3 STRATIFIED SIMPLE RANDOM SAMPLING

- Suppose the population is partitioned into disjoint sets of sampling units called **strata**. If a sample is selected within each stratum, then this sampling procedure is known as **stratified sampling**.

- If we can assume the strata are sampled independently across strata, then

  (i) the estimator of t or $\overline{y}_U$ can be found by combining stratum sample sums or means using appropriate weights

  (ii) the variances of estimators associated with the individual strata can be summed to obtain the variance an estimator associated with the whole population. (Given independence, the variance of a sum equals the sum of the individual variances.)

- (ii) implies that only within-stratum variances contribute to the variance of an estimator. Thus, the basic motivating principle behind using stratification to produce an estimator with small variance is to partition the population so that units within each stratum are as similar as possible. This is known as the **stratification principle**.

- In ecological studies, it is common to stratify a geographical region into subregions that are similar with respect to a known variable such as elevation, animal habitat type, vegetation types, etc. because it is suspected that the y-values may vary greatly across strata while they will tend to be similar within each stratum. Analogously, when sampling people, it is common to stratify on variables such as gender, age groups, income levels, education levels, marital status, etc.

- Sometimes strata are formed based on sampling convenience. For example, suppose a large study region appears to be homogeneous (that is, there are no spatial patterns) and is stratified based on the geographical proximity of sampling units. Taking a stratified sample ensures the sample is spread throughout the study region. It may not, however, lead to any significant reduction in the variance of an estimator.

- But, if the y-values are spatially correlated (y values tend to be similar for neighboring units), geographically determined strata can improve estimation of population parameters.

Notation: H = the number of strata

N_h = number of population units in stratum h $\quad h = 1, 2, \ldots, H$

$N = \sum_{h=1}^{H} N_h$ = the number of units in the population

n_h = number of sampled units in stratum h $\quad h = 1, 2, \ldots, H$

$n = \sum_{h=1}^{H} n_h$ = the total number of units sampled

y_{hj} = the y-value associated with unit j in stratum h

$\overline{y}_h$ = the sample mean for stratum h

$$t_h = \sum_{j=1}^{N_h} y_{hj} = \text{stratum } h \text{ total} \qquad t = \sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H} t_h = \text{the population total}$$

$$\overline{y}_{hU} = \frac{t_h}{N_h} = \text{stratum } h \text{ mean} \qquad \overline{y}_U = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} = \frac{t}{N} = \text{the population mean}$$

- If a simple random sample (SRS) is taken within each stratum, then the sampling design is called **stratified simple random sampling**.

- For stratum h, there are $\binom{N_h}{n_h}$ possible SRSs of size n_h. Therefore, there are $\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_H}{n_H}$ possible stratified SRSs for specified stratum sample sizes $n_1, \cdots, n_H$.

- If $\mathcal{S}_{strat}$ is a stratified SRS, then the probability of selecting $\mathcal{S}_{strat}$ is

$$P(\mathcal{S}_{strat}) \;=\; \prod_{h=1}^{H} \frac{1}{\binom{N_h}{n_h}} \;=\; \frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_H}{n_H}}$$

- Thus, every possible stratified SRS having stratum sample sizes $n_1, \cdots, n_H$ has the same probability of being selected.

3.1 Estimation of $\overline{y}_U$ and t

- Because a SRS was taken within each stratum, we can apply the estimator formulas for simple random sampling to each stratum. We can estimate each stratum population mean $\overline{y}_{hU}$ and each stratum population total t_h. The formulas are:

$$\widehat{\overline{y}_{hU}} = \overline{y}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} y_{hj} \qquad\qquad \widehat{t}_h = N_h\overline{y}_h = \tag{24}$$

- Because each $\widehat{t}_h$ is an unbiased estimator of the stratum total t_h for $i = 1, 2, \ldots, k$, their sum will be an unbiased estimator of the population total t. That is,

$$\widehat{t}_{str} \;=$$

is an unbiased estimator of t. An unbiased estimator of $\overline{y}_U$ is a weighted average of the stratum sample means

$$\widehat{\overline{y}_{U}}_{str} \;=\; \frac{\widehat{t}_{str}}{N} \;=\; \frac{1}{N}\sum_{h=1}^{H} N_h\overline{y}_h \quad \text{or, equivalently,} \quad \widehat{\overline{y}_{U}}_{str} \;=$$

where   is the weighting factor for stratum h.

- Before we can study $V(\widehat{t}_{str})$ and $V(\widehat{\overline{y}_{U}}_{str})$, we need to look at the within-stratum variances.

- Because a SRS is taken within stratum h, we can apply the results for simple random sampling estimators to each stratum. The variances of the stratified SRS estimators of the mean and total are:

$$V(\widehat{\overline{y}_{U}}_{h}) \;= \qquad\qquad\qquad\qquad V(\widehat{t}_h) = \tag{25}$$

where $S_h^2 = \dfrac{1}{N_h - 1}\displaystyle\sum_{j=1}^{N_h}(y_{hj} - \overline{y}_{hU})^2$ is the finite population variance <u>for stratum h</u>.

