Survey Sampling
Statistics 4234/5234 — Fall 2018

Homework 3

Solutions:

1. The following table summarizes the info we are given in this problem.

h	Stratum	N_h	S_h	$N_h S_h$	p_h
1	Houses	35	2τ	70τ	.45
2	Apartments	45	τ	45τ	.25
3	Condos	10	τ	10τ	.12

   (a) For total sample size of $n = 900$, Neyman allocation $n_h \propto N_h S_h$ gives

$$n_1 = (70/125)(900) = 504$$

   and

$$n_2 = (45/125)(900) = 324$$

   and

$$n_3 = (10/125)(900) = 72$$

   (b) The overall proportion of households that practice energy conservation is

$$p = \sum \frac{N_h}{N} p_h = \left(\frac{35}{90}\right)(.45) + \left(\frac{45}{90}\right)(.25) + \left(\frac{10}{90}\right)(.12) = 0.3133$$

   So we have

$$V_{\text{SRS}}(\hat{p}) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right) = \frac{(.3133)(.6867)}{n}\left(\frac{N-n}{N-1}\right) = \frac{.215156}{n}\left(\frac{N-n}{N-1}\right)$$

   In proportional allocation $n_1 = 350$ and $n_2 = 450$ and $n_3 = 100$.

$$\begin{aligned} V_{\text{prop}}(\hat{p}_{\text{str}}) &= \sum \left(\frac{N_h}{N}\right)^2 \frac{p_h(1-p_h)}{n_h}\left(\frac{N_h - n_h}{N_h - 1}\right) \approx \sum \left(\frac{N_h}{N}\right)\frac{p_h(1-p_h)}{n}\left(\frac{N-n}{N-1}\right) \\ &\approx \frac{1}{n}\left[\left(\frac{35}{90}\right)(.45)(.55) + \left(\frac{45}{90}\right)(.25)(.75) + \left(\frac{10}{90}\right)(.12)(.88)\right]\left(\frac{N-n}{N-1}\right) \\ &= \frac{.201733}{n}\left(\frac{N-n}{N-1}\right) \end{aligned}$$

   So the gain in efficiency due to stratification is measure by

$$\frac{V_{\text{SRS}}(\hat{p})}{V_{\text{prop}}(\hat{p}_{\text{str}})} = \frac{.215156}{.201733} = 1.0665$$

   for about a 6.7% gain in efficiency.

2. R function to compute a $100(1-\alpha)\%$ confidence interval for the mean of a finite population, based on simple random sample.

```
> CI.mean <- function(y, N, alpha=.05)
+ {
+  n <- length(y); ybar <- mean(y);
+  SE <- sd(y) / sqrt(n) * sqrt(1 - n/N)
+  z.star <- qnorm(1 - alpha/2)
+  lower <- ybar - z.star * SE
+  upper <- ybar + z.star * SE
+  cbind(lower=lower, upper=upper)
+ }
```

Compute a 95% confidence interval for the total number of refereed publications by the 807 university faculty.

```
> Publications <- rep(0:10, c(28,4,3,4,4,2,1,0,2,1,1))
> CI.mean(y=Publications, N=807)
       lower   upper
[1,] 1.05988 2.50012
> CI.mean(y=Publications, N=807) * 807
        lower    upper
[1,] 855.3229 2017.597
```

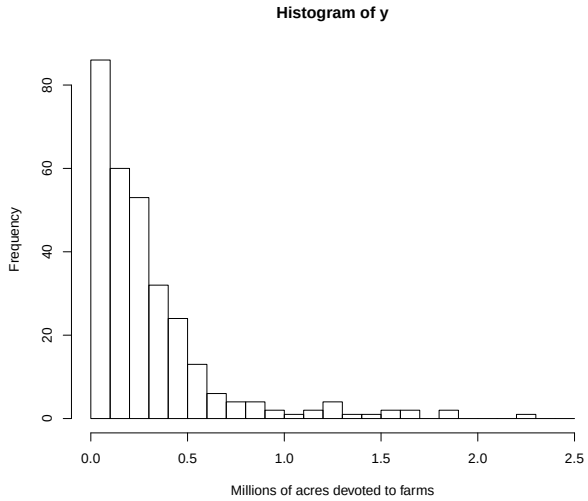
We are 95% confident that the total number of publications for this faculty is between 855 and 2018.

