7.6 Cluster Sampling with Unequal Cluster Sizes

- Suppose the N cluster sizes $M_1, M_2, \ldots, M_N$ are not all equal and that a one-stage cluster sample of n primary sampling units (PSUs) is taken with the goal of estimating t or $\overline{y}_U$.
- Let M_i and t_i $(i = 1, 2, \ldots, n)$ be the sizes and totals of the n sampled PSUs. Let $m = \sum_{i=1}^{n} M_i$ be the total number of secondary sampling units (SSUs) in the sample.
- We will review three methods of estimating $\overline{y}_U$ and t given the unequal cluster sizes. These methods are based on two representations of the population mean $\overline{y}_U$.

(i) $\overline{y}_U$ *as a population ratio*:

$$\overline{y}_U \;=\; \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i} \;=\; \frac{\sum_{i=1}^{N} t_i}{M_0} \tag{89}$$

expresses $\overline{y}_U$ as the ratio of the total of the primary sampling unit values to the total number of secondary sampling units.

(ii) $\overline{y}_U$ *as a mean cluster total*:

$$\overline{y}_U \;=\; \left(\frac{N}{M_0}\right) \sum_{i=1}^{N} \frac{t_i}{N} \tag{90}$$

expresses $\overline{y}_U$ as a multiple of the mean of the cluster totals.

Method 1: **The sample cluster ratio**: Suppose a SRS of clusters is selected without replacement. Substitution of sample values into (89) provides the following ratio estimator for $\overline{y}_U$:

$$\widehat{\overline{y}_{U}}_{c(1)} \;=\; \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} M_i} \;=\; \frac{\sum_{i=1}^{n} t_i}{m} \;=\; r_{clus}$$

which is the ratio of the sum of the sampled cluster totals to the sum of the sampled cluster sizes. Thus, the ratio-based estimator for t is

$$\widehat{t}_{c(1)} \;=\; M_0 \, \widehat{\overline{y}_{U}}_{c(1)} \;=\; M_0 \, r_{clus}$$

- $\widehat{\overline{y}_{U}}_{c(1)}$ is a special case of the SRS ratio estimator presented in Section 5 of the course notes (with $y_i = t_i$ and $x_i = M_i$). Thus, $\widehat{\overline{y}_{U}}_{c(1)}$ is biased with the bias $\rightarrow 0$ as n increases.
- There are no closed-forms for the true variances of $\widehat{\overline{y}_{U}}_{c(1)}$ and $\widehat{t}_{c(1)}$. However, approximations are given in Section 5 of the course notes.
- Estimators of the variances ($\widehat{V}(\widehat{t}_{c(1)})$ and $\widehat{V}(\widehat{\overline{y}_{U}}_{c(1)})$) are in Section 5 of the course notes.
- If M_0 is not known, $\widehat{V}(\widehat{\overline{y}_{U}}_{c(1)})$ can be estimated by replacing M_0 with the estimate $\widehat{M_0} \approx Nm/n$. Then dividing $\widehat{M_0}^2$ (instead of M_0^2 provides an estimate of $\widehat{V}(\widehat{\overline{y}_{U}}_{c(1)})$.
- Furthermore, when estimating t, multiply $\widehat{\overline{y}_{U}}_{c(1)}$ by $\widehat{M_0}$ and $\widehat{V}(\widehat{\overline{y}_{U}}_{c(1)})$ by $\widehat{M_0}^2$.

Figure 10: Cluster Sampling with Unequal-Sized Cluster

The mean $\overline{y}_U = 33.385$. There are $M_0 = 400$ secondary sampling units and $N = 49$ primary sampling units (clusters). There are 9 clusters of size $M_i = 16$, 24 clusters of size $M_i = 8$, and 16 clusters of size $M_i = 4$. The boldfaced values represent the SSUs in the sample.

