36-303: Sampling Surveys and Society
Post-Stratification Challenges

Brian Junker
132E Baker Hall
brian@stat.cmu.edu
March 23, 2010

Contents

1 Post-Stratification

1.1 Basic Post-Stratification

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

$$\textit{weight} = (\textit{population proportion})/(\textit{sample proportion})$$

These categories are called "post-strata", and the weights are called "post-stratification weights".

1.2 Using the Weights

To illustrate, we consider an SRS of 100 college students at a fictional university, taken to find out how many hours they work on class-related things. The survey was done by email/website, and unfortunately the response rate was only 20%, so we have only 20 observations in our sample.

Because so few people responded, we are worried about selection effects in the sample: how are the 20 students who responded different from the 80 people who did not, and how will this affect our estimates? A way to try to make the sample look more like the population is to calculate post-stratification weights and use the weights in all of our calculations.

The basic data is as follows:

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

We have collected some demographic data (male/female and engineering/liberal-arts) on the students, and we find that in the sample we have the following proportions in post-strata created by crossing these categories:

Sex	Eng	Lib	
M	8	2	
F	4	6	
Total			20

On the other hand, in the population of all studetns at this university the counts are

Sex	Eng	Lib	
M	617	380	
F	450	551	
Total			1998

To adjust the sample proportions so that they better reflect the population proportions of each of these four categories, we create post-stratification weights according to the formula:

$$\textit{weight} = (\textit{population proportion})/(\textit{sample proportion})$$

In particular the weights are: Post-strat. weights:

$$w_{M,E} = (617/1998)/(8/20) = 0.77 \qquad w_{M,L} = (380/1998)/(2/20) = 1.90$$
$$w_{F,E} = (450/1998)/(4/20) = 1.13 \qquad w_{F,L} = (551/1998)/(6/20) = 0.92$$

We attach these weights to each of the 20 observations in the original SRS, obtaining

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

Now we can compare the unweighted mean from the SRS,

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

with the weighted mean using the post-stratification weights,

$$\overline{y}_{w} = \frac{\sum_i w_i y_i}{\sum_i w_i} = 29.91\ .$$

There isn't much difference in this case, but you can see that there could be a big difference if the sample proportions are far different from the population ones.

We also need to adjust our variance calculation to account for the weights. There are two commonly-used methods:

Taylor Series Method This method is based on a Taylor Series approximation (known as the "delta method" to statisticians) that was discussed in class. We calculate

$$\begin{aligned} \mathrm{Var}_{TS}(\overline{y}_w) &\approx \frac{1}{(\sum_i w_i)^2}\left[\mathrm{Var}(\sum_i w_i y_i) - 2\overline{y}_w Cov(\sum_i w_i y_i, \sum_i w_i) + (\overline{y}_w)^2 Var(\sum_i w_i)\right] \\ &= 0.46 \end{aligned}$$

where $\overline{y}_w = 29.91$ as above, $\overline{w} = \frac{1}{n}\sum_i w_i = 1.00$, $\overline{wy} = \frac{1}{n}\sum_i w_i y_i = 29.91$, and [assuming the pairs (w_i, y_i) are independent for different i's],

$$\begin{aligned} \mathrm{Var}(\sum_{i=1}^n w_i) &\approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i - \overline{w})^2 = 2.26 \\ \mathrm{Var}(\sum_{i=1}^n y_i w_i) &\approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})^2 = 1788.84 \\ \mathrm{Cov}(\sum_{i=1}^n y_i w_i, \sum_{i=1}^n w_i) &\approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})(w_i - \overline{w}) = 60.64 \end{aligned}$$

Jackknife Method In this method we create n replicate data sets of $n-1$ observations each, by deleting each one of the original observations in turn. We recalculate the post-stratification weights for each replicate sample, and recalculate

