36-303: Sampling, Surveys and Society

Quality in Surveys
Brian Junker
132E Baker Hall
brian@stat.cmu.edu

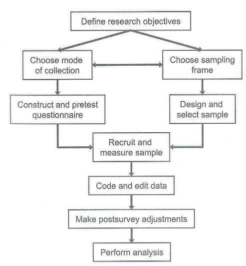
Handouts

- Lecture Notes
- News Stories

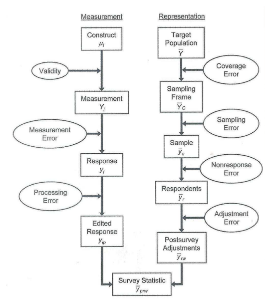
Outline

- Quality in Surveys
- More on the Project Outline Handout
- Reading:
  - Up to today: responsible for Groves Ch's 1, 2
    - Save Groves Ch 3 for later
  - Next week: Groves Ch 5, 11, & 4, in that order
- Lecture notes online at
  http://www.stat.cmu.edu/~brian/303

Quality in Surveys

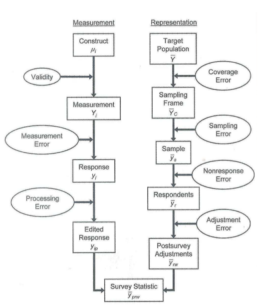


Process Perspective on Surveys



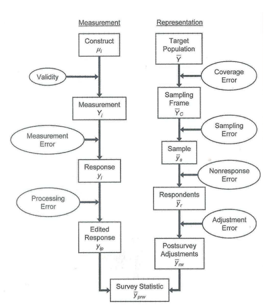
Quality Perspective on Surveys

Quality Overview



- *Total Survey Error*
  - Each of the Quality Components has a verbal description and a statistical formulation
  - The Quality Components are properties of individual survey design and analysis decisions, not of whole surveys
- Our job is to make decisions to minimize error / maximize quality

Measurement Quality



- Working down the left side:
  - *Validity*
  - *Measurement Error*
  - *Processing Error*

Some Notation…

- μ_i = value of the ***construct***. E.g. # of doctor visits for i[th] person in population, i=1, …, N
- Y_i = ***ideal value*** of the ***measurement*** for the i[th] person in the sample, i=1, …, n
- y_i = ***observed value*** (reported number of doctor visits) for i[th] sample person
- y_{ip} = ***observed value after editing/processing***
- y_{it} = value on the t[th] "trial" (t[th] time we run the survey)

Validity

- $Y_i = \mu_i + \epsilon_i$
  - μ_i is the "true value" for the population
  - Y_i is the "ideal measured" value
  - ϵ_i is how much Y_i "deviates" from μ_i
  - Deviation/error is natural. We just have to account for it
- If there are T trials (repeats of the survey), t=1, …, T, we might write

  $Y_{it} = \mu_i + \epsilon_{it}$

  And expect that the errors ϵ_{it} would "average out" over trials…
- A measure of the size of the errors ϵ_i is

  $\mathrm{Corr}(Y_i, \mu_i)$

  This correlation is a measure of the Validity of the measurement

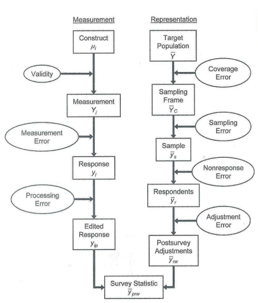
Measurement Error

- $y_i - Y_i$ is the measurement error
  - Y_i is the ideal measurement
  - y_i is the observed measurement
- There are two kinds of measurement error to worry about
  - ***Variability***: $y_i = Y_i + \text{error}_i$, and the error "averages out" over repeated trials: $E_t[y_{it}] = Y_i$
  - ***Bias***: $y_i = Y_i$ + something that doesn't "average out": $E_t[y_{it}] \neq Y_i$

Processing Error

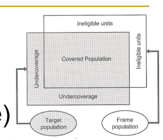
- $y_{ip} - y_i$ is the processing error
  - y_{ip} is the response after editing/processing
  - y_i is the 'raw' response to the measurement
- These errors come in when you have to code, check, or fix survey responses, e.g.
  - Coding a verbal response
  - Range check – can this person have been in High School for 7 years?
  - Clumping, e.g. "income between $10,000 and $30,000"
- These are generally ***bias*** and not ***variability*** issues

