Chapter 7

MH4511: SAMPLING AND SURVEY

Nonresponse

WHY NONRESPONSE OCCUR?

- Nonresponse occurs when the surveyors fail to obtain data on sampling units partly or completely
- In agriculture or wildlife surveys, the term missing data is generally used instead of nonresponse
- The best way to deal with nonresponse is to prevent it
- Main reasons for nonresponse:
  - The surveyors cannot contact the sampling units
  - The sampling units may not be able to respond, for example due to illness
  - The sampling units refuse to participate (throw away the questionnaires)
  - The sampling units participate, but decline to answer some of the questions (such as income)

TWO TYPES OF NONRESPONSE

Generally there are two types of nonresponse

- Unit nonresponse
  - The entire observation unit is missing (persons may refuse to be interviewed)
  - It could possibly use age, gender, race and other characteristics to adjust for the nonresponse
- Item nonresponse
  - Some measurements are present for the observation but at least one item is missing
  - Item nonresponse occurs largely because of refusals

FOUR APPROACHES FOR DEALING WITH NONRESPONSE

- Prevent it by design
  - Reduce the nonresponse to insignificant levels, so that any remaining nonresponse causes little or no harm to the validity of the inference
  - This is the best method
- Take a representative subsample of the nonresponse to make inference about the other nonrespondents
- Use a model to predict the nonrespondents such as weights adjustment, imputation and parametric models
- Ignore the nonresponse (not recommended, but unfortunately common in practice)

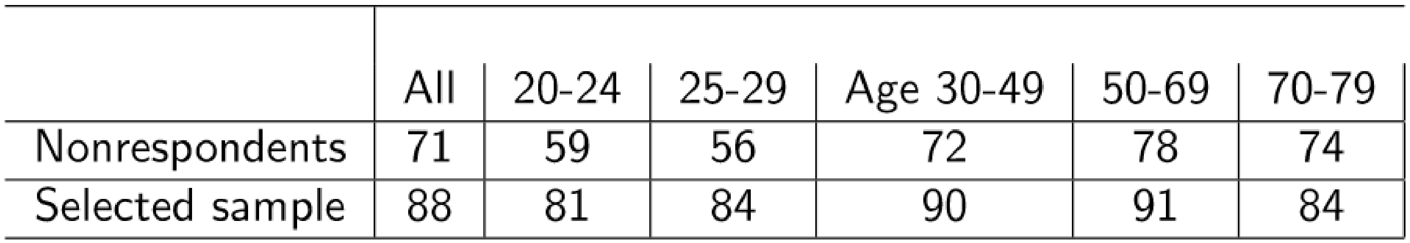
CONSEQUENCES OF IGNORING NONRESPONSE

Example:

- Thomsen and Siring (1983) reported results from a 1969 survey on voting behavior (percentage of persons who voted) carried out by the Central Bureau of Statistics in Norway.
- In that survey, 3 calls were followed by a mail survey.
- The final nonresponse rate was small (9.9%), which often considered to be small nonresponse rate.
- Did the percentage of persons who voted in nonresponse group differ from that in response group?
- Whether a person voted or not could be verified from the voting register.
- The table below shows the results.

VOTING EXAMPLE

Table : Percentage of persons who voted



	All	20-24	25-29	Age 30-49	50-69	70-79
Nonrespondents	71	59	56	72	78	74
Selected sample	88	81	84	90	91	84

Selected sample = all persons selected in the survey

- One can see that 88% of all sampled persons voted, but only 71% of non-respondents voted. (respondent vs non-respondent: 89% vs 71%)
- For person aged 20-24, the difference was even larger (81% vs 59%).
- This shows a nonresponse rate of less than 10% could led to an overestimation of the voting rate.

INCREASE SAMPLE SIZE?

- Increasing sample size without targeting nonresponse cannot mitigate nonresponse bias
- You simply get more observation from the ones who would respond
- the percentage of nonresponse would probably remain the approximately same

REPORT THE RATE OF NONRESPONSE

- Most small surveys ignore any nonresponse that remains after callbacks and follow-ups, and report results based on complete records only.
- Nonresponse is also ignored for many surveys reported in newspapers.
- This relies on the assumptions that the non-respondents are similar to the respondents and that units with missing items are similar to units that have responses for every question.
- However, much evidence indicates that this assumption does not hold true in practice. If you insist on estimating population means and totals using only the complete records and making no adjustment for non-respondents, at the very least you should report the rate of nonresponse.

