One-stage Cluster Sampling

A cluster sample (one-stage) is a probability sample, often a simple random sample, in which each sampling unit is actually a collection, or cluster, of elements.

Why would we want to take a sample in this manner?

1. In many surveys, no reliable list of the elements in a population is available or it would be prohibitively expensive to construct such a list. On the other hand a unit that consists of elements can be easily identified. Thus, we solve the problem of constructing a list of elements as we will sample another type of unit, the cluster, that contains the units of interest. Thus we build a frame of clusters. Clusters are the (primary) sampling unit.
2. Economics may suggest using large clusters as an effective sampling strategy. Sample city blocks rather than individual households, or sample villages rather than individual fishermen. The cost of obtaining observations often increases as the distance separating the elements increases. Thus, we may have gains from intensive sampling of clusters rather than devoting lots of time to travel and search.

Cluster sampling is usually used for gains in efficiency (reduced costs) and rarely for gains in precision.

In stratification we wanted the strata to be as much alike within and as different as possible among strata. In cluster sampling, we will want the clusters to be as heterogeneous as possible within and as homogeneous among clusters.

Single-stage or one-stage cluster sampling is often just a simple random sample of clusters. However, variations exist as we may receive performance gains if we select the clusters on the basis of their "size," say by using PPS sampling.

Let y_{ij} be the observed value for the jth element of the ith cluster and let y_i be the cluster total. We will take a simple random sample of size m clusters from the population of size M clusters.

Each cluster contains N_i elements for a total population of size $N = \sum_{i=1}^{M} N_i$ elements. We will then measure <u>all</u> elements in the selected clusters.

1.0 Clusters of equal size

In this section we will require that $N_i = \bar{N}$ for all clusters. To estimate the population total τ, use

$$\hat{\tau} = \left(\frac{M}{m}\right)\sum_{i=1}^{m} y_i = \left(\frac{M}{m}\right)\sum_{i=1}^{m}\sum_{j=1}^{\bar{N}} y_{ij}$$

with variance

$$\text{var}(\hat{\tau}) = M^2\left(\frac{M-m}{M}\right)\frac{s^2_{1y}}{m}$$

where

$$s^2_{1y} = \frac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{m-1}$$

is the variance of the cluster totals and $\bar{y}$ is the mean of the cluster totals. A similar formula exists for the mean per cluster,

$$\hat{\mu}_c = \frac{1}{M}\hat{\tau} = \bar{y}$$

with variance

$$\text{var}(\hat{\mu}_c) = \frac{s^2_{1y}}{m}\left(\frac{M-m}{M}\right).$$

In general, we would not be all that interested in the cluster mean, but often more interested in the mean per element, μ_e,

$$\hat{\mu}_e = \frac{1}{\bar{N}}\hat{\mu}_c = \frac{\bar{y}}{\bar{N}},$$

where we are dividing the cluster mean among its elements. It is an unbiased estimator of the mean per element. Further, its estimated variance is

$$\text{var}(\hat{\mu}_e) = \frac{s^2_{1y}}{\bar{N}^2 m}\left(\frac{M-m}{M}\right).$$

Example 24: Newspaper circulation.

The circulation manager of a newspaper wishes to estimate the average number of newspapers purchased per household in a given community. Travel costs from household to household are substantial. Therefore, the 4000 households in the community are listed in 400 geographical clusters of 10 households each, and a simple random sample of 4 clusters is selected. Interviews are conducted, and the data are tabulated. Estimate the average number of newspapers per household for the community and place a bound on the error of estimation (From Mendenhall et al. 1990: 256-257).

