36-303: Sampling, Surveys & Society

Post-survey Processing
Brian Junker
132E Baker Hall
brian@stat.cmu.edu

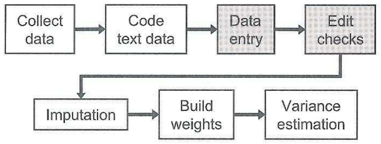
Handouts

- These Lecture Notes
- Group Evaluations (really, this time!)
- Homework 3 [Due Thurs Mar 27]
- Cluster Sampling, Part II (no handout today – we will come back to this next week!)

Outline

- Post-survey Processing [Groves Ch 10]
  - Coding
  - Weighting
  - Imputation
  - Variance Estimation
- Review

Post-survey Processing



- *Top row:* Raw data collection process
  - The order of Coding, Data Entry and Editing will depend on the data collection design (FTF, phone, www, computer assisted, …)
  - Computer-based surveys require you to design the Data Entry and Edit Checks when you build the form in surveymonkey.com, questionpro.com, etc.
- *Bottom row:* Calculations based on the data and/or design

Coding

- Translating non-quantitative or non-categorical data into quantities and categories
- M/C and Likert items *usually* require no coding
  - Indicate your status (check one box only):
    - Full-time student
    - Part-time student
    - Applicant, acceptance letter received
    - Applicant, acceptance letter not received
- When might an M/C item require some coding anyway?

Coding

- Short answer, long answer, graphical response, performance, etc., all require some coding
- On the job crime in the NCVS:
  - What is the name of the (company/government agency/business/nonprofit organization) for which you worked at the time of the accident?
  - What kind of business or industry is this? *(What do they make or do where you worked at the time of the incident?)*
  - What kind of work did you do; that is, what was your occupation at the time of the incident? *(For example: plumber, typist, farmer)*
  - What were your usual activities or duties at this job?

Coding System (Code Structure, Rubric)

- **Each code should include:**
  - A *number or category*, used in statistical analysis
  - A *text label*, describing all answers in that category
- **The set of codes for a response should be:**
  - *Exhaustive:* Every response should be codable into one of the categories
    - Separate codes are needed for skipped, not-asked, off-topic, etc.
  - *Exclusive:* No response should be codable into more than one category
  - *Appropriate* to the purposes of the research
    - Use codes that help you answer your research question(s)
    - If you have more than one research question, you might code the same response using different coding systems
- **Different coders using the same system will produce different codes for the same response**
  - Introduces a kind of "cluster structure" (by coder) into the data
  - Want to construct codes to minimize this (reduce ICC!)

Standard Classification Systems

- If the survey will be compared with other surveys, they should use the same coding scheme
- An ongoing longitudinal or panel survey like NCVS tries to use the same categories in each survey cycle or wave.
- Government or international agencies maintain standard coding systems for common types of information
  - Race/Ethnicity (US Census)
  - Standard Occupational Classification (US Dept of Labor)
  - North American Industry Classification System (US Economic Classification Policy Committee)

Weighting

- Many different sources of weighting in a survey, e.g.
  - Survey design weights
    - In stratified sampling, weights are used to combine stratum means & variance into overall means & variances
    - Other survey designs require weights to account for unequal sampling probabilities, etc.
    - Compute these weights *before the data is collected*
  - Nonresponse and post-stratification weights
    - Try to adjust sample proportions to equal population proportions
    - Compute these weights *after you see the data*
- These are discussed in detail in Groves; below I will only talk about post-stratification weights