DESIGN TO PREVENT NONRESPONSE

- The quality of a survey is largely determined at the design stage
- Recall that this is the best to prevent nonresponse
- A common feature of poor surveys is a lack of time spent on the design and nonresponse follow-up in the survey
- Many people new to surveys simply jump in and start collecting data without considering potential problems in the data-collection process

THE EFFECTS OF NONRESPONSE

- A simple example:
- Suppose $n = 100$, with $yes = 63$, $no = 27$, and 10 nonresponse. The nonresponse rate is 10%
  - If we ignore nonresponse, the proportion of 'Yes' would be estimated by: $\hat{p} = \frac{63}{63+27} = 0.70$, with 95% CI approximately $(0.61, 0.79)$.
  - In the worst case, where all nonresponse are 'No'', we would get $\hat{p} = \frac{63}{63+27+10} = 0.63$, with 95% CI approximately $(0.54, 0.72)$.

- However, if the nonresponse rate is 30%, then $n = 100$, with $yes = 49$, $no = 21$, and 30 nonresponse.
  - If we ignore nonresponse, the proportion of 'Yes' would be estimated by: $\hat{p} = \frac{49}{49+21} = 0.70$, with 95% CI approximately $(0.59, 0.81)$.
  - In the worst case, where all nonresponse are 'No'', we would get $\hat{p} = \frac{49}{49+21+30} = 0.49$, with 95% CI approximately $(0.39, 0.59)$.

HOW TO PREVENT

- Most investigators do not know as much about reasons for nonresponse as they think they do
  - They should discover why the nonresponse occurs and resolve as many of the problems as possible before commencing the survey (such as clarity of questionnaires, accessibility of sampling units)
  - Any book on quality control or design of experiments will help you in your data process
  - Previous researcher's experiments help too

DESIGN A PILOT STUDY TO DETERMINE THE FACTORS OF NONRESPONSE

Example:

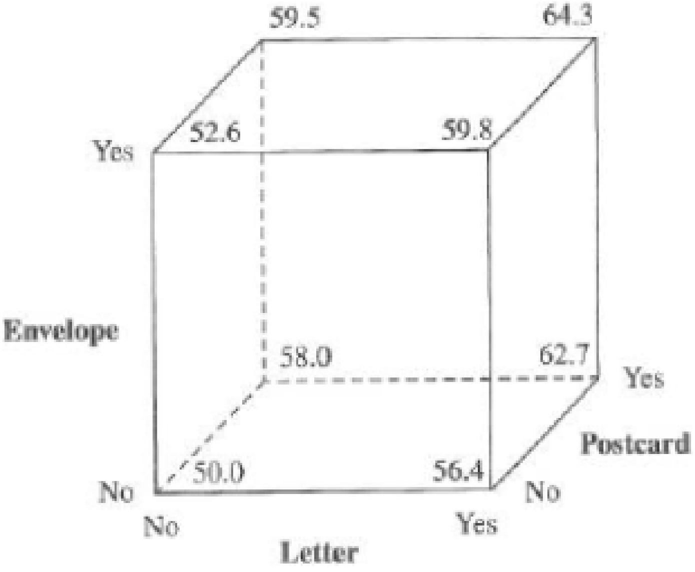
- The 1990 U.S. decennial census attempted to survey each of the over 100 million households in the United States.
- The response rate for the mail survey was 65%.
- Households that did not mail in the questionnaire needed to be contacted in person, adding millions of dollars to the cost of the census.
- Increasing the mail response rate for future censuses would result in tremendous saving.

DESIGN A PILOT STUDY TO DETERMINE THE FACTORS OF NONRESPONSE

- Dillman et al. (1995) reported results of a factorial experiment employed in the 1992 census Implementation Test, designed to explore the individual effects and interactions of the three factors on the response rates.
  - A pre-notice letter alerting the household to the impending arrival of the census form
  - A stamped return envelope included with the census form
  - A reminder postcard sent a few days after the census form
- The results are in the table below.

