36-303 Sampling, Surveys & Society
Homework 05 Solutions

April 8, 2011

Question 1 (taken from S. Lohr's book)

a)

The summary table provides the distinct error proportions for 5 groups of clusters.

$$\bar{y}_{cl} = \frac{1}{85}(0.01860 * 1 + 0.01395 * 1 + 0.009302 * 4 + 0.00465 * 22 + 0 * 57) = 0.002$$

now for the standard error of the estimate:

$$s^2_{\bar{y}_i} = \frac{1}{85-1} * [(0.01860 - 0.002)^2 + (0.01395 - 0.002)^2 + 4 * (0.009302 - 0.002)^2$$
$$+22 * (0.00465 - 0.002)^2 + 57 * (0 - 0.002)^2] = 1.2073 * 10^{-5}$$

$$var(\bar{y}_{cl}) = (1 - \frac{85}{828})\frac{1}{85} * s^2_{\bar{y}_i} = (1 - \frac{85}{828})\frac{1}{85} * 1.2073 * 10^{-5} = 1.274544 * 10^{-7}$$

$$se(\bar{y}_{cl}) = 0.000357$$

b)

an estimate of the total number of errors should be $178020 * 0.002 = 356$ based on our answer from part a, we could find a standard error estimate as

$$var(y_{\text{total}}) = var(178020 * \bar{y}_{cl}) = 178020^2 * var(\bar{y}_{cl})$$

so our standard error estimate should be $178020 * 0.000357 = 63.55314$

c)

In this case we have to think our universe as composed from 'fields'. We have $N = 828 \times 215$ fields and our sample is composed from $n = 85 \times 215$ fields. Assuming the error rate is the same as in the previous case,

$$\hat{p}_{SRS} = \frac{\text{fields with errors}}{n} = \frac{37}{85 \times 215} = 0.002025 \tag{1}$$

The interesting thing happens when we compute the variance assuming that the sample is effectively a SRS *of fields*:

$$\hat{V}[\hat{p}_{SRS}] = \left(1 - \frac{85 \times 215}{828 \times 215}\right) \frac{\hat{p}_{SRS}(1 - \hat{p}_{SRS})}{85 \times 215} = 9.92 \times 10^{-8}$$

If we compare this estimate with the estimate obtained in part a), assuming cluster sampling,

$$\hat{V}[\hat{y}_{cl}] = 1.26172 \times 10^{-7} \tag{2}$$

we see that this last variance estimate is bigger than the one computed assuming SRS. This is a general phenomenon when we have clustered samples. To achieve the same error levels, a clustered sample must be bigger than a SRS. This example also illustrate the problems of analyzing clustered samples using SRS methods: the SE of the estimates will be underestimated. This is dangerous because we (and others) will think that our point estimates are better than they really are.

Question 2

Creating the dataset in R:

```
strata <- data.frame(expand.grid(Sex=factor(c('M','F')),
     College=factor(c('Eng','Lib'))),
  n_h = c(8,4,2,6),
  N_h = c(617,450,380,551),
  sam_w = NA,
  Pop_W = NA
)
strata$Pop_W <- strata$N_h / sum(strata$N_h)
strata$sam_w  <- strata$n_h / sum(strata$n_h)
HrsWk <- c(28,29,23,35,29,30,34,31,30,31,36,33,27,28,29,30,28,28,32,30)
data <- cbind(strata[rep(1:NROW(strata), strata$n_h),],HrsWk)
```

a)

The mean of 'Hrs/Wk' is

```
> mean(data$HrsWk)
[1] 30.05
```

Since this is SRS without replacement, an estimate of the standard error of the mean is

$$\hat{SE}[\bar{y}] = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

where s^2 is the sample variance. Computing this,

```
> fpc <- 1 - sum(strata$n_h)/sum(strata$N_h)
> sqrt(fpc) * sqrt(var(data$HrsWk) / NROW(data))
[1] 0.6634416
```

b)

Computing the post-stratification weights,

```
data <- cbind(data,PSW = data$Pop_W/data$sam_w)
```

And the weighted mean using the post-stratification weights is

```
> weighted.mean(data$HrsWk, w = data$PSW)
[1] 29.9111
```

which is slightly lower than the one without using the population-level information.

c)

To estimate the SE using a first order Taylor series approximation we use the R procedure given in class (don't forget the finite population correction)

```
>tsv <- ts.variance(data$HrsWk, w = data$PSW)
> se.ts <- sqrt(fpc)*sqrt(tsv$var.ts)
> se.ts
[1] 0.6709889
```

Question 3

To estimate the SE using the Jackknife technique we use the R function given in class,

```
> stacked_strata <- data.frame(stratum = paste(data$Sex,
    data$College, sep='.'),data$N_h)
> jkv <- jk.variance(data$HrsWk, stacked_strata$stratum,
    unique(stacked_strata$data.N_h))
> jkv
$ybar.weighted
[1] 29.9111

$ybar.reps
[1] 29.9111

$var.jk
[1] 0.3419697
```

Then, the jacknife estimate of the SE is

```
> se.jk <- sqrt(fpc)*sqrt(jkv$var.jk)
> se.jk
[1] 0.5818476
```