Calculating Post-Stratified Proportions

John Williams
Department of Marketing
University of Otago

July 8, 2009

1 Introduction

In the presentation I gave about my research into support for public funding of the Otago Stadium, I gave an example of post-stratification. In what follows I will explain the details behind that example, and give you a chance to practice in order to check your understanding.

Example: Post-Stratification by Sex

Notation:

N	total population size
N_m	number of males in the population
N_f	number of females in the population
$\widehat{\pi}_m$	proportion of males in favour of public funding
$\widehat{\pi}_f$	proportion of females in favour

The post-stratified estimate of the proportion of Dunedin residents over the age of 18 who are in favour of public funding of the Otago Stadium, $\widehat{\pi}_{\text{post}}$ (where the subscript post denotes "post-stratified"), is:

$$\begin{aligned}
\widehat{\pi}_{\text{post}} &= \frac{N_f}{N}\widehat{\pi}_f + \frac{N_m}{N}\widehat{\pi}_m \\
&= \frac{64,398}{123,516}23.5 + \frac{59,118}{123,516}31.1 \\
&= 27.16 \quad (2\text{dp.})
\end{aligned}$$

From this example you should be able to deduce the general formula for post-stratified proportions (i.e. where there are more than two strata). But before we skip to the end, let's make everything crystal clear, shall we?

2 Notation for Stratification

The notation in this document follows that of Lohr (1999), which is the best book on sampling that I've ever read.

Let the population size be denoted N, the sample size n and the number of strata H. Index each stratum by h [1] and each sampling unit by j. Define $\mathcal{S}_h$ to be the set of n_h sampling units in a simple random sample for stratum h. Finally, the symbol $\in$ is used in set notation to mean "in".

Now we can define the following population quantities:

$$\begin{aligned} y_{hj} &= \text{the value of the } j\text{th unit in stratum } h \\ t_h &= \sum_{j=1}^{N_h} y_{hj} = \text{the population total in stratum } h \\ t &= \sum_{h=1}^{H} t_h = \text{the population total} \end{aligned}$$

3 Mean and Variance of Stratified Samples

With our notation defined, we can now define the sample mean in stratum h:

$$\bar{y}_h = \frac{\sum_{j \in \mathcal{S}_h} y_{hj}}{n_h} \tag{1}$$

We can also define the post-stratified estimate of the overall sample mean:

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \tag{2}$$

What is the variance of this estimate? If certain conditions are met we can use the formula for the variance of a proportional sample[2] to estimate the variance of the post-stratified mean. These conditions are:

1. N and N_h are known (for all h)
2. n_h is "reasonably large", which can be taken to mean "about 30 or more"
3. n is large

The formula is:

$$\widehat{V}_{\text{post}} \approx \left(1 - \frac{n}{N}\right) \sum_{h=1}^{h} \frac{N_h}{N} \frac{s_h^2}{n} \tag{3}$$

where s_h^2 is the sample variance in stratum h and the symbol $\approx$ means "is approximately equal to".

[1] That is, h can take the values $1, 2, \ldots, H$

[2] A proportional sample is where the sample sizes of each stratum are in the same proportions as the population sizes of each stratum

4 Application to Proportions

Although we commonly think of means and proportions as being different, consider that if a variable is coded 0 and 1, then the mean of that variable on one hand, and the proportion of the cases that have the value 1 on the other hand, are exactly equivalent. It turns out that one can use all the equations given so far for proportions as well as means, simply by setting $\bar{y}_h = \widehat{\pi}_h$ and $s_h^2 = [n_h/(n_h - 1)]\widehat{\pi}_h(1 - \widehat{\pi}_h)$

Then an estimate of an overall sample proportion is:

$$\widehat{\pi}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \widehat{\pi}_h \tag{4}$$

And the variance of that estimate is:

$$\widehat{V}\left(\widehat{\pi}_{\text{post}}\right) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\widehat{\pi}(1 - \widehat{\pi})}{n_h - 1} \tag{5}$$

5 Exercises

Got all that straight? Then here's a few exercises for you.

5.1 Calculations

The post-stratified estimates of the proportion of residents in favour of public funding of the Otago Stadium that are published in the survey report are reproduced here in Table 1.

Strata	$\widehat{\pi}_{\text{post}}$
Sex	27.2
Age	30.1
Income	20.5

Table 1: Post-Stratified Estimates of Support

I calculated these estimates using my statistical analysis software (**R** [3]) using the data in Table 2 on the next page. These data are from the 2006 Census and represent Dunedin residents aged 20 and older.

Exercise: Using the data in Table 2 on the following page, check whether the estimates in Table 1 are correct.

5.2 Discussion Questions

The estimates in Table 1 vary quite dramatically. Why is that? Could you explain it to someone who doesn't know much about statistics or mathematics, without using equations? How would you do so?

Can we say that one or more variables are the "right" one(s) to use to post-stratify the estimates? Why or why not?

[3] `http://www.r-project.org`

Strata	N_h	$\widehat{\pi}_h$
SEX		
Female	64,398	23.51
Male	59,118	31.13
Sum	123,516	
AGE		
20 24	14,376	41.18
25 – 29	7,020	27.27
30 – 34	7,434	35.96
35 – 39	7,815	26.88
40 – 44	8,511	30.60
45 – 49	8,469	26.34
50 – 54	7,584	28.81
55 – 59	7,080	31.11
60 – 64	5,222	26.77
65 and over	16,542	23.37
Sum	90,053	
INCOME		
20k or less	42,201	14.05
20k – 30k	14,064	30.69
30k – 40k	8,916	24.24
40k – 50k	4,836	21.92
50k – 70k	4,179	29.91
70k – 100k	1,416	32.53
100k or more	1,179	44.95
Sum	76,791	

Table 2: Demographic Characteristics of Dunedin City. Source: Statistics New Zealand (2006 Census)

6 Further Reading

Most of the material contained here (and a lot more besides!) can be found in Lohr (1999, Chapter 4). In particular, proofs of the unbiasedness of the estimators presented here can be found on page 100 of that book.

References

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.