Cluster	Number of newspapers recorded from 10 households										Cluster Total y_i	Cluster Size N_i
1	1	2	1	3	3	2	1	4	1	1	19	10
2	1	3	2	2	3	1	4	1	1	2	20	10
3	2	1	1	1	1	3	2	1	3	1	16	10
4	1	1	3	2	1	5	1	2	3	1	20	10

$$\sum_{i=1}^{4} y_i = 19 + 20 + 16 + 20 = 75,\ M = 400,\ m = 4,\ \bar{N} = 10,\ \bar{y} = \frac{75}{4} = 18.75$$

$$s_{1y}^2 = \frac{(19 - 18.75)^2 + (20 - 18.75)^2 + (16 - 18.75)^2 + (20 - 18.75)^2}{4 - 1} = 3.58333$$

The estimated mean number of newspapers per household is

$$\hat{\mu}_e = \frac{\bar{y}}{\bar{N}} = \frac{18.75}{10} = 1.875$$

with estimated variance of

$$\text{var}(\hat{\mu}_e) = \frac{1}{10^2}\left(1 - \frac{4}{400}\right)\left(\frac{3.58333}{4}\right) = 0.00887$$

and standard error of

$$\text{s.e.}(\hat{\mu}_e) = \sqrt{0.00887} = 0.09417.$$

This yields a 95% confidence interval of $\hat{\mu}_e \pm 2(\text{s.e.}(\hat{\mu}_e))$ or $(1.6904, 2.0596)$ newspapers per household.

Analysis of newspaper circulation using SURVEYMEANS

```
*****************************************************;
* news.sas -- Cluster Sampling of Households.       *;
*****************************************************;
Options PS=55 LS=78 NoDate PageNo=1 NoCenter
        FORMCHAR='|----|+|---+=|-/\<>*';
Title1 "Cluster Sampling of Households Within Groups of 10";
Data News;
 Input Cluster @;
 SamplingWeight=400/40; /* Clusters of Equal Size */
 Do House=1 To 10;
  Input Papers @;
  Output;
 End;
 Label Cluster="Cluster"
       Papers="Number of Newspapers"
       House="Household"
       SamplingWeight="Sampling Weight";
Datalines;
1  1 2 1 3 3 2 1 4 1 1
2  1 3 2 2 3 1 4 1 1 2
3  2 1 1 1 1 3 2 1 3 1
4  1 1 3 2 1 5 1 2 3 1
;

Proc Sort Data=News;
 By Cluster;
Run;

Proc Print Data=News Split=" ";
Run;

Title2 "Intermediate Values For Calculations";
Proc Means Data=News N Mean Sum Var StdErr;
 Class Cluster;
 Var Papers;
 Output Out=Estimate N=ClusterSize Mean=ElementMean Sum=ClusterMean
Var=S2i;
Run;
Quit;

Proc Means Data=Estimate N Mean Var StdErr;
 Where (_TYPE_=1);
 Var ElementMean ClusterSize ClusterMean;
Run; Quit;

Title2 "Estimation of Population Mean and Total";
Proc SurveyMeans Data=News N=400 Mean CLM Sum CLSum CV;
 Cluster Cluster;
 Var Papers;
 Weight SamplingWeight;
Run;
```

```
Title2 "Decomposition of Sums of Squares";
Proc ANOVA Data=News;
 Class Cluster;
 Model Papers = Cluster;
Run; Quit;

Title2 "Variance Components Estimation";
Proc MIXED Data=News Covtest;
 Class Cluster;
 Model Papers =;
 Random Cluster;
 Parms / Nobound;
 Estimate "Newspapers per Household" Intercept 1 / CL;
Run;
```

Cluster Sampling of Households Within Groups of 10 1

Obs	Cluster	Sampling Weight	Household	Number of Newspapers
1	1	10	1	1
2	1	10	2	2
3	1	10	3	1
4	1	10	4	3
5	1	10	5	3
6	1	10	6	2
7	1	10	7	1
8	1	10	8	4
9	1	10	9	1
10	1	10	10	1
11	2	10	1	1
12	2	10	2	3
13	2	10	3	2
14	2	10	4	2
15	2	10	5	3
16	2	10	6	1
17	2	10	7	4
18	2	10	8	1
19	2	10	9	1
20	2	10	10	2
21	3	10	1	2
22	3	10	2	1
23	3	10	3	1
24	3	10	4	1
25	3	10	5	1
26	3	10	6	3
27	3	10	7	2
28	3	10	8	1
29	3	10	9	3
30	3	10	10	1
31	4	10	1	1
32	4	10	2	1
33	4	10	3	3

