36-303: Sampling, Surveys and Society

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

Handouts

- Today's Articles:
  - Weight, Weight, Don't Tell Me
  - Commercial On-Line Polls and Total Survey Quality
- Lecture Notes

Outline

- TA Office Hours Poll
- Due Next Tuesday:
  - Project Proposals: I.1 on the "Project Schedule" handout.
  - HW01 (find it on http://www.stat.cmu.edu/~brian/303)
- Quality in Surveys
- Reading:
  - Up to today: responsible for Groves Ch's 1, 2, 3
  - Next week:
    - Groves Ch 5
    - Groves Ch 11 (sections 1-6)

    in that order
- Guest Lecturer Next Thu:
  - Dr. Julia Kaufman, on a new technique for writing survey questions

TA Office Hours Poll (Census!)

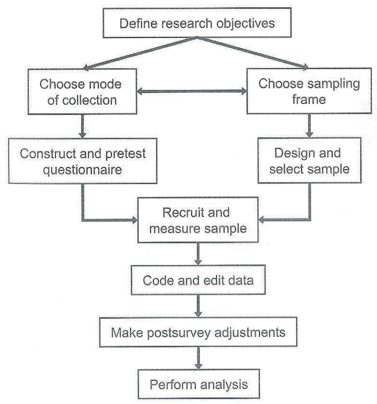
- Put your name on a piece of paper, and write your First and Second choices for office hours:
  - Monday 4-5
  - Monday 5-6
  - Tuesday 5:30-6:30
  - Wednesday 4-5
  - Wednesday 5-6
- If you ***absolutely can't*** make any of these times, write down two hours (first and second choices) during the week that you *can* make.

Q&A on the Project Outline Handout

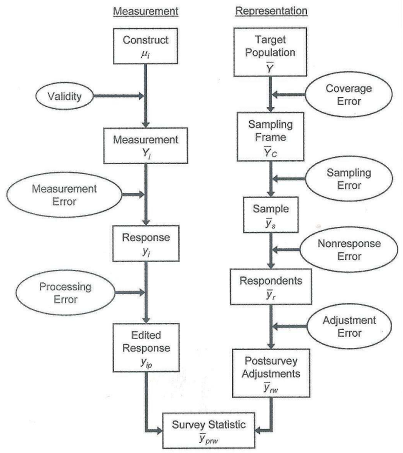
- Posted on www.stat.cmu.edu/~brian/303:
  - Project Outline
  - Some Examples of Project Proposals

- Questions about the projects or teams right now?

Quality in Surveys

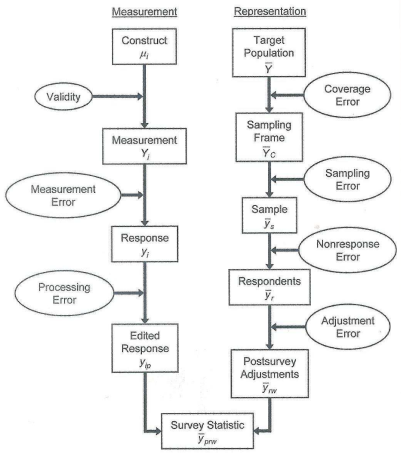

Define research objectives
Choose mode of collection
Choose sampling frame
Construct and pretest questionnaire
Design and select sample
Recruit and measure sample
Code and edit data
Make postsurvey adjustments
Perform analysis

Process Perspective on Surveys



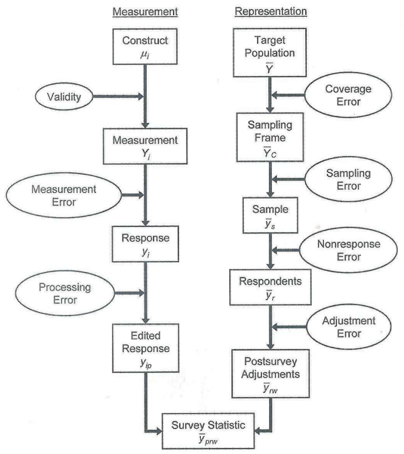
Quality Perspective on Surveys

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 ith person in population, i=1, …, N
- Y_i = *ideal value* of the *measurement* for the ith person in the sample, i=1, …, n
- y_i = *observed value* (reported number of doctor visits) for ith sample person
- y_{ip} = *observed value after editing/processing*
- y_{it} = value on the tth "trial" (tth 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

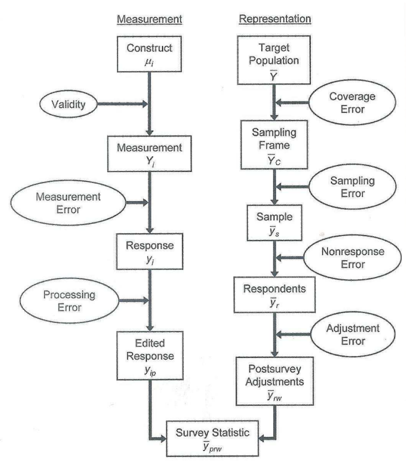
Measurement Error

- $y_i – Y_i$ 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*: $y_i = Y_i$ + 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

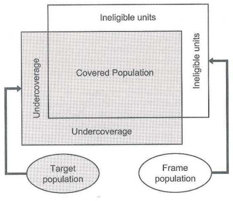
- y_{ip} - y_i is the processing error
  - y_{ip} is the response after editing/processing
  - y_i is the 'raw' response to the measurement
- These errors come in when you have to code, check, or fix survey responses, e.g.
  - Coding a verbal response
  - Range check – can this person have been in High 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*

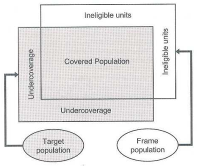
Coverage Error



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

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

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

Adjustment Error

- This usually comes in the form 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 $\overline{y}_w - \overline{Y}$

Review

- TA Office Hours Poll
- Due Next Tuesday:
  - Project Proposals: I.1 on the “Project Schedule” handout.
  - HW01 (find it on http://www.stat.cmu.edu/~brian/303)
- Quality in Surveys
- Reading:
  - Up to today: responsible for Groves Ch’s 1, 2, 3
  - Next week:
    - Groves Ch 5
    - Groves Ch 11 (sections 1-6)

    in that order
- Guest Lecturer Next Thu:
  - Dr. Julia Kaufman, on a new technique for writing survey questions