DESIGN A PILOT STUDY TO DETERMINE THE FACTORS OF NONRESPONSE

FIGURE 8.1
Response rates achieved for each combination of the factors *letter, envelope*, and *postcard*. The observed response rate was 64.3% when all three aids were used and only 50% when none were used.



- The experiment established that while all three factors influenced the response rate, the letter and postcard led to greater gains in response rate than the stamped envelope.

FACTORS INFLUENCING RESPONSE RATE

- Survey content: a survey on sensitive matters (smoking, drug use or financial) may have a large number of refusals. This can be improved by reordering the questions, using self-administered questionnaire or randomized response.
- Time of survey: do you call during holiday, festive season?
- Interviewers: do you provide sufficient training? Variability among interviewers
- Data collection method: telephone, mail, personal, online?
- Questionnaire design: wording, avoid jargons, abbreviations.

FACTORS INFLUENCING RESPONSE RATE

- Respondent burden: persons who respond to a survey are doing you an immense favour, and the survey should be as nonintrusive as possible. Avoid unnecessary details, use short questionnaire, use statistical technique to reduce required sample size.
- Survey introduction: respondent should be told how the data will be used, and assured confidentiality. A good introduction gives people motivation to respond.
- Incentives and disincentives: financial or otherwise, suspension of membership or licenses if refuses to participate (eg. Grab-car)
- Follow-up: send reminders

BIAS DUE TO NONRESPONSE

- Let's see the nonresponse population framework and the source of bias
- Assume that we have a population of N units with population mean

$$\bar{y}_{\mathcal{U}} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- Think of the population being divided into two strata, respondents and non-respondents
- The population respondents are the units that would respond if they were chosen to be in the sample
- Denote the number of population respondents by N_R, which is known

FRAMEWORK FOR NONRESPONSE

- The population parameter

	Size	Total	Mean	Variance
Respondents	N_R	t_R	$\bar{y}_{R\mathcal{U}}$	S_R^2
Non-respondents	N_M	t_M	$\bar{y}_{M\mathcal{U}}$	S_M^2
Entire Population	N	t	$\bar{y}_{\mathcal{U}}$	S^2

- We also have the sample version

	Size	Total	Mean	Variance
Respondents	n_R	$\hat{t}_R = N_R\bar{y}_R$	$\bar{y}_R = \frac{1}{n_R}\sum_{i\in\mathcal{S}_R} y_i$	s_R^2
Non-respondents	$n_M = n - n_R$	$\hat{t}_M$	$\bar{y}_M$	s_M^2
Entire Sample	n	$\hat{t}$	$\bar{y}$	s^2

FRAMEWORK FOR NONRESPONSE

- Relationship among $\bar{y}_{\mathcal{U}}$, $\bar{y}_{\mathcal{U}R}$ and $\bar{y}_{\mathcal{U}M}$:

$$\bar{y}_{\mathcal{U}} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i\in\mathcal{U}} y_i = \frac{1}{N}\left(\sum_{i\in\mathcal{U}_R} y_i + \sum_{i\in\mathcal{U}_M} y_i\right)$$

$$= \frac{1}{N}\left(N_R\frac{\sum_{i\in\mathcal{U}_R} y_i}{N_R} + N_M\frac{\sum_{i\in\mathcal{U}_M} y_i}{N_M}\right)$$

$$= \frac{N_R}{N}\bar{y}_{R\mathcal{U}} + \frac{N_M}{N}\bar{y}_{M\mathcal{U}}$$

FRAMEWORK FOR NONRESPONSE

- From the relationship

$$\bar{y}_{\mathcal{U}} = \frac{N_R}{N}\bar{y}_{R\mathcal{U}} + \frac{N_M}{N}\bar{y}_{M\mathcal{U}}$$

