Some Homework Problem Solutions – HW5

First I give detailed solutions for the two theoretical problems, which people had trouble with. After that, I include the brief Splus data steps supplied by the grader for the other problems.

6.16. You can ignore the $V(\hat{t}_i$ terms because they appear the same way on both sides of the equation (6.13). We showed in class that (in the case of single-stage sampling, with t_i observed whenever PSU i is sampled)

$$V(\hat{t}_{HT}) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij}-\pi_i\,\pi_j}{\pi_i\,\pi_j}\,t_i\,t_j$$

Now, separating out the terms with $i=j$ from those with $i>j$ and those with $i<j$, and using the fact that $\pi_{ii}=\pi_i,$ we get the last expression equal to

$$\sum_{i=1}^{N}\frac{\pi_i(1-\pi_i)}{\pi_i^2}\,t_i^2 \;+\; \sum_{1\le j<i\le N}\frac{\pi_{ij}-\pi_i\,\pi_j}{\pi_i\,\pi_j}\,t_i\,t_j \;+\; \sum_{1\le i<j\le N}\frac{\pi_{ij}-\pi_i\,\pi_j}{\pi_i\,\pi_j}\,t_i\,t_j$$

$$=\; \sum_{i=1}^{N}\Big(\frac{t_i}{\pi_i}\Big)^2\sum_{j:\,j\ne i}(\pi_i\pi_j-\pi_{ij}) \;-\; 2\sum_{1\le i<j\le N}\frac{\pi_i\,\pi_j-\pi_{ij}}{\pi_i\,\pi_j}\,t_i\,t_j \qquad (*)$$

where the first part of the last equality holds because, by (6.10)-(6.11),

$$\sum_{j:\,j\ne i}(\pi_i\pi_j-\pi_{ij}) \;=\; \pi_i\Big(\sum_{j=1}^{N}\pi_j-\pi_i\Big)-\sum_{j:\,j\ne i}\pi_{ij} \;=\; \pi_i(n-\pi_i)-(n-1)\pi_i \;=\; \pi_i(1-\pi_i)$$

Finally, equation (*) is equal to

$$\sum_{i=1}^{N}\Big(\frac{t_i}{\pi_i}\Big)^2 2\sum_{j:\,j>i}(\pi_i\pi_j-\pi_{ij}) \;-\; 2\sum_{1\le i<j\le N}\frac{\pi_i\,\pi_j-\pi_{ij}}{\pi_i\,\pi_j}\,t_i\,t_j \;=$$

$$\sum_{i=1}^{N}\Big(\frac{t_i}{\pi_i}\Big)^2\sum_{j:\,j>i}(\pi_i\pi_j-\pi_{ij})+\sum_{j=1}^{N}\Big(\frac{t_j}{\pi_j}\Big)^2\sum_{i:\,j>i}(\pi_i\pi_j-\pi_{ij})-2\sum_{1\le i<j\le N}\frac{\pi_i\,\pi_j-\pi_{ij}}{\pi_i\,\pi_j}\,t_i\,t_j$$

$$=\; \sum_{1\le i<j\le N}(\pi_i\pi_j-\pi_{ij})\Big(\frac{t_i}{\pi_i}\;-\;\frac{t_j}{\pi_j}\Big)^2$$

6.18. We saw in class that the joint-inclusion probabilities for a single-stage ($t_i=y_i$) stratified SRS sample of sizes n_h from strata $U_h,\ h=1,\ldots,H,$ of respective population sizes N_h, were

$$\pi_i \;=\; \frac{n_h}{N_h} \quad \text{for} \quad i\in U_h \quad , \qquad \pi_{ij} \;=\; \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)} & \text{if} \quad i,j\in U_h \\ \frac{n_h\,n_k}{N_h\,N_k} & \text{if} \quad i\in U_h,\ j\in U_k,\ h\ne k \end{cases}$$

The objective in this problem is to algebraically reduce the general Horvitz-Thompson variance formula with these inclusion probabilities to give the standard stratified-SRS variance formula we covered earlier. Following the same sorts of steps as we covered in class for SRS sampling via Horvitz-Thompson, and indexing all PSU's by (h,i) with $i \in U_h$, we find:

$$V(\hat{t}_{HT}) = \sum_{h=1}^{H}\sum_{k=1}^{H}\sum_{i\in U_h}\sum_{j\in U_k}\frac{N_hN_k}{n_hn_k}\Big(\frac{n_h}{N_h}I_{[h=k,\,i=j]} + \frac{n_h(n_h-1)}{N_h(N_h-1)}I_{[h=k,\,i\neq j]} + \frac{n_hn_k}{N_hN_k}I_{[h\neq k]} - \frac{n_hn_k}{N_hN_k}\Big)t_it_j$$

$$= \sum_{h=1}^{H}\sum_{k=1}^{H} I_{[h=k]}\frac{N_hN_k}{n_hn_k}\sum_{i\in U_h}\sum_{j\in U_k}\Big(\frac{n_h}{N_h}I_{[i=j]} + \frac{n_h(n_h-1)}{N_h(N_h-1)}I_{[i\neq j]} - \frac{n_hn_k}{N_hN_k}\Big)\,t_i\,t_j$$

$$= \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i,j\in U_h}\Big((1-\frac{n_h}{N_h})\,I_{[i=j]} - \frac{N_h-n_h}{N_h(N_h-1)}I_{[i\neq j]}\Big)\,t_i\,t_j$$

and a few more lines following exactly the steps we did in class (but now one stratum at a time) shows the last expression equal to the stratified-SRS-sampling variance

$$\sum_{h=1}^{H}\frac{N_h(N_h-n_h)}{n_h}\frac{1}{N_h-1}\Big(\sum_{i\in U_h}t_i^2 - (\sum_{i\in U_h}t_i)^2/N_h\Big)$$

