

## Computer Exercise 11 for Sample Survey

### Practice Problems

1. Open the data set `p33.dat`.

```
y=scan("http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/p33.dat")
```

Report the total number of observations  $M$ . Write the data into a  $\bar{M}$  called `Mb` by  $N=50$  matrix using `ymat = matrix(y, Mb, N, byrow = T)` so that each of the  $N$  columns represents a systematic sample or a cluster which contains *small, middle and large observations*. Output the vector `ym` of  $N=50$  column means and variance `s2` for each systematic sample. The command for `ym` is `ym = apply(ymat, 2, mean)`.

2. Calculate the between, within and total sum of squares:

$$S_b^2 = \frac{\bar{M}}{N-1} \sum_{j=1}^N (\bar{y}_j - \bar{y})^2, \quad S_w^2 = \frac{\bar{M}-1}{M-N} \sum_{j=1}^N s_j^2, \quad \text{and} \quad S^2 = \frac{1}{M-1} \sum_{j=1}^N \sum_{i=1}^{\bar{M}} (y_{ij} - \bar{y})^2$$

where  $M = \bar{M}N$ . Hence show that a systematic sample or a cluster sample is more efficient than a simple random sample.

3. Regarding the 50 columns as 50 clusters, draw a sample of  $n=10$  clusters and output the sample `ysam` in a `Mb` by `n` matrix.

```
set.seed(12345)
inds1 = sample(1:N, n, replace = FALSE, prob = NULL)
inds1
ysam = ymat[, inds1]
ysam
```

Output the vector of  $n=10$  cluster totals  $y_i$  called `yi`. Hence estimate the mean per cluster  $\hat{Y}_{c1}$  and the mean per element  $\hat{\hat{Y}}_{c1}$  ( $\hat{R}_{c1}$ ) and provide the standard error estimates.

Note that ratio estimators are not used because the cluster sizes  $M_i = \bar{M} = 20$  are all the same and hence the ratio estimators become the ordinary estimators.

4. (Advanced must; Normal optional) Draw a subsample of  $\bar{m}$  called `mb=8` observations from each selected cluster and output the sample into a `mb` by `n` matrix.

```

mb = 8
ysam2 = matrix(NA, mb, n)
for (j in 1:n) {
+ ysam2[, j] = sample(ysam[, j], mb, replace = FALSE, prob = NULL)
+ }
ysam2

```

Output the vector of estimated cluster totals  $\hat{y}_i$  called **yih** and variances **ys2**. Hence estimate the mean per cluster  $\hat{Y}_{c2}$  and the mean per element  $\hat{\hat{Y}}_{c2}$  ( $\hat{R}_{c2}$ ) and provide the standard error estimates.

Comment the additional variance due to  $\hat{y}_i$ . Compare the results with those from one-stage cluster sample in Q3 and with those from SRS, SRS with poststratification and stratified SRS in Practical 10 and 11.