- We can see that the bias occurs if the population mean is different from respondents and non-respondents ($\bar{y}_{R\mathcal{U}} \neq \bar{y}_{M\mathcal{U}}$)
- Let $\bar{y}_R$ be an approximately unbiased estimator of the mean in the respondent's stratum, using only the respondents
- Then,

$$E(\bar{y}_R) - \bar{y}_U \approx \bar{y}_{RU} - \bar{y}_U = \bar{y}_{RU} - \left[\frac{N_R}{N}\bar{y}_{RU} + \frac{N_M}{N}\bar{y}_{MU}\right]$$

$$= \left[\frac{N - N_R}{N}\bar{y}_{RU} - \frac{N_M}{N}\bar{y}_{MU}\right] = \left[\frac{N_M}{N}\bar{y}_{RU} - \frac{N_M}{N}\bar{y}_{MU}\right]$$

FRAMEWORK FOR NONRESPONSE

- 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 determined by two factors
  1. The mean for the non-respondents is close to the mean for the respondents
  2. $\frac{N_M}{N}$ is small, that is, there is little nonresponse
- We generally have no information on the non-respondents to control (#1. above) and thus the only way is to minimize the nonresponse rate.

ADJUSTING SAMPLE SIZE

- If the response rate is low, the actual sample size is low which will result in high variance or big standard error
- It is easy to be remedied by inflating the target sample size
  - Suppose we want to have a target sample size of n=500
  - We guess that the response rate $R = N_R/N$ would be about 80%
  - We then draw a sample of size n/R $(500/0.8 = 625)$so that the response sample size is about n

CALLBACKS AND TWO-PHASE SAMPLING

- Virtually all good surveys rely on callbacks to obtain response from persons not at home for the first try
- Analysis of callback data can provide some information about the biases that can be expected from the remaining non-respondents
- The idea is to subsample the non-respondents and use these data to estimate the population mean for non-respondents $\bar{y}_{MU}$. Then the population mean or total can be estimated
- This is an example of two-phase sampling (also called double sampling, see Chapter 12 of Lohr)

UNBIASED ESTIMATOR FOR THE POPULATION MEAN (1)

- We take an SRS of n units in the population. Of these, n_R respond and n_M do not respond
- Note that n_R and n_M are random variables
- Then make a second SRS of size $100v\%$ from the n_M non-respondents
- Note that the subsampling fraction v does not depend on the data collected.
- The sample mean of respondents is

$$\bar{y}_R = \frac{1}{n_R} \sum_{i \in \mathcal{S}_R} y_i$$

- If all the $n_M \times v$ units respond, the population mean for the non-respondents can be estimated by

$$\bar{y}_M = \frac{1}{v\, n_M} \sum_{i \in \mathcal{S}_M} y_i$$

UNBIASED ESTIMATOR FOR THE POPULATION MEAN (2)

An unbiased estimator for the population mean $\bar{y}_U$ would be

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

$$= \frac{\sum_{i \in \mathcal{S}_R} y_i + \frac{1}{\nu}\sum_{i \in \mathcal{S}_M} y_i}{n}$$

and

$$\hat{t} = N\hat{\bar{y}} = \frac{N}{n}\sum_{i \in \mathcal{S}_R} y_i + \frac{N}{n\nu}\sum_{i \in \mathcal{S}_M} y_i$$

Note that the weight *N/n* and *N/(nv)* are different because only a subsample was taken in the non-respondents stratum (more weights)

VARIANCE ESTIMATE

- If the finite population correction factor (fpc) can be ignore, the estimator for the variance for the population mean is

$$Var(\hat{\bar{y}}) = \frac{n_R - 1}{n - 1} \cdot \frac{s_R^2}{n} + \frac{n_M - 1}{n - 1} \cdot \frac{s_M^2}{nv}$$

$$+ \frac{1}{n - 1}\left[\frac{n_R}{n}(\bar{y}_R - \hat{\bar{y}})^2 + \frac{n_M}{n}(\bar{y}_M - \hat{\bar{y}})^2\right]$$

- The formula is from the general results of two-phase sampling for stratification
- If everyone response in the subsample, two-phase sampling not only removes the nonresponse bias but also accounts for the original nonresponse in the estimated variance

See Chapter 12, page 475 of Lohr.