18	20	15	20	20	15	19	18	**24**	**23**	**20**	**26**	29	28	28	31	31	34	28	32
13	20	16	20	15	23	19	26	**21**	**21**	**24**	**30**	23	26	25	33	31	28	32	38
16	18	20	24	25	26	22	23	**26**	**26**	**22**	**27**	25	25	34	28	37	36	38	31
17	17	16	22	21	23	22	27	**27**	**24**	**28**	**32**	29	33	27	37	37	38	35	33
15	**19**	**23**	**17**	21	23	21	23	24	25	31	26	32	34	32	33	**31**	**31**	36	37
21	**24**	**20**	**21**	28	26	30	22	31	25	29	29	27	30	29	37	**35**	**32**	38	43
23	**17**	**24**	**25**	24	27	31	29	31	34	27	36	29	29	34	39	**37**	**37**	40	36
18	**24**	**21**	**25**	27	22	32	32	31	26	28	34	34	37	35	34	**38**	**38**	37	40
22	26	28	26	24	29	33	26	27	27	34	31	39	32	36	38	**37**	**40**	44	43
23	27	28	29	26	32	25	31	35	34	32	33	37	32	42	40	**40**	**37**	42	44
23	21	31	23	30	27	31	30	32	35	30	40	32	37	37	36	**40**	**44**	44	40
26	29	31	26	30	31	34	36	30	38	36	32	38	38	37	42	**42**	**41**	40	49
28	24	28	27	26	31	32	29	32	33	38	34	39	38	40	37	41	43	**42**	**43**
32	25	31	32	29	29	35	38	38	32	36	35	39	42	39	40	44	42	**41**	**45**
27	29	35	28	35	35	31	40	35	37	38	44	40	40	47	39	49	48	51	49
30	29	32	32	33	30	36	38	42	36	35	38	44	47	45	49	41	43	44	51
28	35	35	34	**34**	**33**	**41**	**33**	34	35	39	44	**44**	**48**	44	50	49	48	**53**	**54**
29	33	32	36	**39**	**33**	**33**	**34**	35	42	46	47	**48**	**47**	46	45	44	52	**54**	**55**
28	37	38	37	33	33	34	37	45	40	39	42	42	46	47	48	52	47	46	53
38	39	39	37	34	38	39	45	39	42	45	41	44	51	46	50	52	51	51	53

M_i	t_i	$\overline{y}_i = t_i/M_i$	M_i	t_i	$\overline{y}_i = t_i/M_i$
16	401	25.0625	16	337	21.0625
8	279	34.8750	8	321	40.1250
8	280	35.0000	4	171	42.7500
4	187	46.7500	4	216	54.0000

$m = 68 \qquad \sum t_i = 2192 \qquad \overline{y} = 274 \qquad \overline{\overline{y}} = 37.453125$

Method 2: **The cluster sample total**: Suppose a SRS of clusters is selected without replacement. Substitution of sample values into (90) provides the following unbiased estimator for $\overline{y}_U$:

$$\widehat{\overline{y}_{Uc(2)}} = \frac{N}{M_0}\frac{\sum_{i=1}^n t_i}{n} = \frac{N}{nM_0}\sum_{i=1}^n t_i$$

- The variance $V(\widehat{\overline{y}_{Uc(2)}}) = \dfrac{(N-n)N}{n(N-1)M_0^2}\displaystyle\sum_{i=1}^N (t_i - \overline{t}_i)^2 = \dfrac{(N-n)N}{nM_0^2} S_t^2$

  where S_t^2 is the population variance of the t_i values.

- An estimate of this variance is given by

$$\widehat{V}(\widehat{\overline{y}_{Uc(2)}}) = \frac{(N-n)N}{n(n-1)M_0^2}\sum_{i=1}^n (t_i - \overline{y})^2 = \frac{(N-n)N}{nM_0^2} s_t^2. \tag{91}$$

  where s_t^2 is the sample variance of the sampled t_i values.

- To estimate t, multiply $\widehat{\overline{y}_{U}}_{c(2)}$ by M_0. For estimated variances, multiply $\widehat{V}(\widehat{\overline{y}_{U}}_{c(2)})$ by M_0^2.
- If $\underline{M_0 \text{ is not known}}$, we can substitute of $\widehat{M_0} = Nm/n$ into (91) and get:

$$\widehat{V}(\widehat{\overline{y}_{U}}_{c(2)}) \;=\; \frac{(N-n)n}{(n-1)Nm^2}\sum_{i=1}^{n}(t_i - \overline{y})^2 \;=\; \frac{(N-n)n}{Nm^2}\, s_t^2. \tag{92}$$

- For Methods 1 and 2, if a SRS of clusters is taken with replacement, then the estimator formulas for t and $\overline{y}_U$ remain unchanged, but the variance formulas need to be adjusted. Simply replace $N(N-n)$ with N^2 in the numerators of the estimated variance formulas.