- Because the simple random samples are independent across the strata, the variance of $\widehat{t}_{str}$ is the sum of the individual stratum variances:

$$V(\widehat{t}_{str}) \;=\; \sum_{i=1}^{H} V(\widehat{t}_h) \;=\; \sum_{i=1}^{H} \tag{26}$$

- Dividing by N^2, gives the $V(\widehat{\overline{y}_U}_{str})$:

$$V(\widehat{\overline{y}_U}_{str}) \;=\; \left(\frac{1}{N^2}\right) V(\widehat{t}_{str}) \;=\; \left(\frac{1}{N^2}\right) \sum_{i=1}^{H} N_h(N_h - n_h)\frac{S_h^2}{n_h} \tag{27}$$

- Because S_h^2 is unknown, we use s_h^2 to get an unbiased estimator of $V(\widehat{t}_h)$:

$$\widehat{V}(\widehat{t}_h) = \tag{28}$$

  where s_h^2 is the sample variance of the n_h y-values sampled from stratum h.

- Substitution of (28) into (26) and (27) produce the estimated variances of the stratified SRS estimators:

$$\widehat{V}(\widehat{t}_{str}) \;=\; \sum_{h=1}^{H} N_h(N_h - n_h)\frac{s_h^2}{n_h} \qquad\qquad \widehat{V}(\widehat{\overline{y}_U}_{str}) \;=\; \left(\frac{1}{N^2}\right) \sum_{h=1}^{H} N_h(N_h - n_h)\frac{s_h^2}{n_h} \tag{29}$$

- Taking a square root of $\widehat{V}(\widehat{t}_{str})$ or $\widehat{V}(\widehat{\overline{y}_U}_{str})$ yields the corresponding **standard error**. This will be used when generating confidence intervals for t or $\overline{y}_U$.

- For the estimated variances of the estimators given in (29), we are assuming that all $n_h > 1$ (because s_h^2 is undefined for $n_h = 1$). Cochran (1977 pages 138-140) discusses two potential methods of dealing with the extreme case where all $n_h = 1$.

Stratification Example with Strong Spatial Correlation

- Abundance counts for the population in Figures 5a and 5b show a strong diagonal spatial correlation. The region has been gridded into a 20×20 grid of 10 m ×10 m quadrats. The total abundance $t = 13354$. This population was stratified in two different ways:

  (i) Into the four 10×10 strata shown in Figure 5a.

  Stratum sizes are $N_h = 100$ and stratum sample sizes are $n_h = 5$ for $h = 1, 2, 3, 4$.

  Stratum sample totals $\sum_{j=1}^{n_h} y_{hj}$ are 124, 158, 172, and 223 for $h = 1, 2, 3, 4$.

  Stratum sample means $\overline{y}_h$ are 24.8, 31.6, 34.4, and 44.6 for $h = 1, 2, 3, 4$.

  Stratum sample variances are $s_1^2 = 21.7$, $s_2^2 = 13.3$, $s_3^2 = 45.3$, and $s_4^2 = 41.3$.

(ii) Into seven unequal size diagonally-oriented strata shown in Figure 5b.

Stratum sizes are $N_1 = N_7 = 45$, $N_2 = N_6 = 60$, $N_3 = N_5 = 66$, and $N_4 = 58$.

Stratum sample sizes are $n_1 = n_7 = 3$, $n_2 = n_3 = n_5 = n_6 = 5$, and $n_4 = 4$.

Stratum sample totals $\sum_{j=1}^{n_h} y_{hj}$ are 65, 122, 153, 143, 178, 203, and 143 for $h = 1, 2, 3, 4, 5, 6, 7$, respectively.

Stratum sample means $\overline{y}_h$ are $21.\overline{6}$, 24.4, 30.6, 35.75, 35.6, 40.6, and $47.\overline{6}$ for $h = 1, 2, 3, 4, 5, 6, 7$, respectively.

Stratum sample variances are $s_1^2 = 10.\overline{3}$, $s_2^2 = 14.8$, $s_3^2 = 19.3$, $s_4^2 = 4.25$, $s_5^2 = 8.3$, $s_6^2 = 10.8$, and $s_7^2 = 26.\overline{3}$, respectively.

- For the stratified SRSs in Figure 5a and Figure 5b:
  - Calculate $\widehat{t}_{str}$, $\widehat{\overline{y}}_{U\,str}$, and their standard errors.
  - Calculate 95% confidence intervals for t and $\overline{y}_U$.