Representation Quality



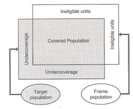
- Working down the right side:
  - *Coverage Error*
  - *Sampling Error*
  - *Nonresponse Error* (later lecture)
  - *Adjustment Error*

Coverage Error



- N = total Target Population (size)
- C = target population covered in frame
- U = target population missed by frame
- $\bar{Y}$ = mean of target population
- $\bar{Y}_C$ = mean of covered population
- $\bar{Y}_U$ = mean of uncovered population
- $\bar{Y}_C - \bar{Y}$ = ***coverage error***
  - Also called ***Coverage Bias***

Coverage Error (Cont'd)



$$\boxed{\overline{Y}_C - \overline{Y} = \frac{U}{N}(\overline{Y}_C - \overline{Y}_U)}$$

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{N}\left(\sum_C Y_i^C + \sum_U Y_i^U\right)$$

$$\begin{aligned}
\overline{Y}_C - \overline{Y} &= \frac{1}{C}\sum_C Y_i^C - \frac{1}{N}\sum_{i=1}^{N} Y_i \\
&= \frac{1}{C}\sum_C Y_i^C - \frac{1}{N}\left(\sum_C Y_i^C + \sum_U Y_i^U\right) \\
&= \left(\frac{1}{C} - \frac{1}{N}\right)\sum_C Y_i^C - \frac{1}{N}\sum_U Y_i^U \\
&= \frac{U}{NC}\sum_C Y_i^C - \frac{U}{N}\cdot\frac{1}{U}\sum_U Y_i^U \\
&= \frac{U}{N}(\overline{Y}_C - \overline{Y}_U)
\end{aligned}$$

Coverage Error/Coverage Bias

- Suppose we are interested in Monthy Mortgage Payment ($0 if you rent)
  - Total population is all adults in (US/Pgh/…)
  - Data collection method is random digit dialling
  - Sampling frame is callable land-line phone #'s
- Renters may be more likely to have only a cell phone than homeowners
  - Renters are undercovered by our frame
  - Our estimate of mean mortgage payment will be too high
  - If we can get an estimate of $\frac{U}{N}(\overline{Y}_C - \overline{Y}_U)$
    Then we can estimate $\overline{Y}_C - \overline{Y}$ and fix the bias!

Sampling Error

- How well does the sample represent the sampling frame?
  - Sampling bias
    - Best to try to anticipate and avoid
    - Can be looked at similarly to coverage bias
    - Another way to deal with is with weights, but this can introduce "adjustment error" (more in a couple pages)
  - Sampling variability – this is a more familiar issue! (see next page)

Sampling Variability

- $\overline{y}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} y_{si}$ is the *mean of the sample*
- $\overline{Y}_C = \frac{1}{C}\sum_C Y_i^C$ is the *mean of the frame*

The *Standard Error* for estimating $\overline{Y}_C$ with $\overline{y}_s$ is

$$SE = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\overline{y}_s - \overline{Y}_C)^2}$$

in case of *simple random sampling* (next week!) we know that

$$SE = SD/\sqrt{n_s} \approx \frac{\sqrt{\frac{1}{n_s-1}\sum_{i=1}^{n_s}(y_{si} - \overline{y}_s)^2}}{\sqrt{n_s}}$$

Adjustment Error

- This usually comes in the forms of weights.
- If the proportion of units in the sample is systematically different from the population, we may weight each unit:

$$\overline{y}_w = \frac{\sum_{i=1}^{n_s} w_i y_i}{\sum_{i=1}^{n_s} w_i}$$

- The main issues are (again) bias and variability of this estimate $\overline{y}_w - \overline{Y}$

More on the Project Outline Handout

- We will go over some parts of the handout now
- This is your chance to ask questions about any parts of the handout that you read, and are concerned about.

Review

- Quality in Surveys
- More on the Project Outline Handout
- Reading:
  - Up to today: responsible for Groves Ch's 1, 2
    - Save Groves Ch 3 for later
  - Next week: Groves Ch 5, 11, & 4, in that order
- Lecture notes online at
  http://www.stat.cmu.edu/~brian/303

Review