Confidence Intervals

- A confidence interval for $\overline{y}_U$ using either Method 1 ($k=1$) or Method 2 ($k=2$) is:

$$\widehat{\overline{y}_{U}}_{c(k)} \;\pm\; t^*\sqrt{\widehat{V}(\widehat{\overline{y}_{U}}_{c(k)})} \qquad \text{for } k = a, b \tag{93}$$

  where t^* is the upper $\alpha/2$ critical value from the $t(n-1)$ distribution.

Method 3: **Primary sampling units selected with pps**:

Suppose that the primary sampling units (PSUs) are selected with replacement with draw-by-draw selection probabilities (p_i) proportional to the sizes of the PSUs, $p_i = M_i/M_0$.

One way to select the PSUs when each of M_i's (the sizes of the PSUs) is known is to

1. Generate N intervals $(0, M_1]$, $(M_1, M_1+M_2]$, $(M_1+M_2, M_1+M_2+M_3], \ldots (M_1+\cdots+M_{N-1},\ M_0]$.
2. Generate a random number U between 0 and M_0. Pick the interval that contains U. If this is the i^{th} interval, then select cluster i.
3. Repeat this n times.

Another way to construct the sampling design if each of the M_0 secondary sampling units can be listed in a sampling frame:

1. Select n SSUs (say, $u_1, u_2, \ldots, u_n$) from the M_0 in the population using simple random sampling with replacement. That is, select n numbers with replacement from $\{1, 2, \ldots, M_0$. Let u_i ($i = 1, 2, \ldots, n$) be the corresponding n secondary sampling units.
2. Then for each u_i ($i = 1, 2, \ldots, n$), sample all SSUs in the cluster containing u_i.

Thus, a PSU is selected every time any of its SSUs is selected.

- Now we can use either the Horvitz-Thompson estimator (Figure 11) or the Hansen-Hurwitz estimator (Figure 12), and their associated variance estimators discussed in Section 6 of the course notes.

Figure 11: Horvitz-Thompson Estimation with Selection Probabilities Proportional to Cluster Size

The mean $\overline{y}_U = 33.385$. There are $M_0 = 400$ secondary sampling units and $M_0 = 49$ primary sampling units (clusters). There are 9 clusters with $M_i = 16$, 24 clusters with $M_i = 8$, and 16 clusters with $M_i = 4$. Five clusters were sampled with replacement. One cluster was sampled twice. The boldfaced values are in the sample.

					Sampled ↓ twice ↓														
18	20	15	20	**20**	**15**	**19**	**18**	24	23	20	26	29	28	28	31	31	34	28	32
13	20	16	20	**15**	**23**	**19**	**26**	21	21	24	30	23	26	25	33	31	28	32	38
16	18	20	24	**25**	**26**	**22**	**23**	26	26	22	27	25	25	34	28	37	36	38	31
17	17	16	22	**21**	**23**	**22**	**27**	27	24	28	32	29	33	27	37	37	38	35	33
15	19	23	17	21	23	21	23	24	25	31	26	**32**	**34**	32	33	31	31	36	37
21	24	20	21	28	26	30	22	31	25	29	29	**27**	**30**	29	37	35	32	38	43
23	17	24	25	24	27	31	29	31	34	27	36	**29**	**29**	34	39	37	37	40	36
18	24	21	25	27	22	32	32	31	26	28	34	**34**	**37**	35	34	38	38	37	40
22	26	28	26	24	29	33	26	27	27	34	31	39	32	36	38	37	40	44	43
23	27	28	29	26	32	25	31	35	34	32	33	37	32	42	40	40	37	42	44
23	21	31	23	30	27	31	30	32	35	30	40	32	37	37	36	40	44	44	40
26	29	31	26	30	31	34	36	30	38	36	32	38	38	37	42	42	41	40	49
28	24	28	27	26	31	32	29	32	33	38	34	39	38	40	37	41	43	42	43
32	25	31	32	29	29	35	38	38	32	36	35	39	42	39	40	44	42	41	45
27	29	35	28	35	35	31	40	35	37	38	44	40	40	47	39	**49**	**48**	51	49
30	29	32	32	33	30	36	38	42	36	35	38	44	47	45	49	**41**	**43**	44	51
28	35	35	34	**34**	**33**	**41**	**33**	34	35	39	44	44	48	44	50	49	48	53	54
29	33	32	36	**39**	**33**	**33**	**34**	35	42	46	47	48	47	46	45	44	52	54	55
28	37	38	37	33	33	34	37	45	40	39	42	42	46	47	48	52	47	46	53
38	39	39	37	34	38	39	45	39	42	45	41	44	51	46	50	52	51	51	53