Post-Stratification Weights

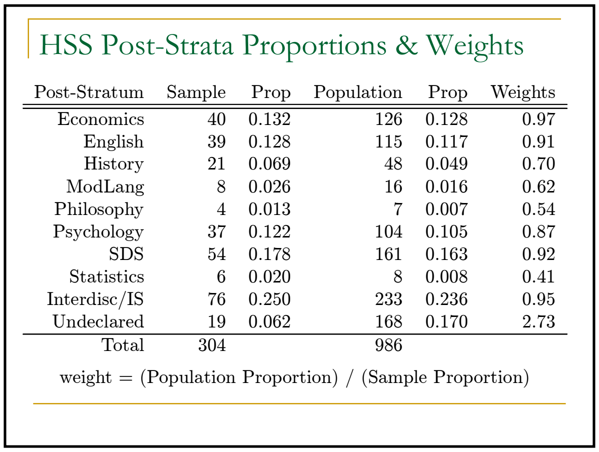
- As part of survey data collection it is a good idea to get general demographic information (e.g. in our surveys: sex, age, class, major, hometown, etc.)
- After data collection we compare the proportions in each of these categories in our sample with the same proportions in the population
- If they agree, great. If the disagree, we may re-weight the sample to make them agree
- These categories are called "post-strata", and the weights are called "post-stratification weights"

Post-Stratification Example

- The 2007 HSS advising satisfaction survey was a simple (no strata, no clusters) web survey of all 986 students in HSS.
- We can separate the responding students by major to see how representative the survey was of each department in HSS.
- If the representativeness was not the same in each department, and if we assume that the *nonresponse is ignorable* (??) within department, we can re-weight the sample data to get more accurate estimates of population quantitites.

HSS Response Rate in Dept Post-Strata

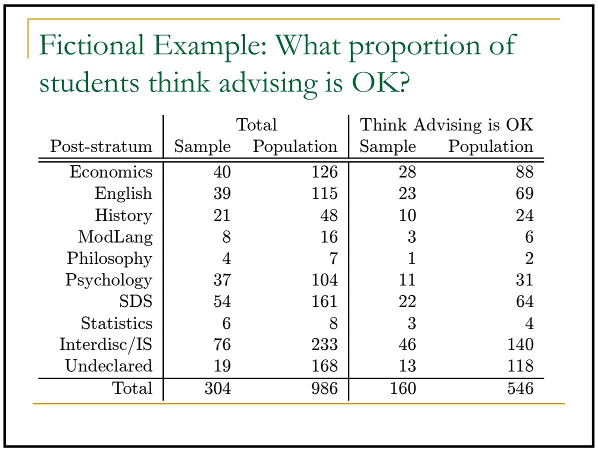
Post-Stratum	Sample	Population	Resp Rate
Economics	40	126	0.32
English	39	115	0.34
History	21	48	0.44
ModLang	8	16	0.50
Philosophy	4	7	0.57
Psychology	37	104	0.36
SDS	54	161	0.34
Statistics	6	8	0.75
Interdisc/IS	76	233	0.33
Undeclared	19	168	0.11
Total	304	986	0.31



HSS Post-Strata Proportions & Weights

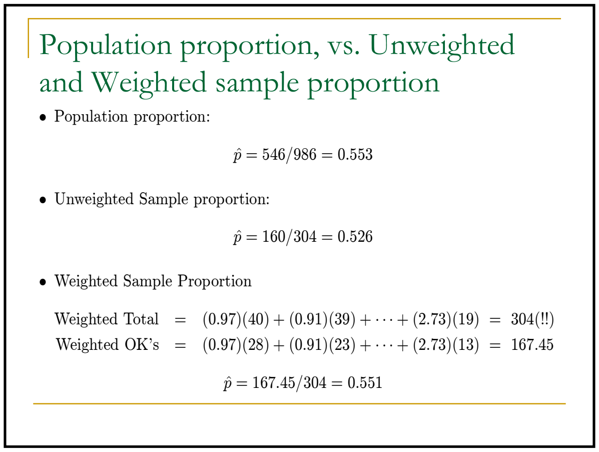
Post-Stratum	Sample	Prop	Population	Prop	Weights
Economics	40	0.132	126	0.128	0.97
English	39	0.128	115	0.117	0.91
History	21	0.069	48	0.049	0.70
ModLang	8	0.026	16	0.016	0.62
Philosophy	4	0.013	7	0.007	0.54
Psychology	37	0.122	104	0.105	0.87
SDS	54	0.178	161	0.163	0.92
Statistics	6	0.020	8	0.008	0.41
Interdisc/IS	76	0.250	233	0.236	0.95
Undeclared	19	0.062	168	0.170	2.73
Total	304		986		

weight = (Population Proportion) / (Sample Proportion)



Fictional Example: What proportion of students think advising is OK?