3. First obtain the data.

```
> 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"
```

Convert from acres to millions of acres, and construct a histogram.

```
> y <- agsrs$acres87 / 1e6
> range(y)
[1] 0.000000 2.266986
> hist(y, breaks=seq(0, 2.5, .1), right=F,
+  xlab="Millions of acres devoted to farms")
```



Heavily right-skewed distribution with most countries having fewer than 500,000 acres devoted to farms, and a handful having many more. One county in the sample (Hudspeth County in Texas) has more than 2 million acres of farmland.

We estimate the mean and median farmland per county by the sample mean and sample median:

```
> mean(y); median(y);
[1] 0.3019537
[1] 0.2063275
```

We estimate the mean farmland per county in 1987 was 302,000 acres, and the median county had about 206,000 acres devoted to farms.

A confidence interval for the population mean can be computed by

$$\bar{y} \pm 1.96 \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

Using R as a calculator we get

```
> ybar <- mean(y); ybar;
[1] 0.3019537
> s <- sd(y); s;
[1] 0.3448296
> n <- length(y); n;
[1] 300
> N <- 3078 # number of counties and county-equivalents in the US
> SE <- s / sqrt(n) * sqrt(1 - n/N); SE;
[1] 0.01891367
> CI.mean <- ybar + c(-1,1) * 1.96 * SE; CI.mean;
[1] 0.2648829 0.3390245
```

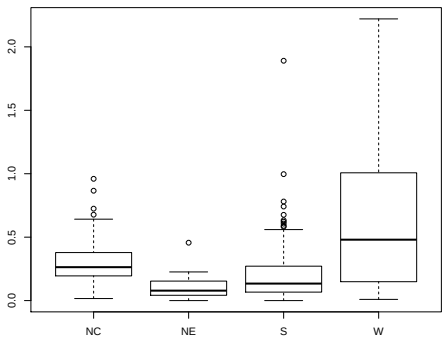
Multiply lower and upper bound by N, to get CI for population total.

```
> N * CI.mean
[1]  815.3097 1043.5174
```

We are 95% confident that, in 1987, there were between 815 million and 1.043 billion acres of farmland.

4. Obtain the data and construct adjacent boxplots of acres of farmland by county, by region.

```
> dim(agstrat); names(agstrat);
[1] 300  17
 [1] "county"   "state"    "acres92"  "acres87"  "acres82"  "farms92"
 [7] "farms87"  "farms82"  "largef92" "largef87" "largef82" "smallf92"
[13] "smallf87" "smallf82" "region"   "rn"       "weight"
> samp <- split(agstrat$acres87 / 1e6, agstrat$region)
> names(samp)
[1] "NC" "NE" "S"  "W"
> boxplot(samp)
```



Estimate population total by $\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h$ and get the following.

```
> ybar.h <- sapply(samp, mean);
> N.h <- c(1054, 220,1382, 422); names(N.h) <- names(ybar.h); N.h;
  NC   NE    S    W
1054  220 1382  422
> N <- sum(N.h); N; # Should be 3078
[1] 3078
> ybar.strat <- sum(N.h/N * ybar.h); ybar.strat;
[1] 0.2985471
> N * ybar.strat;
[1] 918.928
```

We estimate that in 1987 there were about 919 million acres of farmland in the country. The standard error of our estimate is $\text{SE}\left(\hat{t}_{\text{str}}\right) = \sqrt{\hat{V}\left(\hat{t}_{\text{str}}\right)}$, where

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

Using R we get

```
> s.h <- sapply(samp, sd); s.h;
       NC        NE         S         W
0.1715382 0.1002529 0.2387144 0.6143851
> n.h <- sapply(samp, length); n.h;
 NC  NE   S   W
103  21 135  41
> V.hat <- sum(N.h^2 * s.h^2 / n.h * (1 - n.h/N.h)); V.hat;
[1] 2515
> SE <- sqrt(V.hat); SE;
[1] 50.14977
```

The standard error is about 50 million acres. In the previous problem (ordinary SRS no stratification), the standard error was about 58 million.

```
> N*ybar.strat + c(-1,1) * 1.96 * SE
[1]  820.6344 1017.2215
```

We are 95% confident that, in 1987, there were between 820 million and 1.017 billion acres of farmland in the country.

