NOTES ON SAMPLE SIZE CALCULATION UNDER STRATIFIED SAMPLING

In class we had a discussion of sample size calculation under stratified sampling. We discussed several scenarios. One of them was the case where pre-specified levels of precision are desired both for the within-stratum means $\bar{Y}_h$ and the overall population mean $\bar{Y}$. The purpose of these notes is to expand on this situation. The notation here follows that used in class, with some slight differences.

The problem is framed in terms of $100(1 - \alpha)\%$ confidence intervals. We want the half-width of the confidence interval for $\bar{Y}_h$ to be no greater than d_h and the half-width of the confidence interval for $\bar{Y}$ to be no greater than d. That is, we want

$$z_{1-\alpha/2}\sqrt{\mathrm{Var}(\bar{y}_h)} \leq d_h \quad \forall h, \tag{1}$$

$$z_{1-\alpha/2}\sqrt{\mathrm{Var}(\bar{y}_{st})} \leq d. \tag{2}$$

In other words, we want

$$\mathrm{Var}(\bar{y}_h) \leq V_h \quad \forall h, \tag{3}$$

$$\mathrm{Var}(\bar{y}_{st}) \leq V, \tag{4}$$

where

$$V_h = (d_h/z_{1-\alpha/2})^2, \tag{5}$$

$$V = (d/z_{1-\alpha/2})^2. \tag{6}$$

Recall that

$$\mathrm{Var}(\bar{y}_h) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2, \tag{7}$$

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 S_h^2. \tag{8}$$

Thus, the requirement is

$$\frac{S_h^2}{n_h} \leq V_h^* \quad \forall h, \tag{9}$$

$$\sum_{h=1}^{L} \frac{1}{n_h} W_h^2 S_h^2 \leq V^*, \tag{10}$$

where

$$V_h^* = V_h + \frac{S_h^2}{N_h}, \tag{11}$$

$$V^* = V + \sum_{h=1}^{L} \frac{1}{N_h} W_h^2 S_h^2. \tag{12}$$

In practice, as noted in class, we have to replace S_h^2 (which is unknown) with preliminary estimates $\widetilde{S}_h^2$. We redefine V_h^* and V^* to be the values obtained in (11) and (12) after this substitution, and we denote by $(9)'$ and $(10)'$ the versions of (9) and (10) that result after the substitution. We want to meet the conditions $(9)'$ and $(10)'$ with minimum cost, where cost is expressed as $\sum_h c_h n_h$.

We are thus dealing with the following minimization problem.

$$\text{minimize} \sum_{h=1}^{L} c_h n_h$$

subject to

$$\frac{\widetilde{S}_h^2}{n_h} \leq V_h^* \quad \forall h, \tag{13}$$

$$\sum_{h=1}^{L} \frac{1}{n_h} W_h^2 \widetilde{S}_h^2 \leq V^*. \tag{14}$$

Define $n_h^\circ = \widetilde{S}_h^2 / V_h$. These are the values of n_h that will satisfy the constraints (13) with exact equality. If these n_h values also satisfy the constraint (14), then we are done. Otherwise, we need to solve the above minimization problem using a computer routine for solving such problems. One such routine is the Matlab routine `fmincon`.

It is convenient to reformulate the problem. We define $\rho_h = 1/n_h$ and $\rho_h^\circ = 1/n_h^\circ$. We can then write the problem as follows.

$$\text{minimize} \sum_{h=1}^{L} \frac{c_h}{\rho_h}$$

subject to

$$0 \leq \rho_h \leq \rho_h^\circ \quad \forall h, \tag{15}$$

$$\sum_{h=1}^{L} W_h^2 S_h^2 \rho_h \leq V^*. \tag{16}$$

2

This gives us a minimization problem with a nonlinear objective function and linear constraints. We denote the optimal ρ_h values by $\rho_h^*, h = 1, \ldots, L$, and the optimal n_h values by $n_h^*, h = 1, \ldots, L$.

