Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 4a: Weighting and poststratification

Stratum weights and unit weights

- A survey of American adults is conducted, and the respondents are 600 women and 440 men. What unit-level weights should be assigned to the men and the women? (Assume here that you are weighting only on sex, not on any other variable.)
- A survey is done of students at a university. In the resulting weighting, each undergraduate is given a weight of 1 and each graduate student is given a weight of 1.8. The sample includes 200 undergraduates and 300 graduate students. Of these, 30% of the undergraduates and 50% of the graduate students respond Yes to a particular question of interest. Give an estimate and standard error for the proportion of all the students in the population who would answer Yes to this question if asked.

Reminder: weighted averages

- $y_{\text{w.avg}} = 0.2y_1 + 0.3y_2 + 0.5y_3$
- $\text{sd}(y_{\text{w.avg}}) = \sqrt{0.2^2\text{se}_1^2 + 0.3^2\text{se}_2^2 + 0.5^2\text{se}_3^2}$
- Example: survey with 3 strata:
  - In stratum 1, 75% Yes responses out of 200 respondents, se $\sqrt{.75 \cdot .25/200}$
  - In stratum 2, 80% Yes out of 300, se $\sqrt{.80 \cdot .20/300}$
  - In stratum 3, 90% Yes out of 400, se $\sqrt{.90 \cdot .10/400}$
  - Weighted average: $0.2 \cdot 0.75 + 0.30 \cdot 0.80 + 0.5 \cdot 0.90 = 0.84$
  - Standard error: $\sqrt{0.2^2\text{se}_1^2 + 0.3^2\text{se}_2^2 + 0.5^2\text{se}_3^2}$
- Next: example with numerical responses
  - In stratum 1, avg 2.5, sd 0.9, out of 200 respondents, se $0.9/\sqrt{200}$
  - In stratum 2, avg 3.0, sd 0.9, out of 300, se $0.9/\sqrt{300}$
  - In stratum 3, avg 4.0, sd 1.3, out of 400, se $1.3/\sqrt{400}$
  - Weighted average, standard error using same formulas as before

Adjusting exit polls and election-night results

- Exit polls get too many Democrats
- Weight based on election outcome
- "With 11% of precincts reporting, Obama has 44% of the vote in Pennsylvania"

Unit-level weighting in R

- Direct calculations:
  - Unweighted:

    ```
    mean (pew$rvote)
    ```
  - Weighted:

    ```
    sum (pew$pop.weight*pew$rvote)/sum(pew$pop.weight)
    ```
- Using the "survey" package:
  - Unweighted:

    ```
    srs_design <- svydesign (id=~1, data=pew)
    svymean (~rvote, design=srs_design)
    ```
  - Weighted:

    ```
    weighted_design <- svydesign (id=~1,
      weights=~pop.weight, data=pew)
    svymean (~rvote, design=weighted_design)
    ```

Logistic regression using the "survey" package in R

- Unweighted:

```
M1 <- svyglm (rvote ~ factor (eth) + factor (inc) +
  factor (edu) + male + married,
  family=binomial(link="logit"), design=srs_design)
```

- Weighted:

```
M2 <- svyglm (rvote ~ factor (eth) + factor (inc) +
  factor (edu) + male + married, family=quasibinomial
  (link="logit"), design=weighted_design)
```

Survey weighting is a mess

- Using weights
  - Weighted mean: $\bar{y}_w = \sum_{i=1}^n w_i y_i \Big/ \sum_{i=1}^n w_i$
  - Estimating a ratio: $r_w = \sum_{i=1}^n w_i y_i \Big/ \sum_{i=1}^n w_i x_i$
  - Estimating anything more complicated: ???
- Regression modeling as an alternative
  - Need to control for many potential confounders
  - Hierarchical modeling as a (potential) solution

Where do weights come from?

- Survey weights are not inverse probabilities of selection
- Two simple stories
- CBS/New York Times pre-election polls

Story 1: weights for men and women

- Survey of a population with 52% women, 48% men
- Simple random sampling, $n = 100$
  - SRS 1: 52 women, 48 men. Weights are $w_i = 1$ for everyone
  - SRS 2: 60 women, 40 men. Weights are $w_i = \frac{52}{60}$ for women, $\frac{48}{40}$ for men
- We know the population proportions, so the selection probabilities are irrelevant
- Weights depend on the entire survey; the (y_i, w_i) paradigm is inappropriate

Example: CBS/New York Times pre-election polls

id	org	y	state	edu	age	adults	weight
6140	cbsnyt	NA	7	3	1	2	923
6141	cbsnyt	1	39	4	2	2	558
6142	cbsnyt	0	31	2	4	1	448
6143	cbsnyt	0	7	3	1	2	923
6144	cbsnyt	1	33	2	2	1	403

- The weight is listed as just another survey variable
- But they are actually constructed *after* the survey
- Weights $w_i = g(X_i, \theta)$

Story 2: weights for household size

- Telephone survey of households
  - Interview one adult in each sampled household
  - Pr(selection) $\propto 1/(\#$ adults in household)
  - Obvious survey weight: # adults in household
- But . . . we can do better (Gelman and Little, 1998):
  - Instead of weights 1, 2, 3, 4, set weights to 1.0, 1.4, 1.7, 2.0
  - Lower bias *and* lower variance
  - Set weights by matching to census numbers: sampling probabilities don't matter at all!

The poststratification framework

- Goal is to estimate population average, θ
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Point estimate $\hat{\theta} = \frac{\sum_j N_j \hat{\theta}_j}{\sum_j N_j}$
- Cells j might be determined by sex, age, education, ethnicity, ...
- Some estimators:
  - Simple poststratification: $\hat{\theta}_j = \bar{y}_j$
  - Sample mean: $\hat{\theta}_j = \bar{y}$
  - Bayesian compromises: model θ_j given covariates X_j

Complications

- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Many cells j ($2 \times 4 \times 5 \times 4 \times 50$): need complicated model with many levels of interactions
- Adjusting for non-census variables (for example, religion): need to model the N_j's
- Regression of y on x
  - Must model $y|x$ within each cell j
  - Then average over cells to estimate E(y) as a function of x