i	t_i	M_i	$p_i = M_i/M_0$	$\pi_i = 1-(1-p_i)^5$
1	344	16	16/400=.04	$1 - .96^5 = .184627302$
2	252	8	8/400=.02	$1 - .98^5 = .096079203$
3	278	8	8/400=.02	$1 - .98^5 = .096079203$
4	181	4	4/400=.01	$1 - .99^5 = .049009950$

$$
\begin{aligned}
\pi_{12} = \pi_{13} &= [1-(.96^5)] + [1-(.98^5)] - [1-(.94^5)] = .0146105270 \\
\pi_{14} &= [1-(.96^5)] + [1-(.99^5)] - [1-(.95^5)] = .00741819 \\
\pi_{24} = \pi_{34} &= [1-(.98^5)] + [1-(.99^5)] - [1-(.97^5)] = .003823179 \\
\pi_{23} &= [1-(.98^5)] + [1-(.98^5)] - [1-(.96^5)] = .007531104
\end{aligned}
$$

Figure 12: Hansen-Hurwitz Estimation with Selection Probabilities Proportional to Cluster Size

In Figure 11, the total abundance is $t = 13354$. There are $M_0 = 400$ secondary sampling units and $M_0 = 49$ primary sampling units (clusters). There are 9 clusters with $M_i = 16$, 24 clusters with $M_i = 8$, and 16 clusters with $M_i = 4$. The cluster totals t_i for the clusters in Figure 11 are summarized in the figure below. Also included is a cluster label (1 to 49). Eight clusters were sampled with replacement. The sampled units are 2, 6, 6, 16, 25, 30, 32, and 44. Note that cluster 6 was sampled twice. The boldfaced values are in the sample.

1	*2*	*3*	*10*	*11*	*12*	*13*
292	**(344)**	401	218	243	272	267
4	*5*	*6*	*14*	*15*	*16*	*17*
337	418	**((467))**	252	273	**(279)**	307
7	*8*	*9*	*18*	*19*	*20*	*21*
419	475	526	285	308	321	346
22	*26*	*30*	*34*	*35*	*36*	*37*
227	249	**(278)**	158	156	170	171
23	*27*	*31*	*38*	*39*	*40*	*41*
242	278	305	171	180	181	195
24	*28*	*32*	*42*	*43*	*44*	*45*
262	280	**(322)**	187	185	**(193)**	216
25	*29*	*33*	*46*	*47*	*48*	*49*
(293)	293	333	333	183	191	203

Unit i	t_i	p_i	t_i/p_i
2	344	.04	8600
6	467	.04	11675
6	467	.04	11675
16	279	.02	13950
25	293	.02	14650
30	278	.02	13900
32	322	.02	16100
44	193	.01	19300
			109850

7.6.1 Using R and SAS for estimation with unequal cluster sizes

R code for analysis of Figure 10 data: Methods 1 and 2

```
library(survey)
source("c:/courses/st446/rcode/confintt.r")

# Cluster sample with unequal-size clusters (Figure 10)

M0 = 400
N = 49
n =  8
y <- c(401,337,279,321,280,171,187,216)
clusterid <- c(1,2,3,4,5,6,7,8)

# Method 1: SRS of clusters using cluster ratios

fpc1 <- c(rep(N,n))
Mvec <- c(16,16,8,8,8,4,4,4)
ratio_ttl  <- M0*y
ratio_mn   <- y

Fig10 <- data.frame(cbind(fpc1,ratio_ttl,ratio_mn,Mvec))
Fig10

# Create the sampling design
dsgn10 <- svydesign(data=Fig10,id=~1,fpc=~fpc1)

# Method 1: Estimation of the population mean

estmean1 <- svyratio(~ratio_mn,~Mvec,design=dsgn10)
confint.t(estmean1,tdf=n-1,level=.95)

# Method 1: Estimation of the population total

esttotal1 <- svyratio(~ratio_ttl,~Mvec,design=dsgn10)
confint.t(esttotal1,tdf=n-1,level=.95)

# Method 2: SRS of clusters using cluster totals

fpc2 <- c(rep(N,8))
wgt2a <- c(rep(N/n,n))
wgt2b <- wgt2a/M0
#wgt2b <- c(rep(N/(n*M0),n))

Fig10a <- data.frame(cbind(clusterid,y,wgt2a,wgt2b,fpc2))
Fig10a
dsgn10a <-svydesign(ids=~clusterid,weights=~wgt2a,fpc=~fpc2,data=Fig10a)
dsgn10b <-svydesign(ids=~clusterid,weights=~wgt2b,fpc=~fpc2,data=Fig10a)

# Method 2: Estimation of population total

esttotal2 <- svytotal(~y,design=dsgn10a)
print(esttotal2,digits=15)
confint.t(esttotal2,level=.95,tdf=n-1)

# Method 2: Estimation of population mean

estmean2 <- svytotal(~y,design=dsgn10b)
print(estmean2,digits=15)
confint.t(estmean2,level=.95,tdf=n-1)
```

