36-303 Sampling, Surveys & Society
Homework 03 Solutions

April 8, 2008

1 Question 1 (Groves Ch.4 #8)

1.1 (a)

$$\bar{y}_{\text{uw}} = \frac{\sum_h \bar{y}_h}{3} \tag{1}$$

$$= \frac{6+5+8}{3} = 6.333 \tag{2}$$

1.2 (b)

$$\bar{y}_{\text{st}} = \sum_h W_h \bar{y}_h \tag{3}$$

$$= 0.4 \times 6 + 0.5 \times 5 + 0.1 \times 8 = 5.7 \tag{4}$$

1.3 (c)

$$V[\bar{y}_{st}] = \sum_h W_h^2 \left(\frac{1-f_h}{n_h}\right) S_h^2 \tag{5}$$

$$= 0.4^2 \left(\frac{1-0.06}{192}\right) 5 + 0.5^2 \left(\frac{1-0.06}{240}\right) 4 + 0.1^2 \left(\frac{1-0.06}{48}\right) 7 \tag{6}$$

$$= 0.00920 \tag{7}$$

A 95% confidence interval can be constructed as

$$\begin{aligned} CI_{95\%} &= \bar{y}_{st} \pm 1.96\sqrt{V[\bar{y}_{st}]} && (8) \\ &= 5.7 \pm 1.96 \times 0.096 && (9) \\ &= 5.7 \pm 0.188 && (10) \\ &= [5.5, 5.9] && (11) \end{aligned}$$

2 Question 2 (Groves Ch.6 #2)

We evaluate the non-response error as the difference between the statistic that you *would have obtained* had everybody answered and the statistic computed from the incomplete (due to non-response) data. Usually this evaluation is not possible because we don't know the answers from the non-respondents, but in this case that information is given to us "from external sources".

Let's compute the expected total number of schools in the *original* sample that offer sex ed.

- Majority CRG

$$5\% \times 250 + 0\% \times (500 - 250) = 12.5 \tag{12}$$

- Minority CRG

$$50\% \times 600 + 35\% \times (1000 - 600) = 440 \tag{13}$$

Therefore the percentage of schools *the original sample* offering sex. ed. is $(440 + 12.5)/1500 \times 100\% = 30.2\%$. Comparing with the percentage computed using only the respondents, 36.8%, we get a non-response error of $36.8\% - 30.2\% = 6.6\%$.

A caution note: The text of the question says that this was a SRS. However, the "beautiful" numbers 1000 and 500, that supposedly were obtained by chance alone, should call to suspicion and make the analyst wonder if they are a product of a designed stratified sample. Of course, this is a textbook exercise and we know that the numbers are tweaked so they are nice and round, but in real life we should double check with the data producer or pay closer attention to the survey design documentation.

3 Question 3 (Groves Ch.6 # 9)

3.1 (a)

(These are just examples, any reasonable answer will be considered correct.)

1. Pre notifications
2. Incentives
3. Reducing the length of the questionnaire

3.2 (b)

As an example, the incentives method might not be effective with highly affluent people, whom might evaluate that their opportunity cost is much higher than any reasonable incentive that the survey administrator could offer.

4 Question 4 (Groves Ch. 10 #3)

Inputing the data into R:

```
BW <- c(4.4, 4.2, 2.4, 3, 2, 2.4, 2.6, 3, 3, 1.1, 1.3, 1.2, 1.4, 1.5,
1.1, 1.4, 1.8, 1.1, 1.1, 1)
Gender <- c(rep('M',6) ,rep('F',14))
Age <- c(24, 37, 66, 57, 23, 26, 28, 32, 39, 40, 41, 47, 43, 48, 53,
38, 63, 68, 73, 78)
NCC <- c(0, 1, 2, 3, 2, 0, 1, 1, 0, 1, 0, 2, 1, 1, 3, 3, 4, 1, 2, 3)
dat <- data.frame(BW = BW, Gender = factor(Gender), Age= Age,
     NCC = NCC, Wg = NA, nh =NA, PSW =NA)
n <- 20
```

4.1 (a)

Unweighted mean:

```
> mean(dat$NCC)
[1] 1.55
```

Weighted mean:

```
> weighted.mean(dat$NCC, w=dat$BW)
[1] 1.341463
```

4.2 b)

We first input the population proportions for each post stratum

```
dat$Wg[dat$Gender=='M' & dat$Age<40] <- 0.22
dat$Wg[dat$Gender=='M' & dat$Age>=40] <- 0.24
dat$Wg[dat$Gender=='F' & dat$Age<40] <- 0.22
dat$Wg[dat$Gender=='F' & dat$Age>=40] <- 0.32
```

then compute the number of samples in each poststratum

```
dat$nh[dat$Gender=='M' & dat$Age<40] <- NROW(dat$nh[dat$Gender=='M' & dat$Age<40])
dat$nh[dat$Gender=='M' & dat$Age>=40] <- NROW(dat$nh[dat$Gender=='M' & dat$Age>=40])
dat$nh[dat$Gender=='F' & dat$Age<40] <- NROW(dat$nh[dat$Gender=='F' & dat$Age<40])
dat$nh[dat$Gender=='F' & dat$Age>=40] <- NROW(dat$nh[dat$Gender=='F' & dat$Age>=40])
```

And finally compute the post stratification weights:

```
dat$PSW <- (dat$Wg) / (dat$nh/n)
```

The base weights, Post stratification weights and final composite weights are

```
> cbind(dat[c('BW', 'PSW')], dat$BW*dat$PSW)
    BW  PSW dat$BW * dat$PSW
1  4.4 1.10            4.840
2  4.2 1.10            4.620
3  2.4 2.40            5.760
4  3.0 2.40            7.200
5  2.0 1.10            2.200
6  2.4 1.10            2.640
7  2.6 1.10            2.860
8  3.0 1.10            3.300
9  3.0 1.10            3.300
10 1.1 0.64            0.704
11 1.3 0.64            0.832
12 1.2 0.64            0.768
13 1.4 0.64            0.896
14 1.5 0.64            0.960
15 1.1 0.64            0.704
16 1.4 1.10            1.540
17 1.8 0.64            1.152
18 1.1 0.64            0.704
19 1.1 0.64            0.704
20 1.0 0.64            0.640
```

The mean computed with these weights is

```
> weighted.mean(dat$NCC, w = dat$BW*dat$PSW)
[1] 1.462913
```