Statistics 522: Sampling and Survey Techniques

Topic 8

Topic Overview

This topic will cover

- Nonresponse
- Missing data
- Design Issues

Unit nonresponse

- The entire observation is missing.
- A person may refuse to be interviewed.
- 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.
- A person may refuse to answer a particular question

Wildlife surveys (Missing data)

- Some birds may not be found by the researchers; nonrespondents
- A nest may be raided by predators before the investigator can determine how many eggs were laid; item nonresponse.

Four Approaches

1. Prevent it by design (best).

   > "Reduce the nonresponse to insignificant levels, so that any remaining nonresponse causes little or no harm to the validity of the inferences."

2. Take a representative subsample of the nonrespondents.
3. Use a model to predict values for the nonrespondents.
   - Implicit adjustment through weights

    - Imputation
    - Use auxiliary data

4. Ignore the nonresponse
    - Most common
    - Is non-response ignorable?

Effects of ignoring nonresponse

Example 8.1, page 256

- 1969 survey of voting behavior in Norway, e.g. estimate voting rate
- Three phone calls followed by a mail survey
- The nonresponse rate was 9.9%

Voting

- The Norwegian voter register is used to determine who voted.
- 88% of the respondents voted.
- 71% of the nonrespondents voted.
- For persons aged 20-24, it was 81% vs 59%
- Not-at-homes (65% vote) and illness (55% vote) are major causes of nonresponse bias.

Nonresponse in the UK

Lower response rates associated with

- London residents
- Households with no car
- Single people
- Older people
- Divorced and widowed people
- new Commonwealth origin
- lower educational attainment
- self-employed

Increase n?

- Using a larger sample size without targeting the nonresponse will have no effect on the nonresponse bias
- Recall the Literary Digest Survey of 1936, page 7.
  - 2.4 million people
  - 25% response rate
  - predicted Alf Landon would be president with 55% of the vote

The US Census

- Undercoverage rate varies for different demographic groups.
- Some cities filed a lawsuit in the early 1990's.
- "the debate … continues"

Ignoring nonresponse

- Implicitly assumes that the nonrespondents are similar to the respondents; similarly for items
- Evidence suggests that this is not true in general.
- Results reported should be taken as estimating the population of respondents.
- Report the nonresponse rate.

Bias

- Think of the population being divided into two strata, respondents and nonrespondents.
- See page 258 for some equations

$$bias = \mathrm{E}(\bar{y}_{resp}) - \bar{y}_{pop} \approx \frac{N_{missing}}{N}(\bar{y}_{resp,pop} - \bar{y}_{miss,pop})$$

- Bias increases with nonresponse rate
- Small bias
  - Means are close
  - Small nonresponse rate
- Variance estimate often too small

Design issues

- Nonresponse is a neglected problem.
- Surveys reported in many areas have very low response rates, 10-15%.
- To learn about reasons for nonresponse or sources of error, use designed experiments and quality-improvement methods.

1990 Census

- Attempted to survey each of over 100 million households
- Response rate for the mail survey was 65%
- For households not responding to the mail survey, an attempt at a personal contact was made (very expensive).

An experiment

- Factor A: prenotice letter (Y, N)
- Factor B: stamped return envelope (Y, N)
- Factor C: reminder sent a few days after the first mailing (Y, N)
- Response rate was 50% for NNN (nothing used)
- Rate was 64.3% for YYY (all 3 used)

Factors that can influence response rate and data accuracy

- Survey content
  - personal info
  - helped by reordering
- Time of survey
  - Call during Spring Break?
- Interviewers
- Data-collection method
  - telephone vs mail vs personal
- Questionnaire design
  - appeal to cognitive science

- Respondent burden
- Survey introduction
  - why data used
  - assured confidentiality
- Incentives and *disincentives*
- Follow-up
- The attitude of management toward nonresponse.

Quotes

- Lohr: “The quality of a survey is largely determined at the design stage.”
- Fisher: “To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.”

Callbacks

- “Virtually all good surveys rely on callbacks to obtain responses 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 nonrespondents.”
- Can be more effective to take a subsample of nonrespondents

Example 8.3

- Analysis of callback data from two 1984 Michigan polls on preference for presidential candidates
- Response rate 65%
- 21% responded on first call
- up to 30 additional calls

Results

- Later respondents were more likely to be
  - male
  - older
  - Republican
- First call: 48% Reagan, 45% Mondale
- Entire sample: 59% Reagan, 39% Mondale

Polls

- Some of the variation in results for polls is due to different procedures for follow-up
- We would like to assume that late respondents are like nonrespondents.

Stratification

- Hansen and Hurwitz (1946) proposed subsampling nonrespondents and using *two-phase* or *double sampling.*
- Strata are initial respondents and initial nonrespondents.