R output for analysis of Figure 10 data: Methods 1 and 2

```
# Method 1: SRS of clusters using cluster ratios

  fpc1 ratio_ttl ratio_mn Mvec
1   49    160400      401   16
2   49    134800      337   16
3   49    111600      279    8
4   49    128400      321    8
5   49    112000      280    8
6   49     68400      171    4
7   49     74800      187    4
8   49     86400      216    4

> # Method 1: Estimation of the population mean
-------------------------------------------------------------------
mean( ratio_mn/Mvec ) = 32.23529
SE( ratio_mn/Mvec ) = 3.60282

Two-Tailed CI for ratio_mn/Mvec where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  23.71598     40.75460
-------------------------------------------------------------------

> # Method 1: Estimation of the population total
-------------------------------------------------------------------
mean( ratio_ttl/Mvec ) = 12894.11765
SE( ratio_ttl/Mvec ) = 1441.12717

Two-Tailed CI for ratio_ttl/Mvec where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  9486.39340     16301.84189
-------------------------------------------------------------------

> # Method 2: SRS of clusters using cluster totals

  clusterid   y wgt2a     wgt2b fpc2
1         1 401 6.125 0.0153125   49
2         2 337 6.125 0.0153125   49
3         3 279 6.125 0.0153125   49
4         4 321 6.125 0.0153125   49
5         5 280 6.125 0.0153125   49
6         6 171 6.125 0.0153125   49
7         7 187 6.125 0.0153125   49
8         8 216 6.125 0.0153125   49

> # Method 2: Estimation of population total
-------------------------------------------------------------------
mean( y ) = 13426.00000
SE( y ) = 1255.09810

Two-Tailed CI for y where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  10458.16459     16393.83541
-------------------------------------------------------------------

> # Method 2: Estimation of population mean
-------------------------------------------------------------------
mean( y ) = 33.56500
SE( y ) = 3.13775

Two-Tailed CI for y where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  26.14541     40.98459
-------------------------------------------------------------------
```

SAS code for analysis of Figure 10 data (Supplemental)

```
data in;
   NN = 49;
   M0 = 400;
   n =  8;
input _cluster Mi y @@;
     t = M0*y;
datalines;
 1 16 401   2 16 337   3  8 279   4  8 321
 5  8 280   6  4 171   7  4 187   8  4 216
;
TITLE 'SRS of clusters with replacement-- Figure 10';

PROC SURVEYMEANS DATA=in TOTAL=49 ;
   VAR y;
   RATIO y/Mi;
   RATIO t/mi;
TITLE2 'The sample cluster ratio method: Method 1';

DATA in; SET in;
     ytotal = y*NN;
     ymean  = ytotal/M0;
     wgt = 1/M0;

PROC SURVEYMEANS DATA=in TOTAL=49 ;
   WEIGHT wgt;
   VAR ytotal ymean;
TITLE2 'The cluster sample total method: Method 2';
run;
```