3.1.1 Confidence Intervals for $\overline{y}_U$ and t

- If all of the stratum sample sizes n_h are sufficiently large (Thompson suggests $n_h \geq 30$), approximate $100(1-\alpha)\%$ confidence intervals for $\overline{y}_U$ and t are

$$\widehat{\overline{y}}_{U\,str} \pm z^* \sqrt{\widehat{V}(\widehat{\overline{y}}_{U\,str})} \qquad\qquad \widehat{t}_{str} \pm z^* \sqrt{\widehat{V}(\widehat{t}_{str})} \tag{30}$$

where z^* is the upper $\alpha/2$ critical value from the standard normal distribution.

- For smaller sample sizes, the following confidence intervals have been recommended:

$$\widehat{\overline{y}}_{U\,str} \pm t^* \sqrt{\widehat{V}(\widehat{\overline{y}}_{U\,str})} \qquad\qquad \widehat{t}_{str} \pm t^* \sqrt{\widehat{V}(\widehat{t}_{str})} \tag{31}$$

where t^* is the upper $\alpha/2$ critical value from the $t(d)$ distribution. In this case, d is Satterthwaite's (1946) approximate degrees of freedom d where

$$d \;=\; \frac{\left(\sum_{h=1}^{H} a_h s_h^2\right)^2}{\sum_{h=1}^{H} (a_h s_h^2)^2/(n_h-1)} \;=\; \frac{(\widehat{V}(\widehat{t}_{str}))^2}{\sum_{h=1}^{H} (a_h s_h^2)^2/(n_h-1)} \tag{32}$$

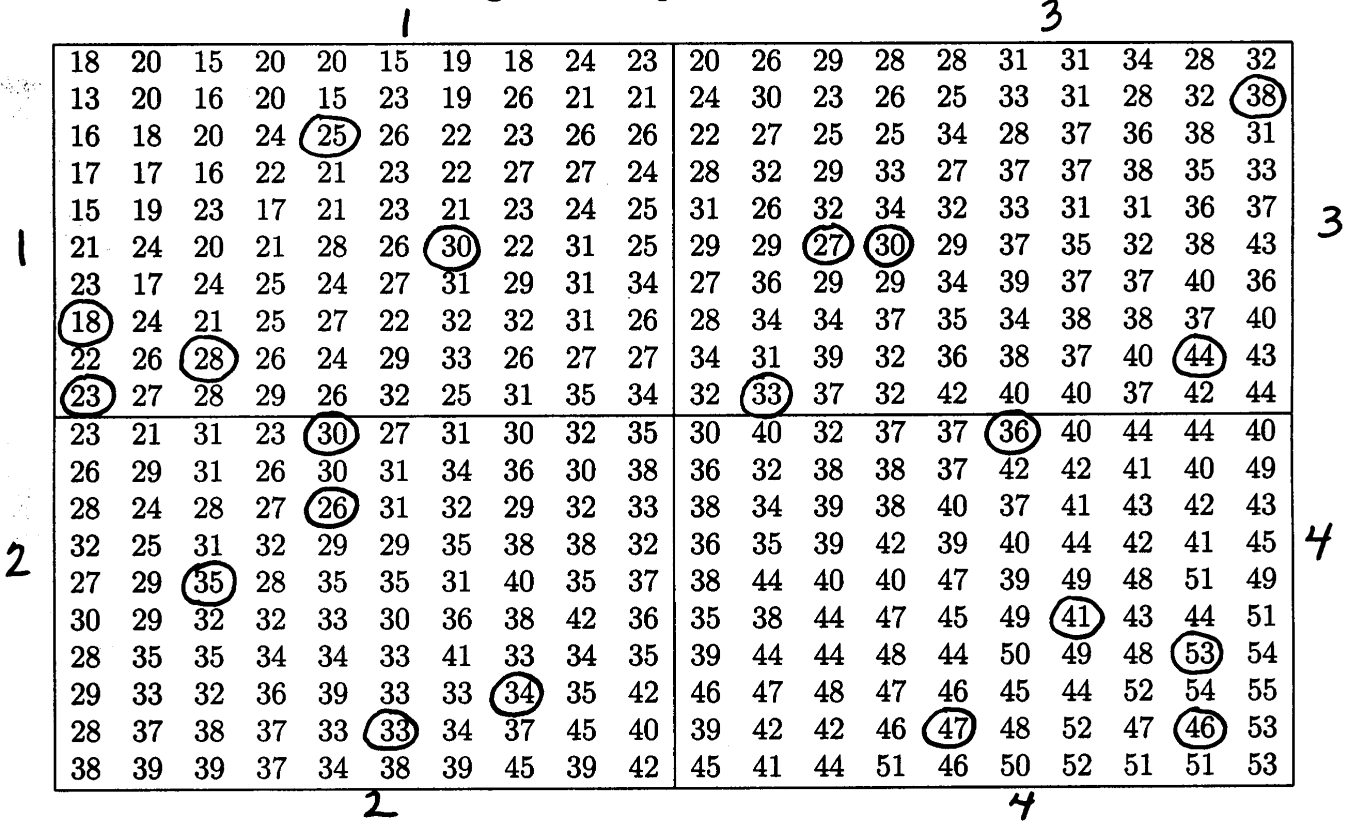
where $a_h = N_h(N_h - n_h)/n_h$.