	Total		Think Advising is OK	
Post-stratum	Sample	Population	Sample	Population
Economics	40	126	28	88
English	39	115	23	69
History	21	48	10	24
ModLang	8	16	3	6
Philosophy	4	7	1	2
Psychology	37	104	11	31
SDS	54	161	22	64
Statistics	6	8	3	4
Interdisc/IS	76	233	46	140
Undeclared	19	168	13	118
Total	304	986	160	546



Population proportion, vs. Unweighted and Weighted sample proportion

- Population proportion:

$$\hat{p} = 546/986 = 0.553$$

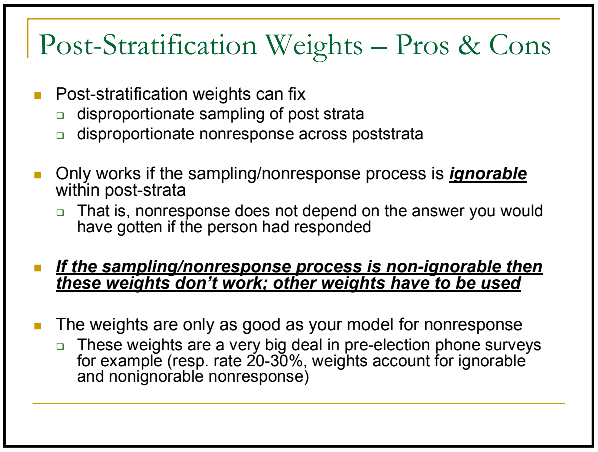
- Unweighted Sample proportion:

$$\hat{p} = 160/304 = 0.526$$

- Weighted Sample Proportion

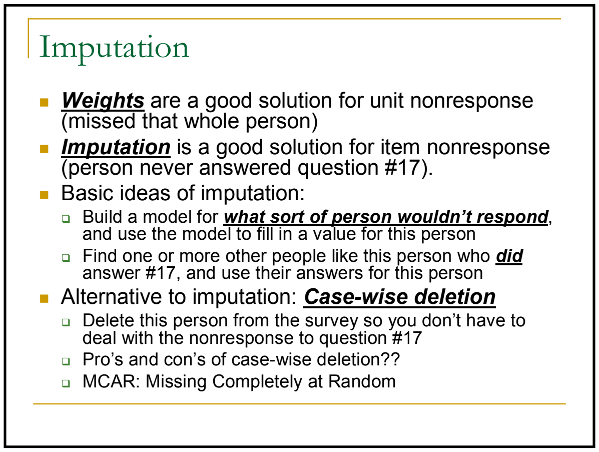
$$\begin{aligned} \text{Weighted Total} &= (0.97)(40) + (0.91)(39) + \cdots + (2.73)(19) &= 304(!!) \\ \text{Weighted OK's} &= (0.97)(28) + (0.91)(23) + \cdots + (2.73)(13) &= 167.45 \end{aligned}$$

$$\hat{p} = 167.45/304 = 0.551$$



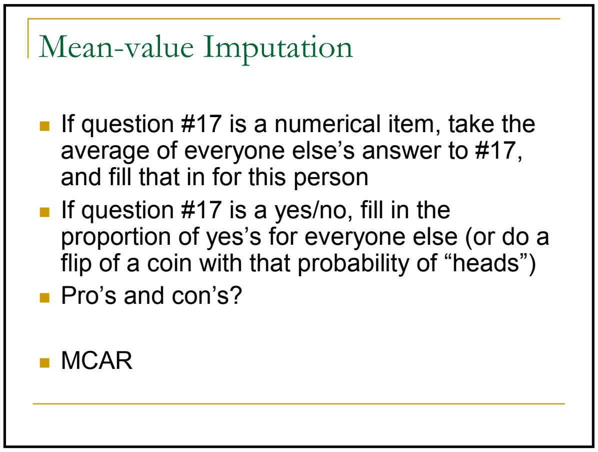
Post-Stratification Weights – Pros & Cons