SAS output for analysis of Figure 10 data

```
SRS of clusters with replacement-- Figure 10
The sample cluster ratio method: Method 1

The SURVEYMEANS Procedure
            Data Summary
Number of Observations             8

                               Statistics

                                                  Std Error
Variable             N           Mean           of Mean     95% CL for Mean
---------------------------------------------------------------------------
y                    8    274.000000      25.614247    213.4319    334.568
Mi                   8      8.500000       1.612406      4.6873     12.313
t                    8        109600          10246  85372.7721 133827.228
---------------------------------------------------------------------------

                          Ratio Analysis
Numerator Denominator              N           Ratio          Std Err
--------------------------------------------------------------------
y         Mi                       8       32.235294         3.602818
--------------------------------------------------------------------

Numerator Denominator     95% CL for Ratio
-----------------------------------------------
y         Mi         23.7159835    40.7546047
-----------------------------------------------

                          Ratio Analysis
Numerator Denominator              N           Ratio          Std Err
--------------------------------------------------------------------
t         Mi                       8           12894      1441.127165
--------------------------------------------------------------------

Numerator Denominator     95% CL for Ratio
-----------------------------------------------
t         Mi         9486.39340    16301.8419
-----------------------------------------------
```

```
SRS of clusters with replacement-- Figure 10
The cluster sample total method: Method 2

The SURVEYMEANS Procedure
            Data Summary
Number of Observations             8
Sum of Weights                  0.02

                               Statistics
                                          Std Error
Variable             N            Mean      of Mean     95% CL for Mean
--------------------------------------------------------------------------
ytotal               8           13426  1255.098104  10458.1646 16393.8354
ymean                8       33.565000     3.137745     26.1454    40.9846
--------------------------------------------------------------------------
```

Using R and SAS for Hansen-Hurwitz Estimation (Method 3) for Figure 12 data

R code for Hansen-Hurwitz Estimation for Figure 12 data

```
library(survey)
source("c:/courses/st446/rcode/confintt.r")

# Cluster sample with unequal-size clusters (Figure 12)

M0 = 400
N = 49
n =  8
y <- c(344,467,467,279,293,278,322,193)
Mvec <- c(16,16,16,8,8,8,8,4)

# Method 3: Sampling proportional to size with replacement

p <- Mvec/M0
t  <- y/p
t2 <- t/M0

Fig10c <- data.frame(t,t2)
Fig10c
dsgn10c <-svydesign(ids=~1,data=Fig10c)

# Method 3: Estimation of population total

esttotal3 <- svymean(~t,design=dsgn10c)
print(esttotal3,digits=15)
confint.t(esttotal3,level=.95,tdf=n-1)

# Method 3: Estimation of population mean

estmean3 <- svymean(~t2,design=dsgn10c)
print(estmean3,digits=15)
confint.t(estmean3,level=.95,tdf=n-1)
```

R output for Hansen-Hurwitz Estimation for Figure 12 data

```
> # Cluster sample with unequal-size clusters (Figure 12)

> # Method 3: Sampling proportional to size with replacement

      t      t2
1  8600 21.5000
2 11675 29.1875
3 11675 29.1875
4 13950 34.8750
5 14650 36.6250
6 13900 34.7500
7 16100 40.2500
8 19300 48.2500

> # Method 3: Estimation of population total
-------------------------------------------------------------------
mean( t ) = 13731.25000
SE( t ) = 1136.47668

Two-Tailed CI for t where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  11043.90968     16418.59032
-------------------------------------------------------------------

> # Method 3: Estimation of population mean
-------------------------------------------------------------------
mean( t2 ) = 34.32812
SE( t2 ) = 2.84119

Two-Tailed CI for t2 where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  27.60977     41.04648
-------------------------------------------------------------------
```

SAS code for Hansen-Hurwitz Estimation for Figure 12 data (supplemental)

```
data in;
   M0 = 400;
input _cluster Mi y @@;
   p = Mi/M0;    t = y/p;   ymean = t/M0;
datalines;
1 16 344   2 16 467   3 16 467   4  8 279
5  8 293   6  8 278   7  8 322   8  4 193
;
PROC SURVEYMEANS DATA=in MEAN CLM ;
   VAR t ymean;
TITLE 'PPS sampling of clusters with replacement-- Figure 12 -- Method 3';
run;
```

SAS output for Hansen-Hurwitz Estimation for Figure 12 data (supplemental)

```
PPS sampling of clusters with replacement-- Figure 12 -- Method 3

The SURVEYMEANS Procedure
            Data Summary
Number of Observations            8

                           Statistics

                               Std Error
Variable             Mean        of Mean       95% CL for Mean
---------------------------------------------------------------
t                   13731     1136.476679    11043.9097 16418.5903
ymean           34.328125        2.841192       27.6098    41.0465
---------------------------------------------------------------
```

7.7 Attribute Proportion Estimation using Cluster Sampling

- Instead of studying a quantitative measure associated with sampling units, we often are interested in an attribute (a qualitative characteristic). Statistically, the goal is to estimate a proportion. The **population proportion** p is the proportion of population units having that attribute.

