36-303: Sampling, Surveys and Society

Variance Calculations for Weights
Brian W. Junker
132E Baker Hall
brian@stat.cmu.edu

Handouts

- These Lecture Notes
  - Also an R handout!
- Post-Stratification Challenges Handout
- Handouts on the Census [read first!]
  - Farley – Statistical, Political and Constitutional Issues in the Census Undercount
  - USAToday: Q&A: 2010 Census
  - Fort Worth Star Telegram: Census Undercount Could Cost Texas Money, Political Clout

Outline/Announcements

- Announcements
  - Groves will be here for most/all of class Thurs
  - Some additional guests attending class also
- A New HW will be posted online tomorrow [Weds]
  - Summary of Groves' visit, including something you learned, and something you still don't understand
  - Problem on cluster sampling
  - Problem on post-stratification weights
- Today: Variance Calculations for Weights
  - Taylor Series
  - Random Partition
  - Jackknife

Variance Calculations for Weights

- Most survey sample estimates have a ratio form: $$\overline{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$$
- Two approaches to $Var(\overline{y}_w)$:
  - Use a ***one-term Taylor approximation*** to "linearize" the survey estimate, and apply CLT.
  - Use a ***replication scheme*** to create "replicate samples" by resampling the real sample and look at the variability among the replicates.
    - Non-overlapping replicates: E.g., *Random Partitions*
    - Overlapping replicates: E.g., *Jackknife Method*

Taylor Series Approximation (Bkgd)

- The ***Delta Method***
  - We know that if
    $$\hat{\theta} - \theta \sim N(0, \sigma^2/n)$$
    then
    $$a(\hat{\theta} - \theta) \sim N(0, a^2\sigma^2/n)$$
  - We can extend this to a nonlinear function
    $$f(\hat{\theta}) - f(\theta) = f'(\theta)(\hat{\theta} - \theta) + (remainder)$$
    so that
    $$f(\hat{\theta}) - f(\theta) \approx f'(\theta)(\hat{\theta} - \theta) \sim N(0, [f'(\theta)]^2\sigma^2/n)$$

Taylor Series Approximation (Bkgd)

- Univariate Delta Method
  $$\text{If} \quad \hat{\theta} - \theta \sim N(0, \sigma^2/n)$$
  $$\text{then } f(\hat{\theta}) - f(\theta) \sim N(0, [f'(\theta)]^2\sigma^2/n)$$
- Multivariate Delta Method
  $$\text{If } \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} - \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{n}\sum \right)$$
  then
  $$f\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} - f\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}$$
  $$\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{n}\left(\frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_2}\right) \sum \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \end{pmatrix} \right)$$

Taylor Series for Ratio Estimator

- Now we consider

$$\overline{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} = \frac{\hat{\theta}_1}{\hat{\theta}_2} = f(\hat{\theta}_1, \hat{\theta}_2)$$

- The gradient of f has components

$$\frac{\partial f}{\partial \theta_1} = 1/\theta_2 \ , \quad \frac{\partial f}{\partial \theta_2} = -\theta_1/\theta_2^2$$

- The Variance/Covariance Matrix for (θ_1, θ_2) is

$$\sum = \left[\begin{array}{cc} Var(\sum_i w_i y_i) & Cov(\sum_i w_i y_i, \sum_i w_i) \\ Cov(\sum_i w_i y_i, \sum_i w_i) & Var(\sum_i w_i) \end{array} \right]$$

Taylor Series Variance for Ratio Estimator

- Applying the Multivariate Delta Method we get

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

- Need to calculate the variances and covariance above – see next slide…

Calculating the Variances for TS Method...

If we assume that each pair $(w_i y_i, w_i)$ is independent of every other pair (not quite true but close!) then

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

where $\overline{w} = \frac{1}{n} \sum_i w_i$. Similarly,

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

where $\overline{wy} = \frac{1}{n} \sum_i w_i y_i$, and

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

Example: HSS Advising Survey…

Post-Strat.	Adv'ing OK	Samp Total	Prop	Pop Total	Prop	Weights
Economics	28	40	0.132	126	0.128	0.97
English	23	39	0.128	115	0.117	0.91
History	10	21	0.069	48	0.049	0.70
ModLang	3	8	0.026	16	0.016	0.62
Philosophy	1	4	0.013	7	0.007	0.54
Psychology	11	37	0.122	104	0.105	0.87
SDS	22	54	0.178	161	0.163	0.92
Statistics	3	6	0.020	8	0.008	0.41
Interdisc/IS	46	76	0.250	233	0.236	0.95
Undeclared	13	19	0.062	168	0.170	2.73
Total	160	304		986		

weight = (Population Proportion) / (Sample Proportion)