- Post-stratification weights can fix
  - disproportionate sampling of post strata
  - disproportionate nonresponse across poststrata
- Only works if the sampling/nonresponse process is ***ignorable*** within post-strata
  - That is, nonresponse does not depend on the answer you would have gotten if the person had responded
- ***If the sampling/nonresponse process is non-ignorable then these weights don't work; other weights have to be used***
- The weights are only as good as your model for nonresponse
  - These weights are a very big deal in pre-election phone surveys for example (resp. rate 20-30%, weights account for ignorable and nonignorable nonresponse)



Imputation

- ***Weights*** are a good solution for unit nonresponse (missed that whole person)
- ***Imputation*** is a good solution for item nonresponse (person never answered question #17).
- Basic ideas of imputation:
  - Build a model for ***what sort of person wouldn't respond***, and use the model to fill in a value for this person
  - Find one or more other people like this person who ***did*** answer #17, and use their answers for this person
- Alternative to imputation: ***Case-wise deletion***
  - Delete this person from the survey so you don't have to deal with the nonresponse to question #17
  - Pro's and con's of case-wise deletion??
  - MCAR: Missing Completely at Random



Mean-value Imputation

- If question #17 is a numerical item, take the average of everyone else's answer to #17, and fill that in for this person
- If question #17 is a yes/no, fill in the proportion of yes's for everyone else (or do a flip of a coin with that probability of "heads")
- Pro's and con's?
- MCAR

Hot-Deck Imputation

- Among all the other people who answered question #17, find the one person who matches this person on important variables
  - age, sex, occupation, answers to other questions, etc. (whatever you think is important to understand this non-response!)
- Fill in that person's answer for this person's #17.
- Pro's? Cons?
- MAR: Missing at Random (within covariates)

Regression Imputation

- Among all the people who answered question #17, fit a regression model (or logistic regression, or whatever) for response to question #17 as a function of other variables:

  $y_{17} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{sex}) + \beta_3(\text{occupation}) + \beta_4(\text{answer to Q3}) + \ldots + \epsilon$
- Use the fitted model to predict what this person would have answered to #17, and fill that value in
- Pro's? Con's?
- MAR

Limitations of Imputation

- You have to have other variables in the survey that help you build a model for the nonresponse (MAR, or better MCAR)
- Nonignorable missingness (MNAR, missing not at random) is much harder
- After you have filled in the missing data
  - You have NOT increased the sample size; this will matter a lot if you are doing a lot of imputation
  - There is some uncertainty in what value to fill in; this can be accounted for by a technique called "***multiple imputation***"

Variance Calculations

- Final weights in a survey will be a combination (usually multiply together) of
  - Design Weights
  - Nonresponse Weights
  - Post-stratification Weights
- They each improve the point estimate ($\overline{y}$) but they have different effects on the variance ($Var(\overline{y})$) and standard error
- Additional variance adjustments are made for imputation
- There are not usually closed-form variance formulas (like our simple formulas for strata and clusters)
- More on this next time!

Review

- The entire survey process consists of
  - Collect/Code/Enter/Edit the data
  - Post-processing of the data
- ***Coding*** is the process of summarizing complex responses into numbers or categories – subject to bias (bad categories) and variability (coder variation)
- ***Post-processing*** includes (usually in this order)
  - Imputation
  - Weighting
  - Variance Calculation

These are calculations on the data to account for various weaknesses of the data.