**Survey Sampling**
**Statistics 4234/5234 — Fall 2018**

**Homework 5**

*Solutions:*

1. Letting

$$y_{ij} = \begin{cases} 1 & \text{error in field } j \text{ of claim } i \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, M = 215$ for $i = 1, \ldots, N = 828$, we are interested in estimating the overall error rate $\bar{y}_U$ and the total number of errors $t$.

This is a one-stage cluster sample from a population of clusters of equal size. Specifically, we have an SRS of $n = 85$ out of $N = 828$ psus, each consisting of $M = 215$ ssus. The observed frequency distribution of the $t_i = \sum_{j=1}^{M_i} y_{ij}$ for $i \in \mathcal{S}$ is

| $t_i$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Frequency | 57 | 22 | 4 | 1 | 1 |

(a) Esimate $\bar{y}_U$ by

$$\hat{\bar{y}} = \frac{1}{nM} \sum_{i \in \mathcal{S}} \sum_{j=1}^{M} y_{ij} = \frac{1}{n} \sum_{i \in \mathcal{S}} \bar{y}_i \ .$$

The standard error is given by the square root of

$$\hat{V}\left(\hat{\bar{y}}\right) = \frac{s_{\bar{y}}^2}{n} \left(1 - \frac{n}{N}\right)$$

where

$$s_{\bar{y}}^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\bar{y}_i - \hat{\bar{y}}\right)^2 \ .$$

```
> N <- 828; M <- 215;
> ybar.i <- rep(0:4, c(57,22,4,1,1)) / M
> ybar.hat <- mean(ybar.i); ybar.hat;
[1] 0.002024624
> 1 / ybar.hat
[1] 493.9189
> n <- length(ybar.i); n;
[1] 85
> SE.ybar.hat <- sd(ybar.i)/sqrt(n) * sqrt(1 - n/N)
> SE.ybar.hat
[1] 0.0003570679
```

We estimate the error rate per field to be 0.002025 (one error every 494 fields), with a standard error of 0.000357.

```
> ybar.hat + c(-1,1) * 1.96 * SE.ybar.hat
[1] 0.001324771 0.002724477
```

We are 95% confident that the proportion of fields with error is between .0013 and .0027.

(b) For the *total* number of errors we have

```
> N * M * ybar.hat
[1] 360.4235
> N * M * SE.ybar.hat
[1] 63.56523
> N * M * (ybar.hat + c(-1,1) * 1.96 * SE.ybar.hat)
[1] 235.8357 485.0114
```

We estimate 360 errors total; we are 95% confident that the total number of errors is between 235 and 485.

Also, the estimator $\hat{t} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i$ has standard error given by the square root of

$$\hat{V}\left(\hat{t}\right) = N^2 \frac{s_t^2}{n} \left(1 - \frac{n}{N}\right)$$

so of course

```
> t.i <- rep(0:4, c(57,22,4,1,1))
> t.hat <- N * mean(t.i); t.hat;
[1] 360.4235
> SE.t.hat <- N * sd(t.i)/sqrt(n) * sqrt(1 - n/N)
> SE.t.hat
[1] 63.56523
```

gives the exact same result.

2. We have $N = 580$ cases, each has $M = 24$ cans; we sample $n = 12$ of the cases, and $m = 3$ cans from each sampled case.

```
> N <- 580; M <- 24; n <- 12; m <- 3;
```

Let $y_{ij}$ = number of worm fragments in the $j$th can of the $i$th case.

```
> ysamp <- matrix(c(1,4,0,3,4,0,5,3,7,3,4,0,
+                   5,2,1,6,9,7,5,0,3,1,7,0,
+                   7,4,2,6,8,3,1,2,5,4,9,0),
+   3, 12, byrow=T)
> rownames(ysamp) <- paste("Can", 1:3)
> colnames(ysamp) <- paste("Case", 1:12)
> ysamp
```

```
      Case 1 Case 2 Case 3 Case 4 Case 5 Case 6 Case 7 Case 8 Case 9
Can 1      1      4      0      3      4      0      5      3      7
Can 2      5      2      1      6      9      7      5      0      3
Can 3      7      4      2      6      8      3      1      2      5
      Case 10 Case 11 Case 12
Can 1       3       4       0
Can 2       1       7       0
Can 3       4       9       0
```