- Lohr (page 79) mentions that some software packages will use $n - H$ degrees of freedom (instead of the approximate degrees of freedom). Both R and SAS use $n - H$ as the default degrees of freedom.

- If the stratum sample sizes n_h are all equal and the stratum sizes N_h are all equal, then the degrees of freedom reduces to $d = n - H$ where $n = \sum n_h$ is the total sample size.

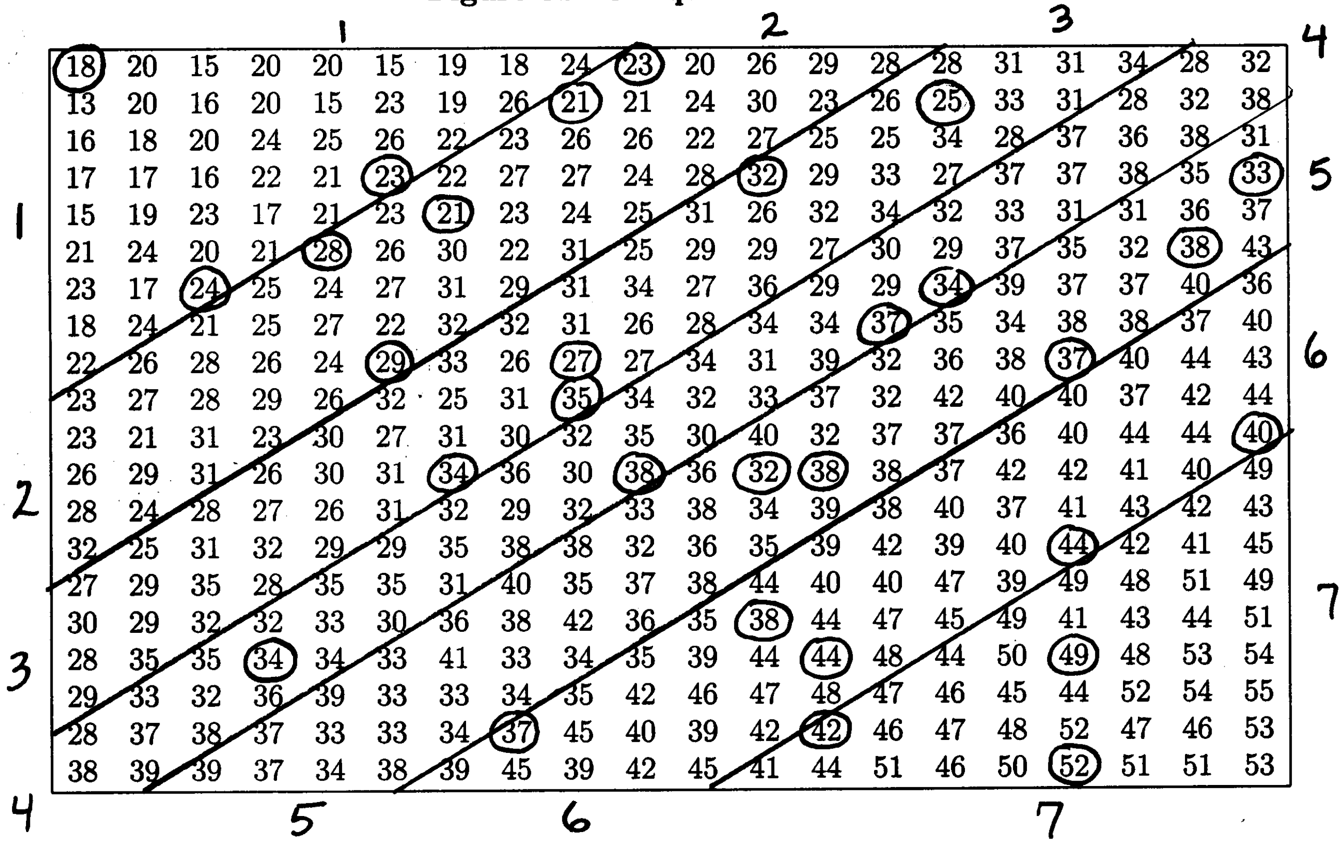
- One-sided confidence intervals can by generated just like those using SRS. Just use t^* using the upper α critical value from the $t(d)$ distribution.

