36-303: Sampling, Surveys and Society

Review for Final Midterm, Apr 13

Brian W. Junker
132E Baker Hall
brian@stat.cmu.edu

Online Handouts

- These Notes
- Formula Sheet(s) for Final Midterm
  - Posted in Week12 on class website
- HW03: I handed these back Tuesday
  - Solutions posted on class website
- HW04: I am handing these back today
  - Solutions posted on class website

Review Session

- Review Session for Final Midterm Exam
  - Friday, 11:00am-12:30pm, PH A18C
  - Bring Questions! Better than me reading overheads to you…
  - Session will end early if we run out of things to talk about

Outline

- Review for Final Midterm Exam
  - Tues Apr 13, 2010
  - Closed book, closed notes
  - Formula sheets (old one plus new one) provided
  - Calculator recommended
  - Cumulative, but concentrating on
    - Groves Ch’s 4, 6, 10
    - Class notes, readings from Weeks 7,8,9,10; hw’s 3 & 4
  - This exam very similar in format to last one

Review

- Good sampling and data collection
- Nonresponse
- Stratified Sampling
- Cluster Sampling
- Post-Survey Processing
- Imputation
- Post-stratification
  - Weights
  - Variance Estimation (Taylor and Jackknife)

Good Sampling and Data Collection (1)

- Adjusting sample size for anticipated response rate
  - Email: 20% is typical – we've seen 12-50% with 25% common
  - Phone: E.g. 2007 Pew Religious Survey had 25%
  - Face to Face: Over 70%; we saw 73%
- Collect demographic variables so you can post-stratify (to check, and if necessary, reweight sample to be "representative")
- SRS from book/list; skip-sampling with random starts; etc.

Good Sampling and Data Collection (2)

- Contacting respondents
  - Once the sample (e.g. SRS) and mode of data collection is chosen (e.g. surveymonkey) is chosen, stick to it
    - You can break the SRS into "waves" and contact people in each wave separately; then if response rate is better than expected, later waves do not have to be contacted.
  - But you can try to contact respondents in any reasonable way: email, phone, Facebook, etc., to improve response rates
  - Followup with nonrespondents directly rather than send out general 2nd and 3rd notices to everyone in sample
  - Late responders can be thought of as being like never-responders.
  - Distinguish refusers vs procrastinators: **"No" means "no"!**
- Personal, polite contacts work best

Nonresponse (1)

- Types of non-response
  - Unit non-response
  - Item non-response
- Reasons
  - MCAR – missing completely at random (ignorable msgnss)
  - MAR – missing at random
  - MNAR – missing not at random (informative missingness)
- What to do about it
  - Ignore it (MCAR!)
  - Prevent it
  - Impute missing responses (MCAR, MAR; hard for MNAR!)

Nonresponse (2)

- Preventing Missingness
  - Survey content
  - Time of survey
  - Interviewer skills
  - Data collection method
  - Questionnaire design
  - Burden on respondent
  - Survey Introduction
  - Incentives
  - Followup
- Imputing missing responses
  - More below on post-processing survey data…

Stratified Sampling (1)

- H strata
  - N_h = population size in each stratum
  - n_h = sample size in each stratum
  - $f_h = n_h/N_h$ = sampling fraction, each stratum
  - $W_h = N_h/N$

$$N = \sum_{h=1}^{H} N_h$$

$$n = \sum_{h=1}^{H} n_h$$

- Mean

$$\overline{y}_{st} \;=\; \sum_{h=1}^{H} W_h \overline{y}_h \;\; unbiased\;\; estim.\;\; of \;\; \overline{y}_{pop} \;=\; \sum_{i=1}^{N} y_i \;=\; \sum_{h=1}^{H} W_h \overline{y}_{h,pop}$$

- Variance

$$Var(\overline{y}_{st}) \;=\; \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