Estimate the average number of worm fragments per can by

$$\hat{\bar{y}}_{\text{unb}} = \frac{N \sum_{i \in \mathcal{S}} M_i \bar{y}_i}{n M_0}$$

which, with clusters of equal size, and equal cluster sample sizes, reduces to the straight sample mean for the 36 cans inspected.

```
> ybar <- apply(ysamp, 2, mean)
> N * sum(M*ybar) / (n*N*M)
[1] 3.638889
> ybar.hat <- mean(ysamp); ybar.hat;
[1] 3.638889
```

We can compute the standard error as $\frac{1}{M_0}\sqrt{\hat{V}(\hat{t})}$ where

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

where

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

```
> s2 <- apply(ysamp, 2, var)
> s2.t <- var(M*ybar)
> V.hat <- N^2 * s2.t/n * (1 - n/N) + N/n * sum(M^2 * s2/m * (1 - m/M))
> SE.ybar <- sqrt(V.hat) / (N*M)
> ybar.hat; SE.ybar;
[1] 3.638889
[1] 0.6101924
```

We estimate that there are, on average, 3.64 worm fragments per can in the latest shipment; the standard error of our estimate is 0.61.

```
> N*M*ybar.hat; N*M*SE.ybar;
[1] 50653.33
[1] 8493.878
```

We estimate that there are a total of 50,653 worm fragments in the the latest shipment; the standard error of our estimate is 8494.

3. Let

$$y_{ij} = \begin{cases} 1 & \text{female student } j \text{ at school } i \text{ smokes} \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, M_i$ for $i = 1, \ldots, N = 29$.

We wish to estimate $\bar{y}_U$, the proportion of female high school students in the region who smoke.

In a population of $N = 29$ psus we have a random sample of $n = 4$ of them. The data:

| School $i$ | $M_i$ | $m_i$ | $\sum_{j \in \mathcal{S}_i} y_{ij}$ |
|---|---|---|---|
| 1 | 792 | 25 | 10 |
| 2 | 447 | 15 | 3 |
| 3 | 511 | 20 | 6 |
| 4 | 800 | 40 | 27 |

We estimate $\bar{y}_U$ by

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

```
> M <- c(792, 447, 511, 800)
> m <- c(25, 15, 20, 40)
> ysum <- c(10, 3, 6, 27)
> ybar.hat <- sum(M * ysum/m) / sum(M)
> ybar.hat
[1] 0.4311765
```

We estimate that 43.12% of female students in the region smoke.

The standard error of our estimator is the square root of

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

where

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

```
> N <- 29; n <- 4;
> s2.r <- var( M*(ysum/m - ybar.hat) )
> s2 <- rep(NA, n)
> for(i in 1:n)
+ {
+  ysamp <- rep(c(1,0), c(ysum[i], m[i]-ysum[i]))
+  s2[i] <- var(ysamp); rm(ysamp); }
```

```
> s2.r; s2;
[1] 17943.04
[1] 0.2500000 0.1714286 0.2210526 0.2250000
> V.sum <- sum(M^2 * s2/m * (1 - m/M))
> V.hat <- 1/(n*mean(M)^2) * (s2.r*(1 - n/N) + (1/N)*V.sum)
> sqrt(V.hat)
[1] 0.09910716
> ybar.hat + c(-1,1) * 1.96 * sqrt(V.hat)
[1] 0.2369264 0.6254265
```

We are 95% confident that the proportion of female high school students in this region who smoke is between 0.237 and 0.625.

For population total we have

```
> t.hat <- N/n * sum(M * ysum/m); t.hat;
[1] 7971.375
> s2.t <- var(M * ysum/m); s2.t;
[1] 40410.14
> V.hat.t <- N^2 * s2.t/n * (1 - n/N) + N/n * sum(M^2 * s2/m * (1 - m/M))
> SE.t <- sqrt(V.hat.t); SE.t;
[1] 2725.67
> t.hat + c(-1,1) * 1.96 * SE.t
[1]  2629.061 13313.689
```

We are 95% confident that the total number of female high school students in the region who smoke is between 2629 and 13314.