TS Variance Estimate, HSS Advising Data…

$$\begin{aligned}
y_i &= 1 \text{ (yes) or } 0 \text{ (no)} \\
\overline{y}_w &= 0.5507865 \\
\overline{w} &= 1.001678 \\
\overline{wy} &= 0.5517105 \\
Var(\sum_i w_i) &= n \cdot s_w^2 = (304)(0.2124) = 64.57 \\
Var(\sum_i w_i y_i) &= n \cdot s_{wy}^2 = (304)(0.4127) = 125.47 \\
Cov(\sum_i w_i y_i, \sum_i w_i) &= n \cdot s_{wy,w} = (304)(0.1637) = 49.75
\end{aligned}$$

So

$$Var_{TS}(\overline{y}_w) = (125.47 - 2(0.5507)(47.75) + (0.5507)^2 * (64.57)/(304 \cdot 1.0017)^2 = 0.000973$$

This is larger (typical!) than the naive variance based on $\hat{p} = \overline{y}$:

$$\hat{p}(1-\hat{p})/n = (0.53)(1-0.53)/(304) = 0.000819$$

We should also multiply by the fpc $= 1 - 304/986 = 0.69$!

Replication Scheme: Jackknife

- From the original sample we create r=1, 2, … n *Jackknife samples* (of size n-1), by deleting one observation at a time from the original data.
- From each jackknife sample
  - Recalculate the weights
  - Recalculate

$$\overline{y}_w^{(r)} = \frac{\sum_{i=1}^n w_i^{(r)} y_i^{(r)}}{\sum_{i=1}^n w_i^{(r)}}$$

- Now calculate

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

Example: HSS Advising Data (Again)

Post-Strat.	Adv'ing OK	Samp Total	Prop	Pop Total	Prop	Weights
Economics	28	40	0.132	126	0.128	0.97
English	23	39	0.128	115	0.117	0.91
History	10	21	0.069	48	0.049	0.70
ModLang	3	8	0.026	16	0.016	0.62
Philosophy	1	4	0.013	7	0.007	0.54
Psychology	11	37	0.122	104	0.105	0.87
SDS	22	54	0.178	161	0.163	0.92
Statistics	3	6	0.020	8	0.008	0.41
Interdisc/IS	46	76	0.250	233	0.236	0.95
Undeclared	13	19	0.062	168	0.170	2.73
Total	160	304		986		

weight = (Population Proportion) / (Sample Proportion)

JK Variance Estimate, HSS Advising Data…

- There are 304 Jackknife samples, of size 303 each.
  - 28 jackknife samples omit one of the Econ 'yes' obs's
  - 12 jackknife samples omit one of the Econ 'no' obs's
  - 23 jackknife samples omit one of the English 'yes' obs's
  - 16 jackknife samples omit one of the English 'no' obs's
  - etc., etc. for the other 8 post-strata
- Calculate $\overline{y}_w^{(r)}$'s
  - The first few unique $\overline{y}_w^{(r)}$ are

    0.5478, 0.5488, 0.5490, 0.5493, 0.5495 …

  - (there are many duplicates!)

JK Variance Estimate, Continued

- Now we calculate

$$\overline{y}_{JK} = \frac{1}{304}\sum_{r=1}^{304}\overline{y}_w^{(r)} = 0.5508 \ (\ = \ \overline{y}_w \ )$$

and

$$Var_{JK}(\overline{y}_w) = \frac{304-1}{304}\sum_{r=1}^{304}(\overline{y}_w^{(r)} - \overline{y}_{JK})^2 = 0.000963$$

- Very similar to TS Variance estimate:

$$Var_{TS}(\overline{y}_w) = 0.000973$$

Actual Calculations…

- See R handout… (is there someone in every group that knows a little R?)
- My recommendation:
  - If you know the formula, ***<u>Taylor Series</u>*** approx is really easy to carry out. However, for a new statistic, have to re-apply Delta Method.
  - ***<u>Jackknife</u>*** is harder to set up, but once it's done, it works for ***<u>all</u>*** possible statistics, not just weighted averages
  - As sample size grows, TS and JK produce same answers
  - (again, we should multiply by fpc = (1-(samp)/(pop)))

Making a Confidence Interval

- Approx 95% confidence interval, based on the Jackknife standard error:

$$(0.5508 - 2 * \sqrt{(1-304/986)(0.000963)} \quad , \quad 0.5508 + 2 * \sqrt{(1-304/986)(0.000963)})$$

$$(0.4992 \quad , \quad 0.6024)$$

- In our fictional example we know the true population proportion:

$$p_{pop} = 546/986 = 0.553$$

- We capture the true mean in this case

Review

- Groves
  - Will be here for most/all of class Thurs
  - Some additional guests attending class also
- A New HW will be posted online tomorrow [Weds]
  - Groves, Cluster Sampling, Post-Stratification Weights
- Fienberg giving the lecture next Tuesday
- Today: Variance Calculations for Weights
  - Taylor Series
  - Random Partition
  - Jackknife