5. Assuming sampling costs are the same in each stratum, optimal allocation gives $n_h \propto N_h S_h$.

```
> N.h
  NC   NE    S    W
1054  220 1382  422
> s.h
       NC        NE         S         W
0.1715382 0.1002529 0.2387144 0.6143851
> 300 * N.h*s.h / sum(N.h*s.h)
        NC         NE          S          W
 68.482674   8.354084 124.958528  98.204714
```

For a total stratified sample size $n = 300$ we should take

$$n_1 = 69 \qquad n_2 = 8 \qquad n_3 = 125 \qquad n_4 = 102$$

6. Obtain the population in R.

```
> y.pop <- split(agpop$acres87, agpop$region);
> names(y.pop);
[1] "NC" "NE" "S"  "W"
> N.h <- sapply(y.pop, length); N.h;
  NC   NE    S    W
1054  220 1382  422
```

And take the stratified sample described above.

```
> n.h <- c(69, 8, 125, 98); names(n.h) <- names(N.h); n.h;
 NC  NE   S   W
 69   8 125  98
> set.seed(5234)
> samp <- list()
> for(h in names(n.h))
+ {
+  samp[[h]] <- sample(y.pop[[h]], n.h[[h]])
+ }
```

Check for missing values, coded in `agpop` as -99.

```
> n.h
 NC  NE   S   W
 69   8 125  98
> for(h in names(samp))
+ {
+  samp[[h]] <- samp[[h]][ samp[[h]] > -99 ]
+ }
> n.h <- sapply(samp, length); n.h;
 NC  NE   S   W
 69   8 124  96
```

One missing value in the South and two in the West.

Calculate the estimate $\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h$ and its standard error.

```
> ybar.h <- sapply(samp, mean); ybar.h;
       NC        NE         S         W
321159.17  64804.38 219689.56 740910.17
> t.hat.str <- sum(N.h * ybar.h) / 1e6
> t.hat.str
[1] 969.0338
```

We estimate that in 1987 the total amount of US land devoted to farming was about 969 million acres.

The standard error of our estimate is given by the square root of $\hat{V}$ in (1).

```
> s.h <- sapply(samp, sd); s.h;
       NC        NE         S         W
215254.28  39150.57 265941.75 704605.39
> V.hat <- sum(N.h^2 * s.h^2/n.h * (1 - n.h/N.h))
> SE <- sqrt(V.hat); SE <- SE / 1e6; SE;
[1] 49.08322
```

The standard error for our estimate is about 49 million, only a little smaller than the one we got (about 50 million) using proportional allocation.

```
> t.hat.str + c(-1,1) * 1.96 * SE
[1]  872.8307 1065.2369
```

We are 95% confident that the total amount of US land devoted to farming in 1987 was between 873 million and 1.065 billion acres.

7. Obtain the data.

```
> dim(otters); names(otters);
[1] 82  3
[1] "section" "habitat" "holts"
> samp <- split(otters$holts, otters$habitat);
```

Estimate the total number of otter dens by $\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h$.

```
> N.h <- c(89,61,40,47); names(N.h) <- names(samp); N.h
 1  2  3  4
89 61 40 47
> ybar.h <- sapply(samp, mean); ybar.h;
        1         2         3         4
 1.736842  1.750000 13.272727  4.095238
> sum(N.h * ybar.h)
[1] 984.7142
```

We estimate that there are 985 otter dens along the Shetland coastline.

The standard error of our estimate is given by the square root of $\hat{V}$ in (1).

```
> s.h <- sapply(samp, sd); s.h;
       1        2        3        4
2.329571 2.613225 7.666761 3.948478
> n.h <- sapply(samp, length); n.h;
 1  2  3  4
19 20 22 21
> V.hat <- sum(N.h^2 * s.h^2/n.h * (1 - n.h/N.h))
> V.hat
[1] 5464.313
> SE <- sqrt(V.hat); SE;
[1] 73.92099
```

The standard error of our estimate is about 74.

```
> t.hat.str <- sum(N.h * ybar.h)
> t.hat.str + c(-1,1) * 1.96 * SE
[1]  839.8291 1129.5994
```

Thus we are about 95% confident that there are between 840 and 1130 otter dens along the Shetland coastline.