- Examples: the proportion of females (or males) in an animal population, the proportion of consumers who own motorcycles, the proportion of married couples with at least 1 child...

- If a one-stage cluster sample is taken, then how do we estimate p?

7.7.1 Estimating p with Equal Cluster Sizes

- Statistically, we use an indicator function that assigns a y_{ij} value to secondary sampling unit j in primary sampling unit (cluster) i as follows:

$$\begin{aligned} y_{ij} &= 1 \quad \text{if unit } j \text{ in cluster } i \text{ possesses the attribute} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Then $t = \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$ = the total number of SSUs in the population that possess the attribute. By definition, the population proportion p is

$$p = \frac{t}{MN} = \frac{t}{M_0}$$

where $M_i = M$ for each cluster, and the proportion for cluster i is

$$p_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}.$$

- By taking a one-stage cluster sample of n equal-sized clusters, we can estimate p as the weighted average of the sampled cluster proportions:

$$\widehat{p}_c = \frac{\sum_{i=1}^{n} p_i}{n}.$$

- $\widehat{p}_c$ is an unbiased estimator of p, and the variance of $\widehat{p}_c$ is

$$V(\widehat{p}_c) = \left(\frac{N-n}{nN}\right)\sum_{i=1}^{N}\frac{(p_i - p)^2}{N-1} = \left(\frac{1-f}{n}\right)\sum_{i=1}^{N}\frac{(p_i - p)^2}{N-1} \tag{94}$$

where $f = n/N$ = the proportion of clusters sampled.

- Because p is unknown, we use $\widehat{p}_c$ as an estimate of p to get the unbiased estimator of $V(\widehat{p}_c)$:

$$\widehat{V}(\widehat{p}_c) = \left(\frac{N-n}{nN}\right)\sum_{i=1}^{n}\frac{(p_i - \widehat{p}_c)^2}{n-1} = \left(\frac{1-f}{n}\right)\sum_{i=1}^{n}\frac{(p_i - \widehat{p}_c)^2}{n-1} \tag{95}$$

7.7.2 Estimating p with Unequal Cluster Sizes

- Suppose the cluster sizes are not all equal. Let M_i be the number of secondary sampling units (SSUs) in cluster i and $t_i = \sum_{j=1}^{M_i} y_{ij}$ = the cluster i total.

- By taking a one-stage cluster sample of n clusters from a population with unequal-sized clusters, we estimate p as the proportion of sampled SSUs that possess the attribute:

$$\widehat{p}_c \;=\; \frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n M_i}.$$

- Note that $\widehat{p}_c$ is a ratio estimator. Therefore, it is a biased estimator. The bias, however, tends to be small for large $\displaystyle\sum_{i=1}^n M_i$.

- The variance $V(\widehat{p}_c)$ is approximated by:

$$V(\widehat{p}_c) \;\approx\; \left(\frac{1-f}{n\overline{M_0}^2}\right) \frac{\sum_{i=1}^N (t_i - pM_i)^2}{N-1} \tag{96}$$

where $\overline{M_0} = \displaystyle\sum_{i=1}^N M_i/N$ = the average number of elements per cluster in the population.

- Because p is unknown, we use $\widehat{p}_c$ as an estimate to get the unbiased estimator of $V(\widehat{p}_c)$:

$$\widehat{V}(\widehat{p}_c) \;\approx\; \left(\frac{1-f}{n\overline{m}^2}\right) \frac{\sum_{i=1}^n (t_i - \widehat{p}_c M_i)^2}{n-1}$$

$$= \left(\frac{1-f}{n\overline{m}^2}\right) \frac{\sum_{i=1}^n t_i^2 - 2\widehat{p}_c \sum_{i=1}^n t_i M_i + \widehat{p}_c^2 \sum_{i=1}^n M_i^2}{n-1} \tag{97}$$

where $\overline{m} = \displaystyle\sum_{i=1}^n M_i/n$ = the average number of elements per cluster in the sample.

Additional References

- Bellhouse, D.R. (1988) Systematic sampling. *Handbook of Statistics, Vol. 6 (Sampling)*. 125-145. Eds: Krishnaiah and Rao. Elsevier Science Publishers. Amsterdam.