IMPUTATION METHOD

- Missing items may occur in surveys for several reasons
  - An interviewer may fail to ask a question
  - A respondent may refuse to answer the question or cannot provide the information
- In a questionnaire, many questions are usually asked. If the respondent fails to answer one question (such as his/her age), it seems a huge waste to throw the whole dataset away
- **Imputation** is commonly used to assign values to the missing items
- A replacement value, often from another person (or an average of several persons) in the survey who is similar to the item non-respondent on other variables, is imputed for the missing value
- we will only discuss some examples, and details will be omitted

AN EXAMPLE FOR IMPUTATION

- The CPS has an overall high household response rate, but some households refuse to answer certain questions.
- The nonresponse rate is about 20% on many income questions.
- Item nonresponse for the income items is highest for low- and high-income household.
- Imputation for the missing data makes it possible to use standard statistical techniques without using special methods.

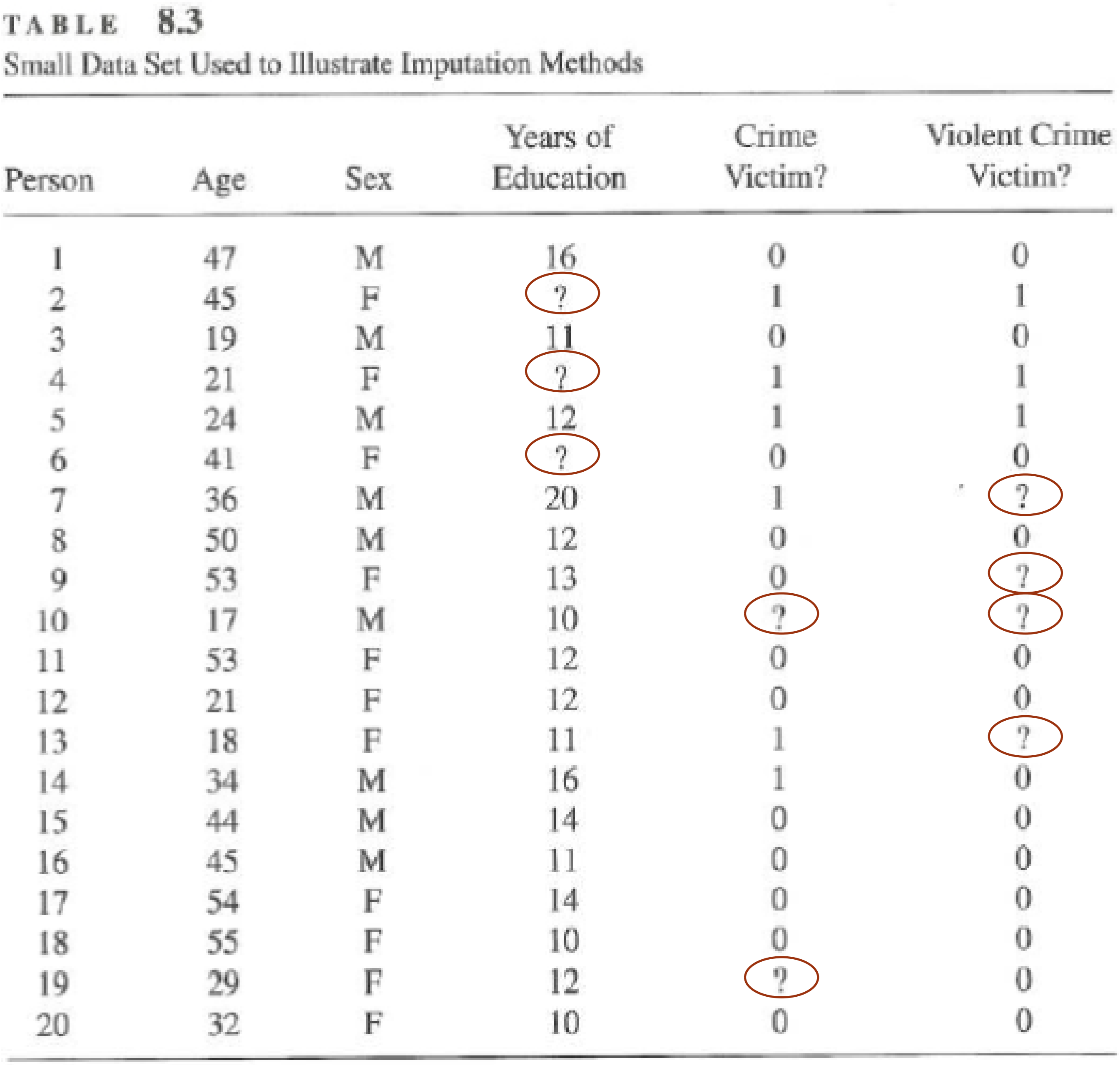


TABLE 8.3

Small Data Set Used to Illustrate Imputation Methods

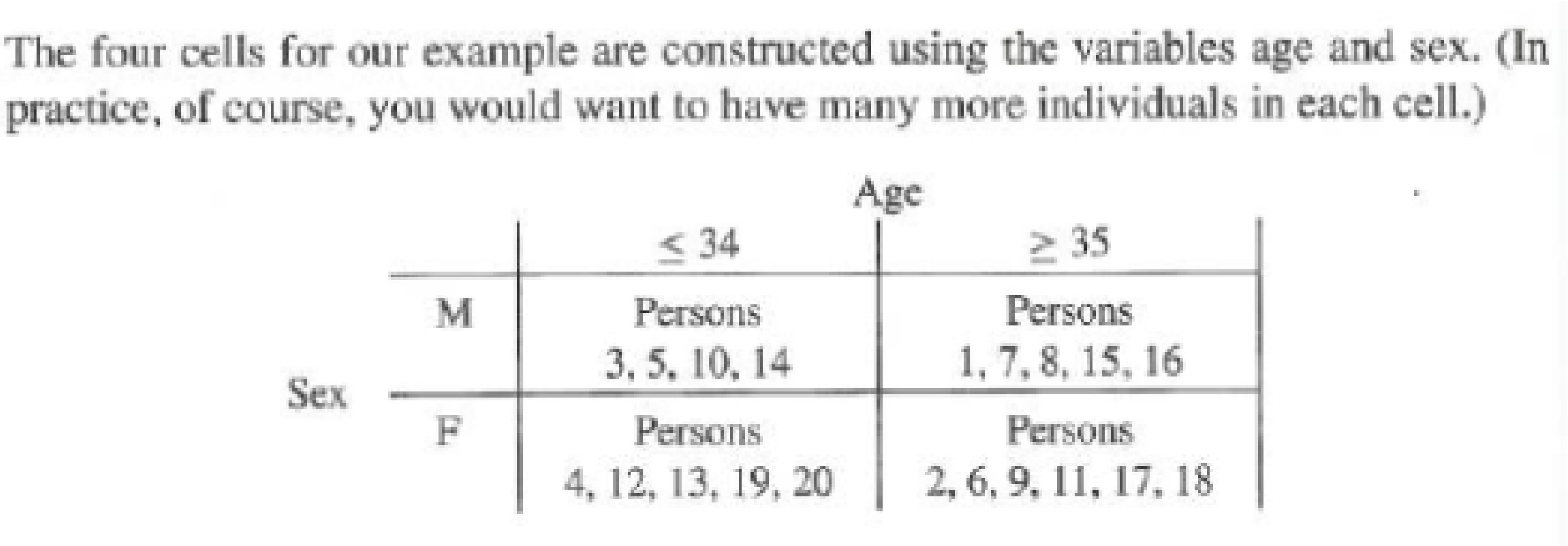
Person	Age	Sex	Years of Education	Crime Victim?	Violent Crime Victim?
1	47	M	16	0	0
2	45	F	?	1	1
3	19	M	11	0	0
4	21	F	?	1	1
5	24	M	12	1	1
6	41	F	?	0	0
7	36	M	20	1	?
8	50	M	12	0	0
9	53	F	13	0	?
10	17	M	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	M	16	1	0
15	44	M	14	0	0
16	45	M	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

DEDUCTIVE IMPUTATION

- Some values may be imputed in the data, using logical relation among the variables.
- Person #9 is missing the response for whether she was a victim of violent crime. But, she had responded that she was not a victim of any crime, so the violent crime response should be changed to 0.
- This technique is also useful for longitudinal surveys. If a woman has two children in year 1 and two children in year 3, but is missing the value for year 2, the logical value to impute would be two.

