Survey Sampling

Statistics 4234/5234 — Fall 2018

Homework 4

Solutions:

1. Hard shell clams in Narragansett Bay, Rhode Island.

   (a) We estimate the total number of bushels of clams in the area by $\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h$. The standard error of our estimate is given by

   $$\text{SE}(\hat{t}_{\text{str}}) = \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) .$$

   We can perform these calculations in R.

```
> N.h <- 25 * c(222.81, 49.61, 50.25, 197.81)
> n.h <- c(4, 6, 3, 5)
> ybar.h <- c(0.44, 1.17, 3.92, 1.80)
> s2.h <- c(.068, .042, 2.146, .794)
> t.hat.str <- sum(N.h * ybar.h)
> V.hat <- sum(N.h^2 * s2.h/n.h * (1 - n.h/N.h))
> SE <- sqrt(V.hat)
> t.hat.str; SE;
[1] 17727.95
[1] 2354.492
```

   We estimate that there are 17,728 bushels of clams in the area, with a standard error of 2354 bushels. The usual confidence interval is probably not valid in this problem since the n_h are so small.

   (b) In the second survey, with only two strata, we get

```
> N.h <- 25 * c(322.67, 197.81)
> n.h <- c(8, 5)
> ybar.h <- c(0.63, 0.40)
> s2.h <- c(.083, .046)
> sum(N.h * ybar.h)
[1] 7060.153
> sqrt(sum(N.h^2 * s2.h/n.h * (1 - n.h/N.h)))
[1] 948.2723
```

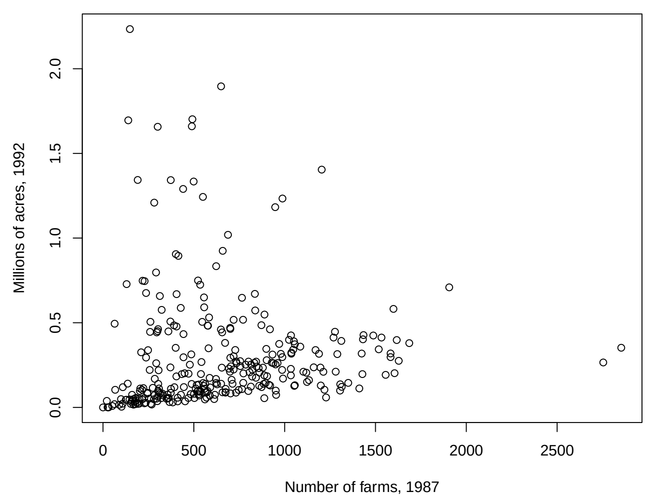
   At the end of the season we estimate only 7060 bushels of clams, with a standard error of 948 clams.

2. The data file `agsrs` contains information on the number of farms and acres devoted to farms, for an SRS of $n = 300$ counties from the population of $N = 3078$ in the United States. In 1987, the United States had a total of 2,087,759 farms.

```
> library(SDaA)
> dim(agsrs); names(agsrs);
[1] 300  14
 [1] "county"   "state"    "acres92"  "acres87"  "acres82"  "farms92"
 [7] "farms87"  "farms82"  "largef92" "largef87" "largef82" "smallf92"
[13] "smallf87" "smallf82"
```

Consider using x_i = number of farms in county i in 1987 as an auxiliary variables for estimating the total of y_i = acres of land devoted to farming in 1992.

```
> x <- agsrs$farms87; y <- agsrs$acres92;
> y <- y / 1e6
> plot(x, y, xlab="Number of farms, 1987", ylab="Millions of acres, 1992")
> y <- y * 1e6
> N <- 3078; xbar.U <- 2087759 / N;
```



(a) Ratio estimation.

```
> ratio.estimator.mean <- function(x.samp, y.samp, N, xbar.U)
+ {
+ n <- length(y.samp)
+ xbar <- mean(x.samp); ybar <- mean(y.samp);
+ B.hat <- ybar / xbar
+ ybar.hat.r <- B.hat * xbar.U
```

```
+ e <- y.samp - B.hat * x.samp
+ V.hat <- (xbar.U/xbar)^2 * var(e)/n * (1 - n/N)
+ SE <- sqrt(V.hat)
+ answer <- c(point.est=ybar.hat.r, std.error=SE)
+ return(answer)
+ }
> mean.farmland <- ratio.estimator.mean(x, y, N, xbar.U)
> N * mean.farmland
point.est std.error
960155061  68446406
> N * mean.farmland / 1e6
point.est std.error
960.15506  68.44641
```

Using ratio estimation we estimate 960 million acres of farmland in 1992, with a standard error of about 68 million; we can be 95% confident that in 1992 there were between 826 million and 1.094 billions of acres of farmland in the United States.