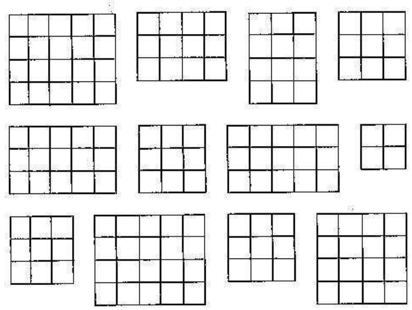
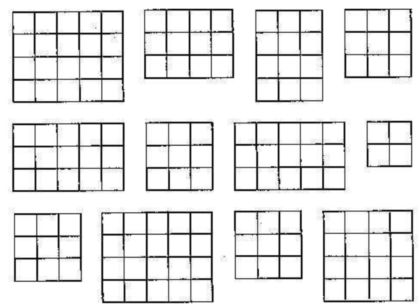
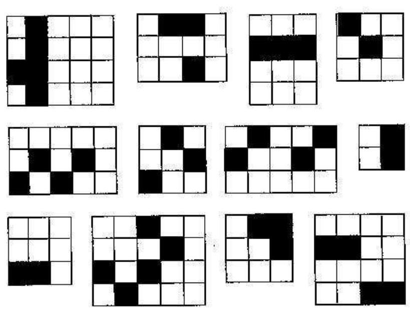
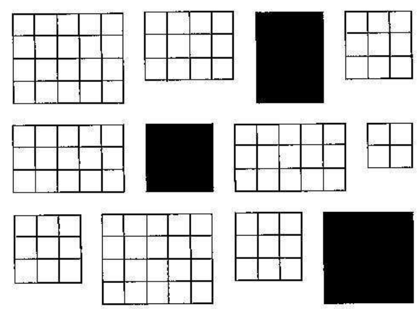
Stratified Sampling (2)

- The *design effect* is a measure of how much better or worse *Stratified* is than *one SRS:*

$$DEFF = \frac{Var(\overline{y}_{st})}{Var(\overline{y}_{srs})} = \frac{\sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}}{(1 - f) \frac{s^2}{n}}$$

- Usually, DEFF < 1, i.e. stratified does better than one big SRS!
  - Usually best if:
    - Elements are more similar to each other within strata than between (e.g., substantively meaningful strata)
    - Proportionate sampling *(f_h same in every stratum)*
  - Cochran (1961) suggests 2-6 strata usually give the best results; greater than 6 OK, but there are diminishing returns

Stratified vs. Cluster Sampling

Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum.	Each element of the population is in exactly one cluster.
Population of H strata; stratum h has n_h elements: 	One-stage cluster sampling; population of N clusters: 
Take an SRS from *every* stratum: 	Take an SRS of clusters; observe all elements within the clusters in the sample: 
Variance of the estimate of $\bar{y}_U$ depends on the variability of values *within* strata.	The cluster is the sampling unit; the more clusters we sample, the smaller the variance. The variance of the estimate of $\bar{y}_U$ depends primarily on the variability *between* cluster means.
For greatest precision, individual elements within each stratum should have similar values, but stratum means should differ from each other as much as possible.	For greatest precision, individual elements within each cluster should be heterogeneous, and cluster means should be similar to one another.

Cluster Sampling (1)

- One-stage clustering, equal cluster sizes:

For each cluster i in the SRS of clusters $\mathcal{S}$, we can calculate the cluster mean

$$\overline{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$$

where M is the cluster size. Since $\mathcal{S}$ is an SRS of n clusters

$$\overline{y}_{cl} = \frac{1}{n}\sum_{i\in\mathcal{S}} \overline{y}_i$$

The standard error (SE) needed for constructing confidence intervals is the *square root of*

$$Var(\overline{y}_{cl}) \approx \left(1-\frac{n}{N}\right)\frac{1}{n}s^2_{\overline{y}_i} = \left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i\in\mathcal{S}}(\overline{y}_i-\overline{y}_{cl})^2\right]$$

Cluster Sampling (2)

