Calculating estimates for cluster random samples

Here is some R code that caluclates two estimators when sampling from clusters.

Here we are assuming that the cluster sizes, the M_i's, vary and the sample sizes, the m_i's, vary as well. For estimating the pop total this calculates the unbiased estimator in (5.21) of the text and its unbiased estimate of variance in (5.25). For the population mean it finds the ratio estimate in (5.28) and its estimate of variance in (5.29). This is done in the context of a simple example which assumes that we have a random sample of 3 clusters from a population consisting of N=15 clusters. The total number of elements in the population is assumed to be unknown.

Note the equation numbers in the above refer to the first edition of the text. The corresponding numbers for the second edition are (5.20), (5.24), (5.27) and (5.28).

These commands should help you do HW assignment 5. Note for some of the bigger problems you will not want to look at all of X.

First we get the population into R.

```
> library("RCurl")
> x<-getURL("http://users.stat.umn.edu/~gmeeden/classes/5201/moredata/cluspop1.txt")
> X<-read.table(textConnection(x), header=T)
> names(X)

[1] "clmemb" "clsize" "yval"

> X

  clmemb clsize yval
1      1      5 12.1
2      1      5 14.3
3      2      8 11.1
4      2      8 13.3
5      2      8 10.4
6      3     10 13.2
7      3     10 14.7
8      3     10 15.1
9      3     10 15.2
```

Now we are ready to do the computations.

```
> yval<-X[,3]
> clsize<-X[,2]
> clmemb<-X[,1]
> foo<-split(yval,clmemb)
> foo
```

```
$`1`
[1] 12.1 14.3

$`2`
[1] 11.1 13.3 10.4

$`3`
[1] 13.2 14.7 15.1 15.2

> foo[[1]]

[1] 12.1 14.3

> nclus<-length(foo)
> nclus

[1] 3

> dum<-split(clsize,clmemb)
> mi<-sapply(dum,length)
> mi

1 2 3
2 3 4

> Mi<-sapply(dum,mean)
> Mi

 1  2  3
 5  8 10

> ssufpc<-1-mi/Mi
> clusmean<-sapply(foo,mean)
> clusvar<-sapply(foo,var)
> N<-15
> that<-(N/nclus)*sum(Mi*clusmean)
> ybarratio<-sum(Mi*clusmean)/sum(Mi)
> s2t<-var(Mi*clusmean)
> varterm2<-sum(ssufpc*(Mi^2)*clusvar/mi)
> s2r<-var(Mi*(clusmean - ybarratio))
> varthat<-N^2*(1-nclus/N)*s2t/nclus + (N/nclus)*varterm2
> varybarr<-( (1 - nclus/N)*s2r/nclus + varterm2/(nclus*N) )/(mean(Mi))^2
> list(that=that,varthat=varthat,yratio=ybarratio,varyratio=varybarr)

$that
[1] 1521.5

$varthat
```

```
[1] 98465.47

$yratio
[1] 13.23043

$varyratio
[1] 0.8042411
```

Here is a way to simplify some of the calculations for problem 12 a. To get an estimate of the variance of our estimator we need (using the terminology from anov) the (between SS)/(size of sample within a cluster) and the within SS. To see how this can be done consider a simple example where we have 2 clusters each with a sample of size 3. In the following y is the set of values for the character of interest and x tells us cluster they belong to. In gl(n,m,k) n is the number of clusters, m is the size of the sample within a cluster and $k = n*m$ is the total number of elements in the sample.

```
> y<-c(2,3,4,6,7,8)
> x<-gl(2,3,6)
> x

[1] 1 1 1 2 2 2
Levels: 1 2

> anova(lm(y~x))

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1     24      24      24 0.00805 **
Residuals  4      4       1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note the Sum Sq due to x in the above will be just the between SS and the residuals Sum Sq is the within SS. So in this example the numbers we need are 24/3 and 4.