Figure 5a: Equal Size Strata



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

Figure 5b: Unequal Size Strata


1
2
3
4
5
6
7
1
2
3
4
5
6
7

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

3.2 Using R and SAS to Analyze a Stratified SRS

Datasets used in the R code

```
R dataset from Figure 5a    R dataset from Figure 5b
------------------------    ------------------------
   count fpc stratum           count fpc stratum
   25 100       1                 18  45       1
   30 100       1                 23  45       1
   18 100       1                 24  45       1
   28 100       1                 23  60       2
   23 100       1                 21  60       2
   30 100       2                 21  60       2
   26 100       2                 28  60       2
   35 100       2                 29  60       2
   34 100       2                 25  66       3
   33 100       2                 32  66       3
   38 100       3                 27  66       3
   27 100       3                 35  66       3
   30 100       3                 34  66       3
   44 100       3                 34  58       4
   33 100       3                 37  58       4
   36 100       4                 38  58       4
   41 100       4                 34  58       4
   53 100       4                 33  66       5
   47 100       4                 38  66       5
   46 100       4                 37  66       5
                                  38  66       5
                                  32  66       5
                                  40  60       6
                                  44  60       6
                                  38  60       6
                                  44  60       6
                                  37  60       6
                                  49  45       7
                                  42  45       7
                                  52  45       7
```

R code for Stratified SRS (Figure 5a)

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

# t-based confidence intervals for SRS in Figure 5a

library(survey)
strat5adat <- read.table("c:/courses/st446/rcode/fig5a.txt", header=T)
# strat5adat

strat_design <- svydesign(id=~1, fpc=~fpc, strata=~stratum, data=strat5adat)
strat_design

esttotal <- svytotal(~count,strat_design)
print(esttotal,digits=15)
confint.t(esttotal,degf(strat_design),level=.95)
confint.t(esttotal,degf(strat_design),level=.95,tails=’lower’)
confint.t(esttotal,degf(strat_design),level=.95,tails=’upper’)

estmean <- svymean(~count,strat_design)
print(estmean,digits=15)
confint.t(estmean,degf(strat_design),level=.95)
confint.t(estmean,degf(strat_design),level=.95,tails=’lower’)
confint.t(estmean,degf(strat_design),level=.95,tails=’upper’)
```

R output for Stratified SRS (Figure 5a)

```
(For the population total)

-------------------------------------------------------------------
mean( count ) = 13540.00000
SE( count ) = 480.66620
Two-Tailed CI for count where alpha = 0.05 with 16 df
    2.5 %        97.5 %
  12521.03317     14558.96683
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 13540.00000
SE( count ) = 480.66620
One-Tailed (Lower) CI for count where alpha = 0.05 with 16 df
     5 %        upper
  12700.81272     infinity
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 13540.00000
SE( count ) = 480.66620
One-Tailed (upper) CI for count where alpha = 0.05 with 16 df
    lower         95 %
  -infinity     14379.18728
-------------------------------------------------------------------



(For the population mean)

-------------------------------------------------------------------
mean( count ) = 33.85000
SE( count ) = 1.20167
Two-Tailed CI for count where alpha = 0.05 with 16 df
    2.5 %        97.5 %
  31.30258     36.39742
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 33.85000
SE( count ) = 1.20167
One-Tailed (Lower) CI for count where alpha = 0.05 with 16 df
     5 %        upper
  31.75203     infinity
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 33.85000
SE( count ) = 1.20167
One-Tailed (upper) CI for count where alpha = 0.05 with 16 df
    lower         95 %
  -infinity     35.94797
-------------------------------------------------------------------
```

R code for Stratified SRS (Figure 5b)

The R code is exactly the same as the R code for the Figure 5a data analysis except you read in the data file **fig5b.txt**.

R output for Stratified SRS (Figure 5b)

```
(For the population total)
-------------------------------------------------------------------
mean( count ) = 13462.70000
SE( count ) = 256.02201
Two-Tailed CI for count where alpha = 0.05 with 23 df
    2.5 %        97.5 %
  12933.07812     13992.32188
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 13462.70000
SE( count ) = 256.02201
One-Tailed (Lower) CI for count where alpha = 0.05 with 23 df
     5 %        upper
  13023.91117     infinity
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 13462.70000
SE( count ) = 256.02201
One-Tailed (upper) CI for count where alpha = 0.05 with 23 df
    lower         95 %
  -infinity     13901.48883
-------------------------------------------------------------------