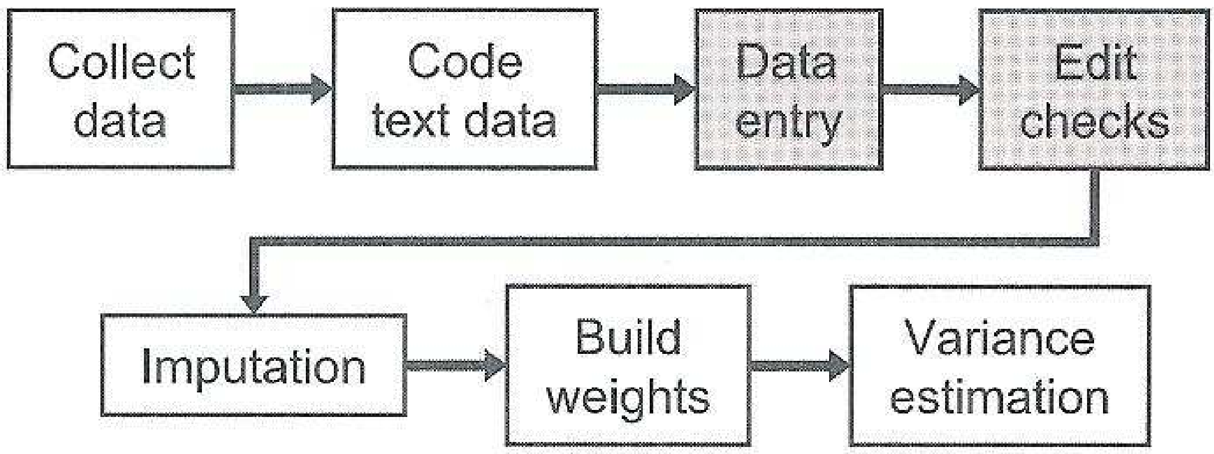
As with stratified sampling we can calculate a *design effect*

$$DEFF = \frac{Var(\bar{y}_{cl})}{Var(\bar{y}_{srs})} = \frac{M s^2_{\bar{y}_i}}{s^2_{y_{ij}}} \approx 1 + (M-1)\rho \ ,$$

where ρ is the *intraclass correlation (ICC)*, to see what the effect on precision of clustering is.

- In stratified sampling we usually get $DEFF < 1$ if we design the strata to have very different means and little variation within stratum.
- In clustered sampling, we usually get $DEFF > 1$. We can make $DEFF \approx 1$ by making the clusters have very similar means and lots of variation within cluster.

Post-survey Processing



- *<u>Top row:</u>* 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.
- *<u>Bottom row:</u>* Calculations based on the data and/or design

Imputation (1)

- ***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

Imputation (2)

- **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
  - MCAR: Missing completely at random
- **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.
  - Fill in that person's answer for this person's #17.
  - 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
  - MAR

Post-Stratification (1)

- 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 they disagree, we may re-weight the sample to make them agree

  weight = (population proportion)/(sample proportion)

- These categories are called “post-strata”, and the weights are called “post-stratification weights”

Post-Stratification (2)

- 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 as low as 5%, weights account for ignorable and nonignorable nonresponse)

Example from HW04

Sex	College	Hrs/Wk
M	Eng	28
M	Eng	29
M	Eng	23
M	Eng	35
M	Eng	29
M	Eng	30
M	Eng	34
M	Eng	31
F	Eng	30
F	Eng	31

Sex	College	Hrs/Wk
F	Eng	36
F	Eng	33
M	Lib	27
M	Lib	28
F	Lib	29
F	Lib	30
F	Lib	28
F	Lib	28
F	Lib	32
F	Lib	30

Sample Post-strata:

Sex	Eng	Lib
M	8	2
F	4	6

Population Post-strata:

Sex	Eng	Lib
M	617	380
F	450	551

Post-strat. weights:

$$(617/1998)/(8/20) = 0.77 \qquad (380/1998)/(2/20) = 1.90$$
$$(450/1998)/(4/20) = 1.13 \qquad (551/1998)/(6/20) = 0.92$$

