GR5234 - HW6

Mathieu Sauterey - UNI: mjs2364

November 11, 2018

Problem 1

A shipment of 100 boxes of frozen food (each box contains 8 separate packages of food) was allowed to thaw during transit. The shipper was worried that some of the boxes could be spoiled. So he took a random sample of 5 boxes and checked all the packages in each box: In 2 of the boxes there were 3 spoiled packages, in one of the boxes there were 2 spoiled packages, and in 2 of the boxes there were no spoiled packages.

(a) Estimate the total number of spoiled packages in the entire shipment, and give a standard error for your estimate.

```
N = 100

n = 5

M = rep(8, n)

m = rep(8, n)

y = data.frame(y4 = rep(c(1,0), times = c(3,5)),
               y3 = rep(c(1,0), times = c(3,5)),
               y2 = rep(c(1,0), times = c(2,6)),
               y1 = rep(c(0), times = c(8)),
               y0 = rep(c(0), times = c(8)))

ybar = apply(y, 2, mean)

s2 = apply(y, 2, var)

t_hat_unb = (N/n) * sum(M * ybar)

s2.t <- var(M * ybar)

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

SE.t <- sqrt(V.hat.t)

print(t_hat_unb)
```

```
## [1] 160
```

```
print(SE.t)
```

```
## [1] 66.10598
```

$\hat{t}_{unb} = 160$ spoiled packages

$\text{SE}[\hat{t}_{unb}] = 66.106$ spoiled packages

(b) Suppose instead that the sampling plan took a simple random sample of size 40 from the population of 800 packages, and 8 spoiled packages were found in the sample. Under this sampling plan estimate the total number of spoiled packages in the shipment, and give a standard error.

```
N = 800

n = 40

spoiled = rep(0:1, times = c(n-8,8))

mean_spoiled = mean(spoiled)
total_spoiled = mean_spoiled*N
print(total_spoiled)
```

```
## [1] 160
```

```
SE_mean_spoiled = sqrt(1-n/N)*sd(spoiled)/sqrt(n)
SE_total_spoiled = SE_mean_spoiled*N
print(SE_total_spoiled)
```

```
## [1] 49.94356
```

$\hat{t}_{SRS} = 160$ spoiled packages

$\text{SE}[\hat{t}_{SRS}] = 49.944$ spoiled packages

As expected, the standard error from the SRS sample is smaller than that from the one-stage cluster sampling.

Problem 2

Consider a population consisting of 20 clusters with a total population size of 940. Suppose we have a simple random sample of four of the clusters from the population which resulted in the following data.

Estimate the population mean and give a standard error for your estimate.

```
N = 20

n = 4

M_0 = 940

M = c(17, 56, 23, 64)

m = c(10, 25, 12, 30)


ybar = c(32.7, 36.1, 30.3, 33.4)
```

```r
s2 = c(26.3, 21.4, 23.6, 29.1)

ybar_hat_r = sum(M * ybar) / sum(M)

s2.r <- var(M * (ybar - ybar_hat_r))

V.term2 <- sum(M^2 * s2/m * (1 - m/M))

V.hat.ybar <- 1/mean(M)^2 * (s2.r/n * (1 - n/N) + 1/(n*N) * V.term2)

SE.ybar <- sqrt(V.hat.ybar)

print(ybar_hat_r)
```

```
## [1] 33.825
```

```r
print(SE.ybar)
```

```
## [1] 1.015215
```

$\hat{\bar{y}}_r = 33.825$

$\text{SE}[\hat{\bar{y}}_r] = 1.015$

Problem 3

A town has four supermarkets, ranging in size from 110 square meters (m2) to 1265 m2. We want to estimate the total amount of sales in the four stores for last month by sampling just two of the stores. Compare unequal-probability sampling with replacement to simple random sampling without replacement:

(a) Suppose we will select two stores with replacement, using psi_i's proportional to store size. Find E[t_hat_psi] and V [t_hat_psi].

```r
N = 4

n = 2

size = c(110, 265, 360, 1265)

sales = c(44, 82, 112, 362)

total_size = sum(size)

total_sales = sum(sales)

psi = size/total_size

u = sales/psi

Var_t_hat_psi = (1/n)*sum((u-total_sales)^2*psi)

print(Var_t_hat_psi)
```

```
## [1] 1410.124
```

$E[\hat{t}_{\psi}]$ = t = \$600,000 (the total size PPS estimator is unbiased)

$Var[\hat{t}_{\psi}]$ = \$1,410,124

(b) Find E[t_hat_SRS] and V [t_hat_SRS].

```
N = 4

n = 2

size = c(110, 265, 360, 1265)

sales = c(44, 82, 112, 362)

total_size = sum(size)

total_sales = sum(sales)

Var_t_hat_srs = N^2*var(sales)*(1-n/N)/n

print(Var_t_hat_srs)
```

```
## [1] 82997.33
```

$E[\hat{t}_{SRS}]$ = t = \$600,000 (the total size SRS estimator is unbiased)

$Var[\hat{t}_{SRS}]$ = \$82,997,330

The variance with SRS sampling is much larger (almost 60 times) than the variance of unequal probability sampling with replacement. This was expected given the large difference in sales volume across stores and the strong positive correlation between stores sales and size.