Example

- Suppose the initial sample size was 1000.
- Initial respondents were $n_R = 800$, nonrespondents (missing) were $n_M = 200$
- Note that n_R and n_M are random variables.
- Sample vn_M of the nonrespondents.
  - v determined beforehand
- Let $\bar{y}_R$ and $\bar{y}_M$ be the means.

$$\hat{\bar{y}} = \frac{800}{1000}\bar{y}_R + \frac{200}{1000}\bar{y}_M$$

Total and variance

- To estimate the total, multiply by N.
- See page 263 for formula for the variance (based material in Chapter 12)

Perspectives on nonresponse

- Cochran (1977)

  > "The division into two distinct strata is, of course, an oversimplification. Chance plays a part in determining whether a unit is found and measured in a given number of attempts. In a more complete specification of the problem, we would attach to each unit a probability representing the chance that it would be measured by a given field method if it fell in the sample."

- "Much of the recent work on estimation in the presence of nonresponse has used the idea expressed in this citation – namely, that the response is stochastic, not deterministic. Under such an outlook, a response distribution is assumed to exist. The known sample selection distribution and the unknown response distribution together form the basis of a randomization theory extended to incorporate nonresponse."

- Conditions of the survey – Dalenius (1983):

  > " . . . it appears utterly unrealistic to postulate fixed 'response probabilities' which are independent of the varying circumstances under which an effort is made to elicit a response. In other words, there is no single unique q value, but possibly a wide spectrum of q values."

- Kalton (1983)

  > "Sampling practitioners do not believe that the nonresponse models on which their adjustments are based hold exactly: they simply hope that they are improvements on the model of data missing at random."

Models for nonresponse

Difficulties due to nonresponse

- We do not know how the response sets are generated.
  - Differs from sampling design
  - Must make assumptions: independence or correlation among similar elements
  - Can't verify these assumptions
- Calculations are more difficult – "heavier".

Definitions

- *Response model* – set of assumptions about true unknown response distribution.
- *Response distribution* – Underlying probability mechanism which produces nonresponse.

- We need a model to adjust for nonresponse.
- Let ϕ_i be the probability that item i will respond (the *propensity* score).
- Let x_i be the data vector for item i.

Three models

1. Missing completely at random (MCAR)
   - ϕ_i does not depend on x_i.
2. Missing at random given covariates or ignorable nonresponse
3. Nonignorable nonresponse

MCAR

- Missing completely at random
- Data entered incorrectly into computer file and appears as an outlier.
- Lab problem with reagents
- MCAR is implicitly adopted when nonresponse is ignored.

MAR

- Missing at random given covariates.
- ϕ_i the probability that item i will respond can depend on an element of the data vector x_i but not on y_i.
- We can use a model here.
- This is sometimes called *ignorable nonresponse.*
- Ignorable means that the model works.

Nonignorable nonresponse

- The probability of nonresponse ϕ_i
  - depends on y_i
  - cannot be completely explained by the x_i
- NCVS
  - suppose a person victimized by a crime is less likely to respond to the survey

Weights

Weights are the reciprocal of the probabilities of selection.

- for stratification we used $w_i = N_h/n_h$
- for unequal probability sampling we used $w_i = 1/\pi_i$.

Weights for nonresponse

- Let π_i be the probability of selection in the sample.
- Let ϕ_i be the probability that the unit will respond.
- Then $\pi_i\phi_i$ is the probability that the unit is selected and responds
    - This assumes that selection and response are independent

Estimated weights

- We use the data to estimate the probability of response $\hat{\phi}_i$
- Use weights $w_i = 1/(\pi_i\hat{\phi}_i)$
- This method assumes MAR.

Weight class adjustment

- Form subgroups of the population
- For an SRS, use the sample proportion of responses in the subgroup as $\hat{\phi}_i$.
- In general, use the ratio of the sum of the weights for the respondents divided by the sum of the weights for the selected sample (in each subgroup).

$$\tilde{w_i} = \begin{cases} \sum \text{classes } c \, \frac{w_i I(i \in \text{ class } c)}{\hat{\phi}_c} \\ 0 \text{ if nonrespondent} \end{cases}$$

Assumptions

To apply the weight class adjustment, we assume

- the probability of response is the same for all members of the subgroup
- MAR (within each subgroup)

Estimates

- Use the usual formulas with weights $w_i = 1/(\pi_i \hat{\phi}_i)$.
- More detailed formulas are on page 267.

$$\begin{aligned}\hat{t} &= \sum \tilde{w}_i y_i \\ \hat{\bar{y}} &= \frac{\hat{t}}{\sum \tilde{w}_i}\end{aligned}$$

Example 8.5 (NCVS)

- The adjustment for nonresponse is the within-household noninterview adjustment factor (WHHNAF) described in Chapter 7 on page 245
- 24 weighting cells are used
- These are based on demographic variables.

Adjustments

- If the weight adjustment factor is too large (greater than 2)
- Or if the number of respondents in a cell is less than 30
- Cells are combined with neighboring cells

Logistic regression

- Logistic regression (or classification methods) can be used to get estimated probabilities of response for each unit.
- Recommendation is to use these estimated probabilities to construct subgroups with similar probabilities of response.

