Mike Lederle
lederle@neo.tamu.edu

STAT 607 — HW 3

September 29, 2012

Exercise 3.6

Suppose that a city has 90000 dwelling units, of which 35000 are houses, 45000 are apartments, and 10000 are condominiums.

a You believe that the mean electricity usage is about twice as much for houses as for apartments or condominiums, and that the standard deviation is proportional to the mean, so that $S_1 = 2S_2 = 2S_3$. How would you allocate a stratified sample of 900 observations if you wanted to estimate the mean electricity consumption for all households in the city?

Use optimal allocation, since the variance between strata are different.

b Now suppose that you take a stratified random sample with proportional allocation and want to estimate the overall proportion of households in which energy conservation is practiced. If 45% of house dwellers, 25% of apartment dwellers, and 3% of condominium residents practice energy conservation, what is p for the population? What gain would the stratified sample with proportional allocation offer over an SRS, that is, what is $V_{prop}(\hat{p}_{str})/V_{SRS}(\hat{p}_{SRS})$?

Using a sample of size 900, $n_1 = 350$, $n_2 = 450$, and $n_3 = 100$. The calculation for $\hat{p}$ is

$$\frac{35000}{90000} \times .45 + \frac{45000}{90000} \times .25 + \frac{10000}{90000} \times .03$$

```
35/90 * 0.45 + 45/90 * 0.25 + 10/90 * 0.03

## [1] 0.3033
```

For SRS:

$$\hat{V}_{SRS}(\hat{p}) = \left(1 - \frac{900}{90000}\right) \frac{.303(1 - .303)}{900 - 1}$$

```
(srs <- (1 - 0.01) * (0.303 * (1 - 0.303))/(900 - 1))

## [1] 0.0002326
```

For proportional allocation:

$$\left(1-\frac{350}{35000}\right)\left(\frac{35000}{90000}\right)^2\frac{.45(1-.45)}{350-1}+\left(1-\frac{450}{45000}\right)\left(\frac{45000}{90000}\right)^2\frac{.25(1-.25)}{450-1}+\left(1-\frac{100}{10000}\right)\left(\frac{10000}{90000}\right)^2\frac{.03(1-.03)}{100-1}$$

```
(prop <- (1 - 350/35000) * (35000/90000)^2 * (0.45 * (1 - 0.45))/(350 - 1) +
    +(1 - 450/45000) * (45000/90000)^2 * (0.25 * (1 - 0.25))/(450 - 1) + +(1 -
    100/10000) * (10000/90000)^2 * (0.03 * (1 - 0.03))/(100 - 1))

## [1] 0.0002131
```

The ratio is

```
prop/srs

## [1] 0.9164
```

Exercise 3.11

Lydersen and Ryg (1991) used stratification techniques to estimate ringed seal populations in Svalbard fjords. The 200 km^2 study area was divided into three zones: Zone 1, outer Sassenfjorden, was covered with relatively new ice during the study period in March 1990, and had little snow cover; Zone 3, Tempelfjorden, had a stable ice cover throughout the year; Zone 2, inner Sassenfjorden, was intermediate between the stable Zone 3 and the unstable Zone 1. Ringed seals need good ice to establish territories with breathing holes, and snow cover enables females to dig out birth lairs. Thus, it was thought that the three zones would have different seal densities. The investigators took a stratified random sample of 20% of the 200 1-km^2 areas. The following tables give the number of plots and the number of plots sampled, in each zone:

Zone	Number of Plots	Plots Sampled
1	68	17
2	84	12
3	48	11
Total	200	40

In each sampled area, Imjak the Siberian husky tracked seal structures by scent; the number of breathing holes in the sampled square was recorded. A total of 199 breathing holes were located in zones 1, 2, and 3 altogether. The data (reconstructed from information given in the paper) are in the file *seals.dat*.

a Estimate the total number of breathing holes in the study region, along with its standard error.

Load the dataset, calculate the means for each zone:

```
library(plyr)
seals <- read.csv("~/Courses/STAT 607/STAT-607/data/Dataset/seals.csv")
ddply(seals, .(zone), colMeans)

##   zone  holes
## 1    1  1.765
## 2    2  4.417
## 3    3 10.545
```

To get the estimate of the total, multiply the mean by the number of sampling units in each strata, and sum:

```
sum(c(68, 84, 48) * ddply(seals, .(zone), colMeans)[[2]])

## [1] 997.2
```

To estimate the variance, we need the strata (sample) variances:

```
ddply(seals, .(zone), sapply, var)

##   zone  holes
## 1    0  3.316
## 2    0 11.538
## 3    0 46.073
```

Do the calculation:

```
N.h <- c(68, 84, 48)
n.h <- ddply(seals, .(zone), nrow)[[2]]
S.h.sq <- ddply(seals, .(zone), sapply, var)[[2]]
(V.hat_t.hat <- sum((1 - n.h/N.h) * N.h^2 * S.h.sq/n.h))

## [1] 13930

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

## [1] 118
```

b If you were designing the survey, how would you allocate observations to strata if the goal is to estimate the total number of breathing holes? If the goal is to compute the density of breathing holes in the three zones?

Exercise 3.15

Hayes (2000) took a stratified sample of New York City food stores. The sampling frame consisted of 1408 food stores with at least 4000 square feet of retail space. The population of stores was stratified into three strata using median houshold income within the zip code. The prices of a "market basket" of goods were determined for each store; the goal of the survey was to investigate whether prices differ among the three strata. Hayes used the logarithm of total price for the basket as the response y. Results are given in the following table:

Stratum, h	N_h	n_h	$\bar{y}_h$	s_h
1 Low income	190	21	3.925	0.037
2 Middle income	407	14	3.938	0.052
3 High income	811	22	3.942	0.070

a The planned sample size was 30 in each stratum; this was not achieved because some stores went out of business while the data were being collected. What are the advantages and disadvantages of sampling the same number of stores in each stratum?

b Estimate $\bar{y}_U$ for these data and give a 95% CI.

c Is there evidence that prices are different in the three strata?