STAT 4234/5234: Calculating estimates from a two-stage cluster sample

Consider a population of N psus, where the ith psu consists of M_i ssus. So the population can be written

$$\{y_{ij} : j = 1, \ldots, M_i \ ; \ i = 1, \ldots, N\} \ .$$

Let $M_0 = \sum_{i=1}^{N} M_i$ denote the total population size (number of ssus). Further let

$$\bar{y}_{iU} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} \qquad \text{and} \qquad S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{iU})^2$$

denote the population mean and population variance, respectively, in psu i, and

$$\bar{y}_U = \frac{1}{M_0} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \qquad \text{and} \qquad S^2 = \frac{1}{M_0 - 1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_U)^2$$

denote the mean and variance for the entire population.

A two-stage cluster sample consists of (i) an SRS of n psus, and (ii) for each sampled psu, i.e., for each $i \in \mathcal{S}$, an SRS of m_i ssus. Let

$$\bar{y}_i = \frac{1}{m_i} \sum_{j \in \mathcal{S}_i} y_{ij} \qquad \text{and} \qquad s_i^2 = \frac{1}{m_i - 1} \sum_{j \in \mathcal{S}_i} (y_{ij} - \bar{y}_i)^2$$

denote the sample mean and sample variance, respectively, in psu i. Note that, in one-stage cluster sampling, $m_i = M_i$ for each $i \in \mathcal{S}$.

Then we estimate the population total $t = \sum_{i=1}^{N} M_i \bar{y}_{iU}$ by

$$\hat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \bar{y}_i \ . \tag{1}$$

The standard error of $\hat{t}_{\text{unb}}$ is the square root of

$$\hat{V}\left(\hat{t}_{\text{unb}}\right) = N^2 \frac{s_t^2}{n} \left(1 - \frac{n}{N}\right) + \frac{N}{n} \sum_{i \in \mathcal{S}} M_i^2 \frac{s_i^2}{m_i} \left(1 - \frac{m_i}{M_i}\right) \tag{2}$$

where

$$s_t^2 = \frac{1}{n - 1} \sum_{i \in \mathcal{S}} \left(M_i \bar{y}_i - \hat{t}_{\text{unb}}/N\right)^2 \ . \tag{3}$$

Point estimate and standard error for estimating the population mean $\bar{y}_U$ are

$$\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{M_0} \qquad \text{and} \qquad \text{SE}\left(\hat{\bar{y}}_{\text{unb}}\right) = \frac{\text{SE}\left(\hat{t}_{\text{unb}}\right)}{M_0} \ .$$

But a better approach, based on ratio estimation, is

$$\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i} \,. \tag{4}$$

The ratio estimator is not unbiased, but will usually have a lower standard error than $\hat{\bar{y}}_{\text{unb}}$, particularly if there is much variability in the size of the clusters (psus).

We have

$$\hat{V}\left(\hat{\bar{y}}_r\right) = \frac{s_r^2}{n\bar{M}^2}\left(1 - \frac{n}{N}\right) + \frac{1}{nN\bar{M}^2} \sum_{i \in \mathcal{S}} M_i^2 \frac{s_i^2}{m_i}\left(1 - \frac{m_i}{M_i}\right) \tag{5}$$

where

$$s_r^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} M_i^2 \left(\bar{y}_i - \hat{\bar{y}}_r\right)^2 \,. \tag{6}$$

Computing

In the Data folder on the Courseworks there is a file called `clustersamp.csv` which contains data from a hypothetical two-stage cluster sample. Suppose the population consists of $N = 15$ clusters (psus), and our random sample of $n = 3$ of them yields clusters 6, 4, and 5. We assume the cluster sizes M_i are unknown except for the sampled clusters $i \in \mathcal{S} = \{6, 4, 5\}$. Thus the total population size M_0 is unknown as well. For the second stage we have random samples of size $m_6 = 2$, $m_4 = 3$ and $m_5 = 4$. Also it is known that $M_6 = 5$, $M_4 = 8$ and $M_5 = 10$.

```
> filename <- "~/Data/clustersamp.csv"
> Data <- read.csv(filename)
> Data
  psu size    y
1   6    5 12.1
2   6    5 14.3
3   4    8 11.1
4   4    8 13.3
5   4    8 10.4
6   5   10 13.2
7   5   10 14.7
8   5   10 15.1
9   5   10 15.2
> dim(Data); names(Data);
[1] 9 3
[1] "psu"  "size" "y"
```

It will be most efficient to work with the `list` version of the data created using the `split` command.

```
> ysamp <- split(Data$y, Data$psu)
> ysamp
$`4`
[1] 11.1 13.3 10.4

$`5`
[1] 13.2 14.7 15.1 15.2

$`6`
[1] 12.1 14.3

> n <- length(ysamp); n;
[1] 3
> m <- sapply(ysamp, length); m;
4 5 6
3 4 2
> ybar <- sapply(ysamp, mean); ybar;
    4     5     6
11.60 14.55 13.20
> s2 <- sapply(ysamp, var); s2;
        4         5         6
2.2900000 0.8566667 2.4200000
> M <- c(8, 10, 5); names(M) <- names(m); M;
 4  5  6
 8 10  5
> N <- 15
```

Compute the estimators in (1) and (4): the unbiased estimator of the population total, and the ratio estimator of the population mean.

```
> t.hat.unb <- (N/n) * sum(M * ybar)
> ybar.hat.r <- sum(M * ybar) / sum(M)
> t.hat.unb; ybar.hat.r;
[1] 1521.5
[1] 13.23043
```

We estimate the population total by $\hat{t}_{\text{unb}} = 1521.5$ and the population mean by $\hat{\bar{y}}_r = 13.23$.

Note that s_t^2 in (3) is the sample variance of the $\{M_i\bar{y}_i : i \in \mathcal{S}\}$, and that s_r^2 in (6) is the sample variance of the $\left\{M_i\left(\bar{y}_i - \hat{\bar{y}}_r\right) : i \in \mathcal{S}\right\}$.

```
> s2.t <- var(M * ybar); s2.t;
[1] 1635.963
> s2.r <- var(M * (ybar - ybar.hat.r)); s2.r;
[1] 172.1404
```

We can now calculate the sample variances in (2) and (5).

```
> V.term2 <- sum(M^2 * s2/m * (1 - m/M))
> V.hat.t <- N^2 * s2.t/n * (1 - n/N) + (N/n) * V.term2
> V.hat.ybar <- 1/mean(M)^2 * (s2.r/n * (1 - n/N) + 1/(n*N) * V.term2)
> V.hat.t; V.hat.ybar;
[1] 98465.47
[1] 0.8042411
```

Thus

```
> SE.t <- sqrt(V.hat.t)
> t.hat; SE.t;
[1] 1521.5
[1] 313.7921
```

we estimate the population total by 1521.5 with a standard error of 313.8, and

```
> SE.ybar <- sqrt(V.hat.ybar)
> ybar.hat.r; SE.ybar;
[1] 13.23043
[1] 0.8967949
```

we estimate the population mean by 13.23 with a standard error of 0.90.