Example from HW04 (cont'd)

Sex	College	Hrs/Wk	Wgt
M	Eng	28	0.77
M	Eng	29	0.77
M	Eng	23	0.77
M	Eng	35	0.77
M	Eng	29	0.77
M	Eng	30	0.77
M	Eng	34	0.77
M	Eng	31	0.77
F	Eng	30	1.13
F	Eng	31	1.13

Sex	College	Hrs/Wk	Wgt
F	Eng	36	1.13
F	Eng	33	1.13
M	Lib	27	1.90
M	Lib	28	1.90
F	Lib	29	0.92
F	Lib	30	0.92
F	Lib	28	0.92
F	Lib	28	0.92
F	Lib	32	0.92
F	Lib	30	0.92

Unweighted mean:

$$\overline{y}_{srs} = \frac{1}{20}\sum_{i=1}^{20} y_i = 30.05$$

Weighted mean:

$$\overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i} = 29.91$$

Example from HW04 (cont'd)

- Taylor Series Variance Approximation

$$Var_{TS}(\overline{y}_w) \approx 0.46 =$$

$$\frac{1}{(\sum_i w_i)^2}\left[Var(\sum_i w_i y_i) - 2\overline{y}_w Cov(\sum_i w_i y_i, \sum_i w_i) + (\overline{y}_w)^2 Var(\sum_i w_i)\right]$$

where $\overline{y}_w = 29.91$, $\overline{w} = 1.00$, $\overline{wy} = 29.91$ and

$$Var(\sum_{i=1}^{n} w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^{n}(w_i - \overline{w})^2 = 2.26$$

$$Var(\sum_{i=1}^{n} y_i w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^{n}(w_i y_i - \overline{wy})^2 = 1788.84$$

$$Cov(\sum_{i=1}^{n} y_i w_i, \sum_{i=1}^{n} w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^{n}(w_i y_i - \overline{wy})(w_i - \overline{w}) = 60.64$$

Example from HW04 (cont'd)

- Jackknife Variance Approximation:
  - Replicate $\overline{y}_w^{(r)} = \dfrac{\sum_{i=1}^n w_i^{(r)} y_i^{(r)}}{\sum_{i=1}^n w_i^{(r)}}$'s:

```
29.99382 29.94970 30.21439 29.68501 29.94970 29.90558 29.72912
29.86147 30.09879 30.02371 29.64834 29.87356 30.00619 29.81600
29.93868 29.88352 29.99383 29.99383 29.77321 29.88352
```

  - Calculate

$$\overline{y}_{JK} = \frac{1}{n}\sum_{r=1}^{n} \overline{y}_w^{(r)} = 29.91$$

$$Var_{JK}(\overline{y}_w) = \frac{n-1}{n}\sum_{r=1}^{n} (\overline{y}_w^{(r)} - \overline{y}_{jk})^2 = 0.34$$

Example from HW04 (cont'd)

- Now confidence intervals can be calculated in the usual way, e.g.

$$\left(\overline{y}_w - 2\sqrt{(1-n/N)Var(\overline{y}_w)}\ ,\ \ \overline{y}_w + 2\sqrt{(1-n/N)Var(\overline{y}_w)}\right)$$

for either the Taylor Series or Jackknife estimate of variance.

Review

- Review Session for Final Midterm Exam
  - Friday, 11:00am-12:30pm, PH A18C
  - Bring Questions! Better than me reading overheads to you…
  - Session will end early if we run out of things to talk about
- Final Midterm Exam
  - Tues Apr 13, 2010, in class
  - Closed book, closed notes
  - Formula sheets (old one plus new one) provided
  - Calculator recommended
  - Cumulative, but concentrating on
    - Groves Ch's 4, 6, 10
    - Class notes, readings from Weeks 7,8,9,10; hw's 3 & 4