- Murthy, M.N. and Rao, T.J. (1988) Systematic sampling with illustrative examples. *Handbook of Statistics, Vol. 6 (Sampling)*. 147-185. Eds: Krishnaiah and Rao. Elsevier Science Publishers. Amsterdam.

- Wolter, K.M. (1984) An investigation of some estimators of variance for systematic sampling. *J. of the American Statistical Association.* **79** 781-790.

Example of cluster sampling of attributes with unequal size clusters:

A simple random sample of $n = 30$ households (clusters) was drawn from a health district in Baltimore (USA) that contains $N = 15,000$ households. Using the following data, estimate the proportion p of people in this health district that visited a doctor last year.

Household Number	Household Size (M_i)	Number who visited doctor last year (t_i)	Household Number	Household Size (M_i)	Number who visited doctor last year (t_i)
1	5	5	16	6	0
2	3	2	17	3	3
3	2	0	18	3	0
4	3	0	19	3	0
5	4	0	20	4	0
6	3	0	21	2	0
7	7	0	22	4	4
8	3	1	23	5	2
9	4	0	24	4	0
10	3	1	25	3	3
11	4	2	26	3	0
12	3	0	27	1	0
13	2	2	28	4	2
14	3	0	29	4	2
15	2	0	30	4	1
			Totals	104	30

7.7.3 Using R and SAS for proportion estimation with unequal-sized clusters

R code for proportion estimation with unequal-sized clusters

```
library(survey)
source("c:/courses/st446/rcode/confintt.r")

# Cluster sample with unequal-size clusters - proportion estimation)

N = 15000
n = 30

Mvec <- c(5,6,3,3,2,3,3,3,4,4,3,2,7,4,3,5,4,4,3,3,4,3,3,1,2,4,3,4,2,4)
y <- c(5,0,2,3,0,0,0,0,0,0,0,0,0,4,1,2,0,0,1,3,2,0,0,0,2,2,0,2,0,1)
fpc <- c(rep(N,n))

propest <- data.frame(cbind(fpc,y,Mvec))

# Create the sampling design
dsgn <- svydesign(data=propest,id=~1,fpc=~fpc)

estmean1 <- svyratio(~y,~Mvec,design=dsgn)
confint.t(estmean1,tdf=n-1,level=.95)
```

R output for proportion estimation with unequal-sized clusters

```
----------------------------------------------------------------------
mean( y/Mvec ) = 0.28846
SE( y/Mvec ) = 0.07208
Two-Tailed CI for y/Mvec where alpha = 0.05 with 29 df
    2.5 %        97.5 %
  0.14105     0.43587
----------------------------------------------------------------------
```

SAS code for proportion estimation with unequal-sized clusters (supplemental)

```
data in;
    n =  30;
   NN = 15000;
    m = 104;
input _cluster Mi y @@;
datalines;
 1 5 5   2 6 0   3 3 2   4 3 3   5 2 0   6 3 0
 7 3 0   8 3 0   9 4 0  10 4 0  11 3 0  12 2 0
13 7 0  14 4 4  15 3 1  16 5 2  17 4 0  18 4 0
19 3 1  20 3 3  21 4 2  22 3 0  23 3 0  24 1 0
25 2 2  26 4 2  27 3 0  28 4 2  29 2 0  30 4 1
;
PROC SURVEYMEANS DATA=in TOTAL = 15000 MEAN CLM ;
   VAR y;
   RATIO y/Mi;
TITLE 'SRS of clusters without replacement-- Estimating p for household
data';
run;
```

SAS output for proportion estimation with unequal-sized clusters (supplemental)

```
SRS of clusters without replacement-- Estimating p for household data

The SURVEYMEANS Procedure
                          Statistics

                             Std Error
Variable           Mean        of Mean      95% CL for Mean
-------------------------------------------------------------
y              1.000000       0.253454   0.48162776 1.51837224
Mi             3.466667       0.223297   3.00997205 3.92336128
-------------------------------------------------------------


                  Ratio Analysis

Numerator Denominator        Ratio         Std Err
-------------------------------------------------
y         Mi              0.288462        0.072076 <-- for proportion p
-------------------------------------------------

Numerator Denominator     95% CL for Ratio
---------------------------------------------
y         Mi          0.14104945    0.43587362     <-- for proportion p
---------------------------------------------
```