repectively for the i^{th} data file , $i = 1,2,\ldots M$.
- In fact, $\hat{Q}_I^{(i)}$ and $U^{(i)}$ are different versions of $\hat{Q}$ and U respectively

MULTIPLE IMPUTATION

- **The multiple imputed estimator of Q is given by**

$$\bar{Q}_{IM} = \frac{1}{M}\sum_{i=1}^{M}\hat{Q}_I^{(i)}$$

- **The total variance of $\hat{Q}_{IM}$, denoted by T_M, is given by**

$$T_M = \bar{U}_M + \left(1 + \frac{1}{M}\right)B_M$$

where $\bar{U}_M = \frac{1}{M}\sum_{i=1}^{M}U^{(i)}$ and $B_M = \frac{1}{M-1}\sum_{i=1}^{M}\left[\hat{Q}_I^{(i)} - \bar{Q}_{IM}\right]^2$

MULTIPLE IMPUTATION

- $\bar{U}_M$: **within-imputation variance**
- B_M : **between-imputation variance**

- **We have**

$$\frac{Q-\bar{Q}_{IM}}{\sqrt{T_M}} \sim t_{\upsilon_M}$$

where t_{υ_M} is the student distribution with υ_M degrees of freedom,

$$\upsilon_M = (M-1)\left(1+\frac{1}{r_M}\right)^2 \quad \text{and} \quad r_M = \frac{(1+1/M)B_M}{\bar{U}_M}.$$

MULTIPLE IMPUTATION

Remarks:

(1) Clearly, imputation must be stochastic

(2) r_M is the relative increase in variance due to nonresponse

(3) In the case of full response, we have $r_M = 0$ so $\upsilon_M = \infty$, which bring us back to the starting assumption for multiple imputation

(4) A $100(1-\alpha)\%$ confidence interval for Q is then given by

$$\overline{Q}_{IM} \pm t_{\upsilon_M,\alpha/2}\sqrt{T_M}$$

MULTIPLE IMPUTATION

(5) Rubin (1987) defines the fraction of missing information as

$$\gamma_M = \frac{r_M + 2/(\upsilon_M + 3)}{r_M + 1}$$

(6) The typical conclusions associated with multiple imputation are given by

$$\lim_{n\to\infty} E(\bar{Q}_{I\infty} - Q) = 0$$

and

$$\lim_{n\to\infty} \left[E(T_\infty) - V(\bar{Q}_{I\infty} - Q)\right] = 0$$

where $\bar{Q}_{I\infty} = \lim_{M\to\infty} \bar{Q}_{IM}$ and $T_\infty = \lim_{M\to\infty} T_M$

MULTIPLE IMPUTATION

(10) How many imputation should one use ?

The use of a finite number of imputation M versus the use of and infinite number of imputation may be measured by

	γ_M				
M	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

$$RE = \left(1 + \frac{\gamma_M}{M}\right)^{-1}$$

MULTIPLE IMPUTATION

- In order for the inference to be valid in the frequentist framework, imputation should be proper (Rubin, 1987)
- Except in trivial cases, it can be extremenly difficult to determine whether an imputation method is proper
- Binder and Sun (1996) question whether or not proper imputation can be achieved in the case of complex surveys (stratified multistage designs)
- The imputation methods used in practice (regression imputation, random hot-deck,…) are not proper in the sense of Rubin

MULTIPLE IMPUTATION

Example: Proper imputation: Approximate Bayesian Bootstrap

(i) For each imputation ($k = 1, 2, \ldots, M$), draw $\mathbf{Y}^*_{(k)} = \left\{Y^*_{1(k)}, Y^*_{2(k)}, \ldots Y^*_{r(k)}\right\}$ **with replacement and equal probability of selection from the set of respondents** $Y_{obs} = \{Y_i : i \in s_r\}$. **The set is then the set of respondents for the k^{th} imputation.** $\mathbf{Y}^*_{(k)}$

(ii) For each missing value $j \in s_m$, draw $Y^{**}_{(k)}$ from $\mathbf{Y}^*_{(k)}$ **with replacement and equal probability of selection and use** $Y^{**}_{(k)}$ **as the k^{th} imputed value.**

(iii) Repeat steps (i) et (ii) M independent times.

MULTIPLE IMPUTATION

- Schafer (1997) gives a diiferent definition of proper imputation, called Bayesianly proper imputation
- A imputation method is said to be Bayesianly proper if the imputed values independent realizations of $P(Y_{mis} \mid Y_{obs})$.
- To generate Bayesianly proper imputations, one must
  - Model the variable of interest by including all the appropriate auxiliary information
  - The imputed values may then be generated from by using, for example, MCMC methods (Data augmentation by Tanner and Wong, 1987)
- A Bayesianly proper imputation does not guarantee that the imputation method is proper in the sense of Rubin (1987)

MULTIPLE IMPUTATION

- **Wang and Robins (1998) showed that, when the nonresponse rate is high and the number of imputation is small, then multiple imputation leads to inconsistent variance estimates.**

	Coverage probability	Length of the C.I.
Type A $p = 0.25$	97.8	1.00
Type B $p = 0.25$	98.0	1.87
Type A $p = 0.5$	96.5	1.00
Type B $p = 0.5$	94.3	1.52

– Type A: Improper

– Type B: Proper

Evaluation

What should I report??

- The amount of missing data
- The impact of the missing data
- The solution of the missing data problem
- The impact of the missing data solution

Discussion

- Consider the context of the missing data when deciding how to correct it
- Consider more than one solution and select the most appropriate for your study
- There is no “single solution”