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

for the r^{th} replicate sample (based on deleting the r^{th} observation from the original data set). In the case of our fictional college work survey we get the following $y_w^{(r)}$'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
```

These can be combined to create a new estimate of the average hours of work

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

(which in most cases should be equal to the original $\overline{y}_w$, as it is here), and a new estimates of the variance

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

(Note that we multiply the sum of squared differences by $(n-1)/n$ because the $y_w^{(r)}$'s are all positively correlated (they are calculated from mostly the same data), rather than the usual $1/(n-1)$ which would be appropriate if the $y_w^{(r)}$'s were independent of each other).

In most cases, $\text{Var}_{TS}(\overline{y}_w)$ and $\text{Var}_{JK}(\overline{y}_w)$ will be slightly larger than $\text{Var}(\overline{y}_{srs})$, reflecting extra uncertainty in calculating the weights.

Once we have the variance, we can calculate confidence intervals, etc., as usual, for example an approximate 95% interval would be:

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

For example, plugging in $\overline{y}_w = 29.91$, $\text{Var}_{TS}(\overline{y}_w) = 0.46$, $n = 20$, and $N = 1998$, we obtain the interval

$$(28.56\,,\;\; 31.26)$$

for the overall average number of hours worked per week by students at this university.

1.3 Pro's & Con's

Post-stratification weights can fix

- Disproportionate sampling of post strata;
- Disproportionate nonresponse across poststrata

However, the weights we are calculating here, which simply adjust the sample proportions so that they equal the population proportions, only work 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. These new weights would be based on a model that you build, to explain why the non-responders are not responding. Weights for non-ignorable non-response 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).

2 What to Do About Empty Strata

As in the example above, you can "cross" several different demographic variables to make post-strata. However, you will quickly run out of sample observations to fill each post-stratum, if you try to combine very many variables, or variables with many categories, this way.

For example, if you have collected data on CMU students for

- Class: Freshman, Sophomore, Junior, Senior, 5^{th}-year
- School: HSS, CIT, MCS, CFA, Tepper, Heinz, CS

then you will have $5 \times 7 = 35$ post-strata. If your sample size is only around 100 students, you will very likely have some empty post-strata in the sample, even though there are students in every post-stratum in the population.

Here are two strategies to try, in this case:

Use fewer variables to post-stratify.

For example, maybe the sample proportions of class (Fr/So/Jr/etc) are a worse match to the population proportions than for school. Then try to post-sratify just on class. In many cases, this will also help correct proportions for school.

Build weights for each variable in succession.

For example,

- First build post-stratification weights for *class*, using the sample counts and population counts for class.
- Then build post-stratification weights for *school*, using the sample counts and population counts for school.

Your final weights will be the product of these two sets of weights.

Using the college work survey example above, we find

w_E = [(617+450)/1998]/[(8+4)/20] = 0.8901 , w_L = [(380+551)/1998]/[(2+6)/20] = 1.1649

and

w_M = [(617+380)/1998]/[(8+2)/20] = 0.9980 , w_F = [(450+551)/1998]/[(4+6)/20] = 1.0020

so the combined weights would be

$$w^*_{M,E} = (0.9980)(0.8901) = 0.8883 \qquad w^*_{M,L} = (0.9980)(1.1649) = 1.1647$$
$$w^*_{F,E} = (1.0020)(0.8901) = 0.8919 \qquad w^*_{F,L} = (1.0020)(1.1649) = 1.1672$$

These weights differ a bit from the original weights we calculated

$$w_{M,E} = (617/1998)/(8/20) = 0.77 \qquad w_{M,L} = (380/1998)/(2/20) = 1.90$$
$$w_{F,E} = (450/1998)/(4/20) = 1.13 \qquad w_{F,L} = (551/1998)/(6/20) = 0.92$$

because the w^*'s assume that sex and college are independent in the sample and population, whereas the w's do not assume independence.

It is probably better to stick with the first strategy if you can. Even professional surveys, such as the 2007 Pew survey on Religion in America (that we looked at briefly in class a few weeks ago), uses this method.

3 How to Post-Stratify Within Sampling Strata

3.1 Sampling (Design) Strata

Sampling strata (also called design strata or pre-strata) are strata that you create before you do the survey. For example, perhaps you want to stratify on school within CMU, so you take a separate SRS within each of CMU's seven schools. The schools are *sampling strata*.

3.2 Checking Post-Stratum Proportions Within Sampling Strata

If you are worried about nonresponse within sampling strata, you can also post-stratify within sampling strata. For example, you could post-stratify on class (Fr/So/Jr/etc.) within each sampling stratum.

You can then check to see if the sampling proportions of each class (post-stratum) match the population proportions, separately within each school (sampling stratum).

3.3 Post-Stratum Weights Within Stratified Sampling Calculations

If you find that the post-stratum proportions do not match within each sampling stratum, then you should create

- Post-stratification weights separately within each stratum h;
- $\overline{y}_{h,w}$'s separately within each stratum h;
- $\text{Var}(\overline{y}_{h,w})$'s separately within each stratum h.

These can then be combined using the usual formulas for stratified sampling:

$$\overline{y}_{st,w} = \sum_{h=1}^{H} W_h \overline{y}_{h,w} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \overline{y}_{h,w}$$

and

$$\text{Var}\,(\overline{y}_{st,w}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \text{Var}\,(\overline{y}_{h,w}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - n_h/N_h) \text{Var}\,(\overline{y}_{h,w})$$

where $\text{Var}\,(\overline{y}_{h,w})$ can be calculated using the TS or JK method, within each stratum h.