(For the population mean)
-------------------------------------------------------------------
mean( count ) = 33.65675
SE( count ) = 0.64006
Two-Tailed CI for count where alpha = 0.05 with 23 df
    2.5 %        97.5 %
  32.33270     34.98080
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 33.65675
SE( count ) = 0.64006
One-Tailed (Lower) CI for count where alpha = 0.05 with 23 df
     5 %       upper
  32.55978     infinity
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 33.65675
SE( count ) = 0.64006
One-Tailed (upper) CI for count where alpha = 0.05 with 23 df
    lower         95 %
  -infinity     34.75372
-------------------------------------------------------------------
```

Using Proc Surveymeans in SAS:

- When the stratum unit totals (N_d) are known, you must create a variable called **_total_** that assigns N_h to each stratum level. It must be called **_total_**. In the following examples, the stratum variable is called **Area**.
- You also need to create a weight variable which takes on the value N_h/n_h. In the following examples, the weight variable is called **W** and it appears in the **Weight** statement.
- Include the option **total=(dataname)** in the Proc Surveymeans statement. (dataname) is the name of the data set. In the first example, the dataname is **fig_5a**. In the second example, the dataname is **fig_5b**.
- Include a **Stratum** statement that contains the stratum variable.
- In the **Var** statement, include the response variable y. In these examples, y is **Count**.
- If you want one-sided confidence intervals for $\overline{y}_U$ or t, in the Proc Surveymeans statement enter **lclm** or **uclm** for $\overline{y}_U$ and **lclmsum** or **uclmsum** for t. In the second example, I included all 4 options.
- The **list** option in the Stratum statement produces a table containing information about each stratum.

Analysis of the Stratified SRS in Figure 5a

```
data fig5a;
   input Area Count @@;
datalines;
  1 18   1 23   1 28   1 25   1 30
  2 35   2 30   2 26   2 33   2 34
  3 33   3 27   3 30   3 44   3 38
  4 47   4 36   4 41   4 53   4 46
;
data fig5a; set fig5a;

  if Area = 1 then _total_= 100;  *** _total_ = Nh ;
  if Area = 2 then _total_= 100;
  if Area = 3 then _total_= 100;
  if Area = 4 then _total_= 100;

  if Area=1 then W = 100/5;       *** W = Nh / nh ;
  if Area=2 then W = 100/5;
  if Area=3 then W = 100/5;
  if Area=4 then W = 100/5;
title1 'Analysis of Stratified SRS in Figure 5a';

proc surveymeans data=fig5a total=fig5a mean clm sum clsum df;
    Stratum Area / list;
    Var Count;
    Weight W;
run;
===============================================================
                 Analysis of Stratified SRS in Figure 5a

                        The SURVEYMEANS Procedure

                               Data Summary

                   Number of Strata                4
                   Number of Observations         20
                   Sum of Weights                400
```

```
                           Stratum Information

  Stratum           Population   Sampling
   Index     Area        Total       Rate      N Obs   Variable          N
  ------------------------------------------------------------------------
     1          1          100      5.00%          5   Count             5
     2          2          100      5.00%          5   Count             5
     3          3          100      5.00%          5   Count             5
     4          4          100      5.00%          5   Count             5
  ------------------------------------------------------------------------

                               Statistics

                                    Std Error
  Variable     DF        Mean         of Mean       95% CL for Mean
  --------------------------------------------------------------------
  Count        16   33.850000        1.201666    31.3025829 36.3974171
  --------------------------------------------------------------------

  Variable               Sum         Std Dev        95% CL for Sum
  -------------------------------------------------------------------
  Count                13540      480.666204    12521.0332 14558.9668
  -------------------------------------------------------------------
```

Analysis of Stratified SRS in Figure 5b

```
data fig5b;
   input Area Count @@;
datalines;
  1 18   1 24   1 23
  2 28   2 29   2 21   2 21   2 23
  3 34   3 27   3 35   3 32   3 25
  4 34   4 38   4 37   4 34
  5 32   5 38   5 37   5 38   5 33
  6 37   6 38   6 44   6 44   6 40
  7 42   7 49   7 52
;
data fig5b; set fig5b;

  if Area = 1 then _total_= 45;  *** _total_ = Nh ;
  if Area = 2 then _total_= 60;
  if Area = 3 then _total_= 66;
  if Area = 4 then _total_= 58;
  if Area = 5 then _total_= 66;
  if Area = 6 then _total_= 60;
  if Area = 7 then _total_= 45;

  if Area=1 then W = 45/3;       *** W = Nh / nh ;
  if Area=2 then W = 60/5;
  if Area=3 then W = 66/5;
  if Area=4 then W = 58/4;
  if Area=5 then W = 66/5;
  if Area=6 then W = 60/5;
  if Area=7 then W = 45/3;

title1 'Analysis of Stratified SRS in Figure 5b';

proc surveymeans data=fig5b total=fig5b mean clm sum clsum df
                 lclm uclm lclmsum uclmsum ;
    Stratum Area / list;
    Var Count;
    Weight W;
run;
```

```
                 Analysis of Stratified SRS in Figure 5b

                        The SURVEYMEANS Procedure

                               Data Summary

                   Number of Strata                    7
                   Number of Observations             30
                   Sum of Weights                    400