CELL MEAN IMPUTATION

- The sample is divided into classes (cells) using known variables such as sex, age, race, and other demographic characteristics.
- Then the average of the values for the responding units in the cell will be used to substitute the missing value.



The four cells for our example are constructed using the variables age and sex. (In practice, of course, you would want to have many more individuals in each cell.)

		Age	
		≤ 34	≥ 35
Sex	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, missing the value for years of education, would be assigned the mean value (12.25) for the four women aged 35 or older who responded to the questions.

OTHER IMPUTATION METHODS

- Hot-deck imputation (using the last recorded value, random, or nearest neighbour).
- Cold-deck imputation (using previous survey or other information).
- Regression imputation (predict the missing value using a regression of the item of interest on variable observed for all cases). Linear or logistic regression.
- Cautions:
  - Imputation creates "clean" data set for analysis.
  - Future data analysis may not be able to distinguish between the original and imputed values.
  - Imputed data may be good guesses, but they are not real.
  - The true variance will be larger than that estimated using imputed data.

RANDOMIZED RESPONSE TO SENSITIVE QUESTIONS

- Sometimes you want to conduct a survey asking some sensitive questions such as
  - Do you have AIDS?
  - Have you had any extramarital affairs?
  - Do you use cocaine?
  - Have you ever shoplifted?
  - Have you ever cheated in the exams?
  - Did you understate your income on your tax return?
- In these circumstances, you may expect that many people will find it very uncomfortable to answer such questions and so may choose not to answer them, or if they do, the answers may be very evasive
- In these cases, we may resort to **randomized responses**

HOW?

- We shall illustrate how to do this by an example
- Example: the respondent throws a coin with $Pr(head) = P$
- But, the interviewer does not know the outcome of coin tossing
  - If it is head, he answers the original (sensitive) question (e.g. have you ever cheated on an exam?)
  - If it is tail, he answers a totally unrelated (innocent) question (e.g. were you born in July?)

RELEVANT PROBABILITY

- Let

  $P = \Pr(\text{asked sensitive question})$

  $p_s = \Pr(\text{say "Yes" | asked sensitive question})$

  $p_I = \Pr(\text{say "Yes" | asked innocent question})$

  $n = \text{total number of people being asked, that is, sample size}$

- Therefore,

  $\psi = \Pr(\text{respondent replies "Yes" })$

  $= \Pr(\text{"Yes" | asked sensitive question})\times\Pr(\text{asked sensitive question})$
  $\quad +\Pr(\text{"Yes" | asked innocent question})\times\Pr(\text{asked innocent question})$

  $= p_s \times P + p_I \times (1 - P)$

- Hence,

$$p_s = \frac{\psi - (1 - P)\, p_I}{P}$$

ESTIMATOR OF THE PROPORTION

- Here, P and p_I are known, but ψ is unknown
- ψ can be estimated from the sample, denoted by

  $\hat{\psi}$ = estimated proportions of "Yes" from the sample
- Thus,

$$\hat{p}_s = \frac{\hat{\psi} - (1 - P)\, p_I}{P}$$

- Clearly,

$$E(\hat{p}_s) = p_s, \qquad \text{and}$$

$$Var(\hat{p}_s) = \frac{Var(\hat{\psi})}{P^2}$$

OTHER COMMENTS

- The “penalty” for randomized response appears in the factor $\frac{1}{P^2}$ in the variance expression.
- If $p = \frac{1}{3}$ for example, the standard deviation is 3 times as great as it would have been, had everyone in the sample being asked the sensitive question and responded truthfully
- One needs to think before choosing P : the larger the P is, the smaller the variance of $\hat{p}_S$. But if P is too large, respondents may think that the interviewer will know which question is being answered
- Some respondents may think that $P = 0.5$ is fair
- Instead of throwing a coin, respondent could also be asked to draw a card from a deck of 52 cards. If it is “heart” or a “spade”, he/she answers the sensitive question. Otherwise, he/she answers the innocent question