Post-stratification

- Weight-adjusted classes should be constructed like strata.
- Method is similar to weight class adjustment but population counts are used to adjust the weights.

$$\begin{aligned}\bar{y}_{post} &= \sum \left(\frac{N_h}{N}\right) \bar{y}_{h,R} \\ \bar{y}_{wei} &= \sum \left(\frac{n_h}{n}\right) \bar{y}_{h,R}\end{aligned}$$

- With the weighted estimator N_h is unknown and is estimated by $N(n_h/n)$

Assumptions

- Within each poststratum unit, each unit selected to be in the sample has the same probability of being a respondent.
- Response/nonresponse is independent across units.
- Nonrespondents are like respondents.
- The data are MCAR within each stratum.

Example (NCVS)

- In the second stage poststratification is used to adjust the weights.
- This is SSF (see Chapter 7, page 246)
- 72 post strata are based on age race and sex and the values of N_h are taken from the Bureau of Census counts

Raking adjustments

- Used when poststrata are formed using more than one variable (e.g., race and sex) but only the marginal population totals for each variable are known.
- Iterative adjustment procedure
- See Section 8.5.2.2 on pages 269-271

Other methods

- Estimate the probability of nonresponse by asking the respondent something such as whether or not they were home on the four preceding weeknights.
- Politz-Simmons method
- See Section 8.5.3 on page 271 for details

Final Word of Caution

- Weighting adjustments uses lots of assumptions.
- May be implausible
- Should state assumed model and justify
- Usually used for unit non-response.

Imputation

- Used for item nonresponse
- Values are assigned for missing items
- Often we use a value from a similar item
- Data set should include an indicator for the imputation.
- Goals
  - reduce nonresponse bias
  - create a clean data set

General criteria

- Imputed values should be consistent.
- Should reduce nonresponse bias and preserve relationships between items as far as possible
- Suitable for any pattern of missing items
- Set up ahead of time
- Effects on bias and prediction of the estimate can be evaluated.

Deductive imputation

- Use common sense
- Table 8.3, person 9 responded that she was not a victim of any crime but did not respond to the question about violent crime.
- (Duplicate records)
- Study the data.
- Replace missing values by the last present value for longitudinal studies.

Cell mean imputation

- Divide the sample into subgroups
- Compute the mean for the respondents in each subgroup.
- For non respondents, substitute this mean for the response
- Implicitly assumes MAR (MCAR within the subgroups)

Evaluation

- Mean imputation gives the same estimates for means, totals, and proportions as the weighting class adjustments.
- Variances can be very different.
- Estimation of relationships can be distorted.

Hot-deck imputation

- Sample units are divided into subgroups or classes.
- Replace missing values by values from a similar unit.
- With computer 'cards' or sequential files, use the previous record.
- Or randomly select a unit from the same class.
- Or select the 'closest' unit

Regression imputation

Use units with complete data for the variables in question to determine a regression equation for estimation of missing values.

Cold-deck imputation

Use values from a previous study to replace missing values.

Substitution

- Used for unit nonresponse
- Randomly select from a predetermined a list of substitutes
- See text for evaluation

Problems with imputation

- Example SRS with 50% nonresponse
- Assume MAR
- Impute the mean of the respondents; replace values with the mean
- What happens to the variance?
- And the standard error?

Multiple imputation

- A possible solution to the variance problem
- Impute values plus random noise
- Do this several times
- Combine results using standard statistical methods
- Most interesting theory

Parametric models for nonresponse

- Model for complete data with nonresponse mechanism
- Advantages
  - flexible, particular when model nonresponse
  - Assumptions must be stated and evaluated.
  - Variance estimation can account for nonresponse.
- Often use likelihood-based methods
- Spotted owls in Washington, Oregon and California
- Read Example 8.10 on pages 278-280

Acceptable response rate

- Hard to get objective measures of bias, but can quantify nonresponse
- However, there are different measures that can be declared.

$$\frac{\sum_{response} 1/\text{probability of being sampled}}{\sum_{sample} 1/\text{probability of being sampled}}$$

$$\frac{\text{number of completed responses}}{\text{number of units in sample}}$$

$$\frac{\text{number of completed responses}}{\text{number of units contacted}}$$

$$\frac{\text{completed responses}+\text{ineligible units}}{\text{contacted units}}$$

$$\frac{\text{completed responses}}{\text{contacted units}-(\text{ineligible units})}$$

$$\frac{\text{completed responses}}{\text{contacted units}-(\text{ineligible units})-\text{refusals}}$$

$$\frac{\sum_{response} x_k}{\sum_{sample} x_k}$$

$$\frac{\sum_{response} x_k/\pi_k}{\sum_{sample} x_k/\pi_k}$$

- Be aware that response rates can be manipulated
- Read Section 8.8
- See recommendations from Maria Gonzales *et al* (1994) on page 282