STAT 4234/5234 Survey Sampling: Calculating estimates from stratified random samples

Here is some R code to take a stratified random sample, estimate the population mean $\bar{y}_U$ by the stratifed estimator $\bar{y}_{\text{str}}$, and calculate the standard error of the estimate.

On the Courseworks there is a file stratpop1.txt containing a simple population with two strata — stratum a of size 5 and stratum b of size 7. Save this file to your 'Data' folder, then read the data into R by

```
> Population <- read.csv("~/Data/stratpop1.csv", header=T)
> Population
      y strat
1  10.21     b
2   6.05     a
3   8.80     b
4   3.35     a
5   5.94     a
6  11.22     b
7   7.87     b
8   5.57     a
9  13.02     b
10  9.69     b
11  6.60     a
12  9.23     b
```

The best way to work with stratified populations and samples in R is to use the `split` command, as illustrated here.

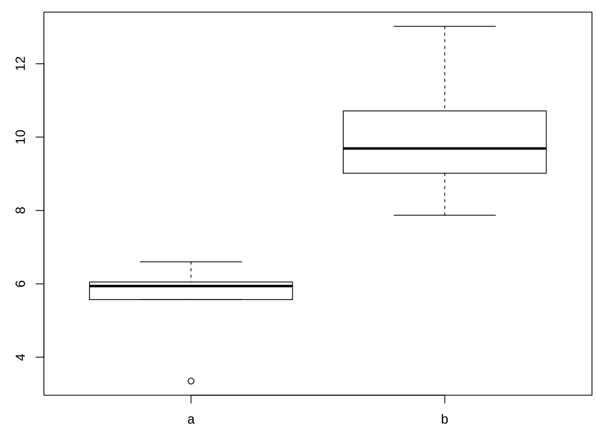
```
> names(Population)
[1] "y"     "strat"
> y <- split(Population$y, Population$strat)
> y
$a
[1] 6.05 3.35 5.94 5.57 6.60

$b
[1] 10.21  8.80 11.22  7.87 13.02  9.69  9.23

> rm(Population)
> names(y)
[1] "a" "b"
```

The object `y` is a `list` in R, each element of the list is a stratum. We use the `sapply` command to find the stratum sizes and stratum means.

```
> N.h <- sapply(y, length); N.h;
a b
5 7
> N <- sum(N.h); N;
[1] 12
> ( ybar.hU <- sapply(y, mean) )
       a        b
 5.50200 10.00571
> (ybar.U <- sum(N.h/N * ybar.hU))
[1] 8.129167
```

The stratum sizes are $N_1 = 5$ and $N_2 = 7$, total population size is $N = 12$; the stratum means are $\bar{y}_{1U} = 5.502$ and $\bar{y}_{2U} = 10.006$, and population mean is $\bar{y}_U = 8.129$.

```
> boxplot(y)
```



Suppose we wish to take a stratified random sample by sampling $n_1 = 2$ units from the first stratum a and n_3 units from the second stratum b.

```
> n.h <- c(2,3); names(n.h) <- names(N.h); n.h;
a b
2 3
```

Now take the sample. We will set the seed manually, to ensure we get the same answer every time. This is useful for writing handouts and homework situations, but a bad idea in practice.

```
> set.seed(5234)
> samp <- list()
> for(h in names(n.h))
+ {
+  samp[[h]] <- sample(y[[h]], n.h[[h]])
+ }
```

```
> samp
$a
[1] 6.60 5.94

$b
[1]  8.80 13.02 11.22
```

Use `sapply` to find the sample mean for each stratum.

```
> ( ybar.h <- sapply(samp, mean) )
       a        b
 6.27000 11.01333
```

Calcuate $\bar{y}_{\text{str}} = \sum_h N_h \bar{y}_h / N$.

```
> (ybar.str <- sum(N.h/N * ybar.h))
[1] 9.036944
```

Get $\bar{y}_{\text{str}} = 9.037$.

Now calculate the standard error for this stratified estimator.

First we need the sample standard deviations,

```
> s.h <- sapply(samp, sd); s.h;
        a         b
0.4666905 2.1175772
```

so we can calculate

$$\hat{V}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

by the following R code.

```
> (V.hat <- sum((N.h/N)^2 * (s.h^2)/n.h * (1 - n.h/N.h)))
[1] 0.301982
```

We get $\hat{V}(\bar{y}_{\text{str}}) = 0.302$.

An approximate 95% confidence interval for the population mean $\bar{y}_U$ is

```
> CI.lower <- ybar.str - 1.96 * sqrt(V.hat)
> CI.upper <- ybar.str + 1.96 * sqrt(V.hat)
> c(CI.lower, CI.upper)
[1]  7.959868 10.114021
```

or we can go

```
> ybar.str + c(-1,1) * 1.96 * sqrt(V.hat)
[1]  7.959868 10.114021
```

We are 95% confident that the population mean $\bar{y}_U$ is between 7.96 and 10.11. (In fact we have seen that it is equal to $\bar{y}_U = 8.13$.)