36-303: Sampling, Surveys and Society

Stratified Samples and Sample Size Calculations

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

Handouts

- These Lecture Notes
- U.S.Religious Landscape Survey 2008:
  - Washington Post Article
  - Survey Methodology Appendix from Full Report
- Handout on Stratified Sampling – Will be on Web!
- Handout on Sampling Details – Will be on Web!
  - Selecting an SRS from C-Book
  - Contacting respondents
  - Nonresponse followup on surveymonkey.com
- Reading:
  - Stratified Sampling: Groves Sect 4.5,
  - Nonresponse: Groves Ch 6

Outline

- Team Projects This Week
- US Religious Landscape Survey: A Phone Survey
- Midterm Exam Progress Report
- Stratification
  - What is it; Notation
  - Weights and Proportionate Sampling
  - Variances and Design Effect
  - Examples

Team Projects This Week

- Team Working Agreements Due Today (email)
- I.5a Due Thursday (email)
  - Include a paragraph or so on your research question
  - Decide on a sampling scheme (e.g., SRS, Stratified random sample, etc.) and explain why you chose it.
  - Write a questionnaire with 20-30 questions. Some of you have already started this process. Pretend I haven't seen any of your previous attempts.
    - 10 or so demographic questions
    - 10-20 substantive questions
  - Give some idea of the sample size you will require and how you arrived at this number (talk about the margin of error for inferences you want to make).
    - Compromise between sample size calculation, and how big a sample you can afford to collect and process!

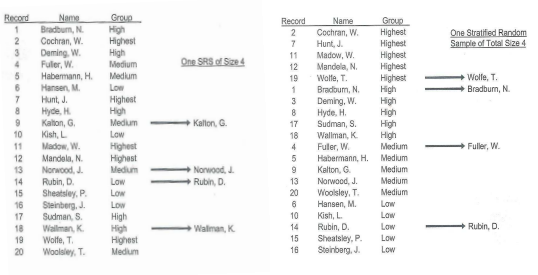
Midterm Exam Progress Report

- Two makeup exams are still being graded
- I have not had a chance to look at any of the graded exams yet, and I want to do that before handing the exams back.

Stratified Sampling

- Strata are just subgroups of the target population that have some feature in common (gender, major, region, income, …)
- Why stratify?
  - We need to make a separate inference for each stratum (e.g. we want to estimate men's and women's incomes separately)
  - Different sampling schemes would be used in each stratum (PA voters in PA, vs PA voters in Iraq)
  - Population is geographically diverse (Minnesota, Illinois, Ohio, Pennsylvania)
  - Reduce variance of estimates (and reduce *sample size*) by exploiting similarity among members of the same stratum

What is Stratification?



Unstratified Sample

Stratified Sample

Some Basic Notation

- H strata
  - N_h = population size in each stratum $N = \sum_{h=1}^{H} N_h$
  - n_h = sample size in each stratum $n = \sum_{h=1}^{H} n_h$
  - $f_h = n_h/N_h$ = sampling fraction, each stratum
- The population average

$$\bar{y}_{pop} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{H}\frac{N_h}{N}\frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_{h,pop}$$

- In stratified sampling we mimic this

$$\bar{y}_{st} = \frac{1}{n}\sum_{i=1}^{n} y_i = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h \text{ where } \bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$$

Weights, and Proportionate Sampling

- Let $W_h = N_h/N$. Then

$$\bar{y}_{pop} = \sum_{h=1}^{H} W_h \bar{y}_{h,pop} \text{ and } \bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$$

- In proportionate sampling we let $f_h = n_h/N_h = f$ for all strata h. Then $n_h/n = N_h/N$ (why??)
  - The sample is called "self-weighting"
  - Sample mean is "simple" for self-weighting

$$\bar{y}_{srs} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} y_{hi} = \sum_{h=1}^{H}\frac{n_h}{n}\frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$$
$$= \sum_{h=1}^{H}\frac{n_h}{n}\bar{y}_h = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h = \sum_{h=1}^{N} W_h \bar{y}_h = \bar{y}_{st}$$

Sampling Variances (SRS w/o replacement in each stratum)

- Within each stratum it's the same old answer

$$Var(\bar{y}_h) = (1 - f_h)\frac{s_h^2}{n_h} \text{ where } s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$$

- Then we combine across strata using weights $(W_h)^2$:

$$Var(\bar{y}_{st}) = Var\left(\sum_{h=1}^{H} W_h \bar{y}_h\right)$$
$$= \sum_{h=1}^{H} Var(W_h \bar{y}_h) = \sum_{h=1}^{H} W_h^2 Var(\bar{y}_h)$$
$$= \sum_{h=1}^{H} W_h^2 (1 - f_h)\frac{s_h^2}{n_h}$$

Design Effect

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

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

- Usually, $d^2 < 1$, 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
  - Cochran (1961) suggests 2-6 strata usually give the best results; greater than 6 OK, but there are diminishing returns

Handout on Stratified Sampling

(Briefly) Handout on Sampling Details

Review

- Team Projects This Week
- US Religious Landscape Survey: A Phone Survey
- Midterm Exam Feedback
- Stratification
  - What is it; Notation
  - Weights and Proportionate Sampling
  - Variances and Design Effect
  - Handout on Stratified Sampling
- Handout on Sampling Details – Will be on Web!
- Reading:
  - Stratified Sampling: Groves Sect 4.5,
  - Nonresponse: Groves Ch 6