                            Stratum Information

Stratum               Population   Sampling
 Index         Area        Total       Rate      N Obs   Variable      N
-----------------------------------------------------------------------
   1              1           45      6.67%          3   Count         3
   2              2           60      8.33%          5   Count         5
   3              3           66      7.58%          5   Count         5
   4              4           58      6.90%          4   Count         4
   5              5           66      7.58%          5   Count         5
   6              6           60      8.33%          5   Count         5
   7              7           45      6.67%          3   Count         3
-----------------------------------------------------------------------

                               Statistics

                                          Std Error
Variable      DF           Mean             of Mean       95% CL for Mean
------------------------------------------------------------------------
Count         23      33.656750            0.640055    32.3326953 34.9808047
------------------------------------------------------------------------

             Lower 95%        Upper 95%
          One-Sided CL     One-Sided CL
Variable      for Mean         for Mean             Sum           Std Dev
------------------------------------------------------------------------
Count        32.559778        34.753722           13463        256.022011
------------------------------------------------------------------------

                                           Lower 95%        Upper 95%
                                          One-Sided        One-Sided
 Variable          95% CL for Sum        CL for Sum       CL for Sum
 ------------------------------------------------------------------
 Count        12933.0781 13992.3219           13024            13901
 ------------------------------------------------------------------
```

3.3 Efficiency of Stratified Simple Random Sampling

- Because the variance formulas for $\widehat{t}_{str}$ and $\widehat{\overline{y}_U}_{str}$ are determined only from within-stratum variances, the precision of the estimators can be improved by forming strata with small S_h^2 values (strata with similar y-values within each stratum). We will compare $\widehat{V}(\widehat{\overline{y}_U})$ from a SRS to $\widehat{V}(\widehat{\overline{y}_{str}})$ from a stratified SRS.

- The population variance can be rewritten as the weighted sum of within-stratum and between-stratum variabilities:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\overline{y}_U)^2 \\ &= \frac{1}{N-1}\left[\sum_{h=1}^{H}(N_h-1)S_h^2+\sum_{h=1}^{H}N_h(\overline{y}_{hU}-\overline{y}_U)^2\right] \end{aligned}$$

- By substituting this alternative form of S^2 into $V(\widehat{\overline{y}_U})$ and $V(\widehat{\overline{y}_U}_{str})$, it can be shown that:

$$V(\widehat{\overline{y}_U}) - V(\widehat{\overline{y}_U}_{str}) = \frac{N-n}{Nn(N-1)} \left[\sum_{h=1}^{H} N_h(\overline{y}_{hU} - \overline{y}_U)^2 - \frac{1}{N}\sum_{h=1}^{H}(N - N_h)S_h^2 \right].$$

- If this difference in variances is positive, or, equivalently, if

$$\sum_{h=1}^{H} N_h(\overline{y}_{hU} - \overline{y}_U)^2 \; > \; \frac{1}{N}\sum_{h=1}^{H}(N - N_h)S_h^2 \; ,$$

  then we say that $\widehat{\overline{y}_U}_{str}$ **is more efficient** than $\widehat{\overline{y}_U}$.

- A stratified SRS estimator will be more efficient than the SRS estimator of $\overline{y}_U$ or t if the variability between stratum means is sufficiently large relative to the within-stratum variability. This is what happened with the stratification used in Figures 5a and 5b.

3.4 Allocation of Sampling Units

- Given that we have enough resources to allocate n units among the H strata, how do we determine the stratum sample sizes n_h?

- **Situation 1**: If <u>all strata are the same size</u> and no prior information is available about the population, a reasonable choice would be to assign equal (or nearly equal) sample sizes to the strata. That is, $n_h \approx$.

  - Example: Consider the stratified population in Figure 5a. Suppose there are enough resources to take a sample of size $n = 50$. How many samples should be taken for each stratum assuming Situation 1?

- **Situation 2**: If <u>the strata are not all the same size</u> and no prior information is available about the population, a reasonable choice would be to assign sample sizes proportional to the sizes of the strata relative to the population size N. That is, n_h . This is known as **proportional allocation**.

  - Example: Consider the stratified population in Figure 5b. Suppose there are enough resources to take a sample of size $n = 50$. How many samples should be taken for each stratum assuming proportional allocation?