Below is a Matlab function to carry out the calculation. The required inputs to the function are as follows.

w = a vector containing the W_h values

$s2til$ = a vector containing the $\widetilde{S}_h^2$ values

$vhstar$ = a vector containing the V_h^* values

$vstar$ = the value of V^*

c = a vector containing the c_h values

For an explanation of the syntax of `fmincon`, see the accompanying document.

```
****************************************************************************

% main program
function [ ] = sam_all(w,s2til,vhstar,vstar,c);
global c;
format compact;
siz = size(w);
l = siz(2);
w
s2til
vhstar
vstar
c
rho_circ = (vhstar./s2til)';
rho_min = zeros(1,l)+0.00001;
a = (w.^2).*s2til;
aeq = [ ];
beq = [ ];
rho0 = rho_circ;
rho = fmincon(@cost,rho0,a,vstar,aeq,beq,rho_min,rho_circ);
ncirc = round(1./rho_circ)
rho_star = rho
nstar = round(1./rho_star)
opt_cost = cost(1./nstar)
% end of main program

% cost function
function [cstval] = cost(rho);
global c;
cstval = sum(c./rho');

****************************************************************************
```

The outputs are as follows (after echoing the inputs).

$$\texttt{ncirc} = \text{a vector containing the } n_h^{\circ} \text{ values}$$

$$\texttt{rho\_star} = \text{a vector containing the } \rho_h^* \text{ values}$$

$$\texttt{nstar} = \text{a vector containing the } n_h^* \text{ values}$$

$$\texttt{opt\_cost} = \text{the total cost at the optimal solution}$$

We now present an example. Consider a stratified survey with three strata. Suppose that:

$$W_1 = 0.4, W_2 = 0.4, W_3 = 0.2$$

$$\widetilde{S}_1^2 = 10, \widetilde{S}_2^2 = 12, \widetilde{S}_3^2 = 20$$

$$d_1 = d_2 = d_3 = 0.5, d = 0.2$$

$$c_1 = 2, c_2 = 3, c_3 = 6 \text{ (in dollars)}$$

Suppose further than the N_h's are all very large, so that the terms involving N_h in (11) and (12) can be ignored, and that the desired coverage level for the confidence intervals is 95%. The computation of the solution in Matlab proceeds as follows.

****

```
>> format compact
>> global c
>> w = [0.4 0.4 0.2];
>> s2til = [10 12 20];
>> dh = [0.5 0.5 0.5];
>> d = 0.2;
>> c = [2 3 6];
>> coverage = 0.95;
>> alpha = 1-coverage
alpha =
    0.0500
>> z = norminv(1-alpha./2)
z =
    1.9600
```

```
>> vhstar = (dh./z).^2
vhstar =
    0.0651    0.0651    0.0651
>> vstar = (d./z).^2
vstar =
    0.0104
>> sam_all(w,s2til,vhstar,vstar,c)
Warning: The value of local variables may have been changed to match the
        globals.  Future versions of MATLAB will require that you declare
        a variable to be global before you use that variable.
> In sam_all at 3
w =
    0.4000    0.4000    0.2000
s2til =
    10    12    20
vhstar =
    0.0651    0.0651    0.0651
vstar =
    0.0104
c =
     2     3     6
Warning: Large-scale (trust region) method does not currently solve this type
of problem, switching to medium-scale (line search).
> In fmincon at 260
  In sam_all at 18
Optimization terminated: magnitude of search direction less than 2*options.TolX
 and maximum constraint violation is less than options.TolCon.
Active inequalities (to within options.TolCon = 1e-006):
  lower      upper      ineqlin    ineqnonlin
               3          1
ncirc =
   154
   184
   307
rho_star =
    0.0021
    0.0023
    0.0033
nstar =
   480
   429
   307
opt_cost =
       4089


***********************************************************************
```

Thus, to achieve the within-stratum precision requirements, the required sample sizes are $n_1^\circ = 154, n_2^\circ = 184$, and $n_3^\circ = 307$. These sample sizes do not meet the precision requirement for the overall mean; larger sample sizes are needed to accomplish this. The minimum cost solution is to increase the sample sizes in Strata 1 and 2 to 480 and 429, respectively, while leaving the Stratum 3 sample size unchanged at 307. The reason why the Stratum 3 sample size is left unchanged is the relatively low frequency of Stratum 3 in the overall population combined with the relatively high cost of sampling from this stratum.

Using the theory presented in class, we can calculate the optimal sample sizes needed to meet the precision requirement for the overall mean alone, dropping the within-stratum requirements. You should go through the calculation as an exercise. The optimal allocation and optimal sample sizes are found to be as follows.

Stratum	Allocation	Sample Size
Stratum 1	0.4343	548
Stratum 2	0.3884	490
Stratum 3	0.1773	224

The cost of this sampling scheme is \$3,910, as compared with the \$4,089 required to meet both the overall and the within-stratum precision requirements.