(b) Regression estimation.

```
> regression.estimator.mean <- function(x.samp, y.samp, N, xbar.U)
+ {
+ n <- length(y.samp)
+ xbar <- mean(x.samp); ybar <- mean(y.samp);
+ foo <- lsfit(x.samp, y.samp)
+ B1.hat <- as.numeric(foo$coefficients)[2]
+ ybar.hat.reg <- ybar + B1.hat * (xbar.U - xbar)
+ resids <- foo$residuals
+ V.hat <- var(resids) / n * (1 - n/N)
+ SE <- sqrt(V.hat)
+ answer <- c(point.est=ybar.hat.reg, std.error=SE)
+ return(answer)
+ }
> mean.farmland <- regression.estimator.mean(x, y, N, xbar.U)
> N * mean.farmland
point.est std.error
921406265  58065813
> N * mean.farmland / 1e6
point.est std.error
921.40627  58.06581
```

Using regression estimation we estimate 921 million acres of farmland in 1992, with a standard error of about 58 million; we can be 95% confident that in 1992 there were between 807 million and 1.035 billion acres of farmland in the United States.

3. Domain estimation.

```
> library(SDaA); rm(list=ls());
> dim(agsrs); names(agsrs);
[1] 300  14
 [1] "county"   "state"    "acres92"  "acres87"  "acres82"  "farms92"
 [7] "farms87"  "farms82"  "largef92" "largef87" "largef82" "smallf92"
[13] "smallf87" "smallf82"
> x <- agsrs$farms92; y <- agsrs$acres92;
> n <- length(y); n
[1] 300
> domain <- ifelse(x < 600, 1, 2)
> table(domain)
domain
  1   2
171 129
```

So there are 171 counties in our domain 1 (fewer than 600 farms) and 129 sampled counties in our domain 2 (600 or more).

```
> domain.estimation <- function(y.samp, domain.samp, d, N)
+ {
+ n <- length(y.samp); n.d <- sum(domain.samp==d);
+ y.samp.d <- y.samp[domain.samp==d]
+ ybar.d <- mean(y.samp.d); s2.yd <- var(y.samp.d);
+ V.hat <- n*(n.d-1)/(n.d*(n-1)) * s2.yd/n.d * (1 - n/N)
+ SE <- sqrt(V.hat)
+ answer <- c(point.est=ybar.d, std.error=SE)
+ return(answer)
+ }
> domain.estimation(y, domain, d=1, N=3078)
point.est std.error
283813.71  28852.24
> domain.estimation(y, domain, d=2, N=3078)
point.est std.error
316565.65  21553.21
```

We estimate the average farmland per county with less than 600 farms to be 283,814 acres, with a standard error of 28,852 acres.

We estimate the average farmland per county with 600 or more farms to be 316,566 acres, with a standard error of 21,553 acres.

Oops! The problem asked for estimation of *total number* of acres devoted to farming, not average per county. Domain estimation for population totals is a bit tricky — we'd like to just go $N_d \bar{y}_d$ but the N_d are not known. So instead we use

$$\hat{t}_{yd} = \hat{t}_u = N\bar{u}$$

and

$$\hat{V}(\hat{t}_{yd}) = N^2 \hat{V}(\bar{u}) = N^2 \frac{s_u^2}{n}\left(1 - \frac{n}{N}\right)$$

where

$$u_i = \begin{cases} y_i & i \in \mathcal{U}_d \\ 0 & i \notin \mathcal{U}_d \end{cases}$$

```
> N <- 3078
> u1 <- ifelse(domain==1, y, 0); u2 <- ifelse(domain==2, y, 0);
> N * mean(u1) / 1e6
[1] 497.9398
> N * sd(u1)/sqrt(n) * sqrt(1 - n/N) / 1e6
[1] 55.91952
```

We estimate about 498 million acres devoted to farming in counties with fewer than 600 farms, with a standard error of about 56 million.

```
> N * mean(u2) / 1e6
[1] 418.9873
> N * sd(u2)/sqrt(n) * sqrt(1 - n/N) / 1e6
[1] 38.93828
```

We estimate about 419 million acres devoted to farming in counties with 600 or more farms, with a standard error of about 39 million.