DATA PROBLEMS --- SPLUS CODE AND NUMERICS

```
##Chapter 6##

#Problem 8

import.data(DataFrame="statepps",FileName="c:/LOHR/SPSS/statepps.por",
FileType= "SPSS_POR")
#(a)

prob <- statepps$LANDAREA/sum(statepps$LANDAREA)

c1<-matrix(1,51,1)
c2<-matrix(0,51,1)
for (i in 1:50)
        {
        c1[i+1]<-statepps$LANDAREA[i]+c1[i]
        c2[i]<-c1[i+1]-1
}
c2[51]<-sum(statepps$LANDAREA)
```

```
cumulateM<-cbind(c1,c2)
> cumulateM
         [,1]    [,2]
 [1,]       1   50750
 [2,]   50751  621124
 [3,]  621125  734766
 [4,]  734767  786841
 [5,]  786842  942814
 [6,]  942815 1046543
 [7,] 1046544 1051388
...
[47,] 3254597 3294194
[48,] 3294195 3360775
[49,] 3360776 3384862
[50,] 3384863 3439176
[51,] 3439177 3536281
n<-10
x<-c(1:3536281)
number<-sample(x,n,replace=T)
number
 [1] 3397518 1330229 1755842  949512 1115423  245036
     1334146 3462603 2837745 1415486

#leads to the sample {50,15,25,6,11,2,15,51,42,17}#####

nump<-c(50,15,25,6,11,2,15,51,42,17)
statename<-statepps$STATE[nump]
pip<-prob[nump]
Wisconsin   Indiana      Mississippi   Colorado     Georgia
0.01535907  0.01014342   0.01326648    0.02933279    0.01637851
Alaska       Indiana      Wyoming      South Dakota  Kansas
0.16129205   0.01014342   0.02745964   0.02146210    0.02313815

#(b)#

probp<-statepps$POPN/sum(statepps$POPN)

p1<-matrix(1,51,1)
p2<-matrix(0,51,1)
for (i in 1:50)
        {
        p1[i+1]<-statepps$POPN[i]+p1[i]
        p2[i]<-p1[i+1]-1

}
p2[51]<-sum(statepps$POPN)
```

```
cumulateMp<-cbind(p1/1000,p2/1000)
> cumulateMp
            [,1]       [,2]
 [1,]     0.0001   413.7511
 [2,]   413.7512   472.5277
 [3,]   472.5278   855.7645
 [4,]   855.7646  1095.1898
 [5,]  1095.1899  4184.7254
 [6,]  4184.7255  4531.1929
...
[46,] 23570.2297 23627.3630
[47,] 23627.3631 24266.8111
[48,] 24266.8112 24781.0857
[49,] 24781.0858 24961.9717
[50,] 24961.9718 25461.2381
[51,] 25461.2382 25507.7117

n<-10
x<-c(1:255077117)
numberp<-sample(x,n,replace=T)

numberp
[1]  4507 11071  2429 21157 14214  9890 19293 21702  4553 24342

##leads to the sample {6,22,5,43,31,19,38,44,7,48}############
numbi<-c(6,22,5,43,31,19,38,44,7,48)
statename<-statepps$STATE[numbi]
pii<-probp[numbi]
Colorado      Massachusetts  California    Tennessee     New Jersey
0.01358285    0.02349373     0.12112163    0.01970095    0.03065841
Louisiana     Oregon         Texas         Connecticut   Washington
0.01677488    0.01164968     0.06932232    0.01285539    0.02016153

#(c)#

#two samples are quite different since the same state has different
probabilities to be included in sample (a)
and (b). For sample (a) tend to choose states having large land
area (Alaska) and sample (b) tends to choose states having
large population (California).######

#Problem 12##

import.data(DataFrame="statepop",FileName="c:/LOHR/SPSS/statepop.por",
FileType= "SPSS_POR")
```

```
#a#
x<-statepop$POPN/sum(statepps$POPN)
plot(x,statepop$VETERANS,xlab='psi for county',ylab='veterans in county')
cor(x,statepop$VETERANS)
[1] 0.9871941
#There's strong linear relationship and we expect the unequal
probability sampling to be very efficient here.#

#b#
that<-(1/100)*(sum(statepop$VETERANS/x))
27914134
y<-(statepop$VETERANS/x-that)^2
se<-sqrt((1/100)*(sum(y)/99))
1087451

#c#
t<-(statepop$PERCVIET*statepop$VETERANS)/100
that<-(1/100)*(sum(t/x))
8050464
y2<-(t/x-that)^2
se<-sqrt((1/100)*(sum(y2)/99))
327336.7
```