- **Situation 3**: The allocation scheme that minimizes $V(\widehat{t}_{str})$ is called **optimum allocation** and requires

$$n_h \; =$$

  Because the S_h^2 values are unknown, we would need prior estimates (possibly from past data or published studies) to attempt optimum allocation.

  - Example: Consider the stratified population in Figure 5b. Suppose there are enough resources to take a sample of size $n = 50$ and we have prior estimates of $s_1 = 3.2$, $s_2 = 3.8$, $s_3 = 4.4$, $s_4 = 2.1$, $s_5 = 2.9$, $s_6 = 3.3$, and $s_7 = 5.1$. How many samples should be taken for each stratum assuming optimum allocation?

- **Situation 4**: In some cases, if the cost of sampling units varies from stratum to stratum, then the total cost of taking a stratified SRS may determine how to allocate units to strata.
  - Let c_0 be the fixed (also called "overhead") cost of the survey that does not depend on what units are in sample. Let c_h be the cost to sample a unit from stratum h. The total cost C of the sample will be $C = c_0 + \sum_{h=1}^{H} c_h n_h$.
  - Case I: **For a fixed total cost** C, the smallest variance $V(\widehat{\overline{y}_U})$ or $V(\widehat{t})$ is achieved by choosing n_h such that:

$$n_h = \frac{(C - c_0)N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h \sqrt{c_h}}$$

  - Case II: **For a fixed (specified) variance** $V(\widehat{\overline{y}_U})$, the smallest cost is achieved by first determining the total sample size n such that

$$n = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{H} N_h S_h/\sqrt{c_h}\right)}{N^2 V + \sum_{h=1}^{H} N_h S_h^2}$$

    where V is the fixed variance specified by the researcher. Then, the stratum sample size n_h for $h = 1, 2, \ldots, H$ is

$$n_h = \frac{n N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h/\sqrt{c_h}}$$

  For a fixed $V(\widehat{t})$, use $V = V(\widehat{t})/N^2$ in the formula.

- If all of the costs (c_h) are the same, then the total sample size formula reduces to

$$n = \frac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{N^2 V + \sum_{h=1}^{H} N_h S_h^2}$$

- Because the S_h values are unknown in either Case I or Case II, we would need prior estimates (possibly from past data or published studies) to attempt optimum allocation.

<u>**Example of Situation 4: Case I**</u>: Suppose there is a fixed total cost $C = \$3000$ and a fixed overhead cost of $c_0 = \$500$. Consider the stratification used in Figure 5b. The unit sampling costs are

$$\begin{aligned} c_1 &= \text{\$20 per unit from stratum 1} \\ c_2 = c_3 &= \text{\$25 per unit from stratum 2 or 3} \\ c_4 &= \text{\$30 per unit from stratum 4} \\ c_5 = c_6 &= \text{\$35 per unit from stratum 5 or 6} \\ c_7 &= \text{\$40 per unit from stratum 7} \end{aligned}$$

Then, using s_h^2 as an estimate of S_h^2: $\quad n_h = \dfrac{(C - c_0)N_h s_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h s_h \sqrt{c_h}} = \dfrac{2500\, N_h s_h/\sqrt{c_h}}{7657.776}$

						rounded	projected	
Stratum	N_h	s_h	c_h	$N_h s_h \sqrt{c_h}$	$N_h s_h/\sqrt{c_h}$	n_h	n_h	cost
1	45	3.215	20	647.006	32.350	10.6	11	$220
2	60	3.847	25	1154.100	46.164	15.1	15	$375
3	66	4.393	25	1449.690	57.988	18.9	19	$475
4	58	2.062	30	655.054	21.835	7.1	7	$210
5	66	2.881	35	1124.919	32.141	10.5	10	$350
6	60	3.286	35	1166.414	33.326	10.9	11	$385
7	45	5.132	40	1460.593	36.515	11.9	12	$480

The estimated total cost is $\$ \qquad + c_0 = \qquad + \$500 = \qquad$ requiring sampling units.

Example of Situation 4: Case II: Suppose there is a fixed variance of $V = V(\widehat{t}) = .35$. Consider the stratification used in Figure 5b. The costs are the same as Case I.

Then, using s_h as an estimate of S_h:

$$n = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{H} N_h S_h/\sqrt{c_h}\right)}{N^2 V + \sum_{h=1}^{H} N_h S_h^2} \approx \frac{(7657.776)(260.319)}{(400^2)(.35) + 5254.1} =$$

Then, substitution yields

$$n_h = \frac{n N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h/\sqrt{c_h}} = \frac{(37.433) N_h S_h/\sqrt{c_h}}{260.319} \approx \qquad N_h S_h/\sqrt{c_h}.$$

							rounded	projected	
Stratum	N_h	s_h	c_h	$N_h s_h \sqrt{c_h}$	$N_h s_h/\sqrt{c_h}$	$N_h S_h^2$	n_h	n_h	cost
1	45	3.215	20	647.006	32.350	465.0	4.65	5	$100
2	60	3.847	25	1154.100	46.164	888.0	6.64	7	$175
3	66	4.393	25	1449.690	57.988	1273.8	8.34	8	$200
4	58	2.062	30	655.054	21.835	246.5	3.14	3	$ 90
5	66	2.881	35	1124.919	32.141	547.8	4.62	5	$175
6	60	3.286	35	1166.414	33.326	648.0	4.79	5	$175
7	45	5.132	40	1460.593	36.515	1185.0	5.25	5	$200
				7657.776	260.319				

Thus, the minimum cost to achieve V is $\$ \qquad + c_0 = \qquad + \$500 = \qquad$ requiring a total of sampling units.