34	4	10	4	2
35	4	10	5	1
36	4	10	6	5
37	4	10	7	1
38	4	10	8	2
39	4	10	9	3
40	4	10	10	1

Cluster Sampling of Households Within Groups of 10 2
Intermediate Values For Calculations

The MEANS Procedure
Analysis Variable : Papers Number of Newspapers

Cluster	N Obs	N	Mean	Sum	Variance	Std Error
1	10	10	1.9000000	19.0000000	1.2111111	0.3480102
2	10	10	2.0000000	20.0000000	1.1111111	0.3333333
3	10	10	1.6000000	16.0000000	0.7111111	0.2666667
4	10	10	2.0000000	20.0000000	1.7777778	0.4216370

Cluster Sampling of Households Within Groups of 10 3
Intermediate Values For Calculations

The MEANS Procedure

Variable	Label	N	Mean	Variance	Std Error
ElementMean	Number of Newspapers	4	1.8750000	0.0358333	0.0946485
ClusterSize	Number of Newspapers	4	10.0000000	0	0
ClusterMean	Number of Newspapers	4	18.7500000	3.5833333	0.9464847

Cluster Sampling of Households Within Groups of 10 4
Estimation of Population Mean and Total

The SURVEYMEANS Procedure
Data Summary

Number of Clusters	4
Number of Observations	40
Sum of Weights	400

Statistics

Variable	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Coeff of Variation
Papers	1.875000	0.094174	1.575296	2.174704	0.050226

Statistics

Variable	Sum	Std Dev	Lower 95% CL for Sum	Upper 95% CL for Sum
Papers	750.000000	37.669616	630.118468	869.881532

The ANOVA Procedure

Class Level Information

Class	Levels	Values
Cluster	4	1 2 3 4

Number of observations 40

The ANOVA Procedure

Dependent Variable: Papers Number of Newspapers

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.07500000	0.35833333	0.30	0.8266
Error	36	43.30000000	1.20277778		
Corrected Total	39	44.37500000			

R-Square	Coeff Var	Root MSE	Papers Mean
0.024225	58.49132	1.096712	1.875000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Cluster	3	1.07500000	0.35833333	0.30	0.8266

The Mixed Procedure

Class Level Information

Class	Levels	Values
Cluster	4	1 2 3 4

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z
Cluster	-0.08444	0.04074	-2.07	0.0382
Residual	1.2028	0.2835	4.24	<.0001

The Mixed Procedure

Fit Statistics

Res Log Likelihood	-59.0
Akaike's Information Criterion	-61.0
Schwarz's Bayesian Criterion	-60.4
-2 Res Log Likelihood	117.9

PARMS Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	1.47	0.2257

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Newspapers per Household	1.8750	0.09465	3	19.81	0.0003	0.05

Estimates

Label	Lower	Upper
Newspapers per Household	1.5738	2.1762

1.1 Comparisons with Simple Random Sampling

Let S^2 be the usual variance among the elements. Thus, had we taken a simple random sample that had the same number of elements as our cluster sample, i.e. $m\bar{N}$, then the variance of the mean would have been

$$\mathrm{Var}(\bar{y}) = \left(1 - \frac{m\bar{N}}{M\bar{N}}\right)\frac{S^2}{m\bar{N}}$$

where $S^2 = \dfrac{\sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(y_{ij} - \tilde{\bar{Y}})^2}{M\bar{N} - 1}$ and $\tilde{\bar{Y}}$ is the population mean per element. We can decompose this "total" variation into components between and within clusters. Let's do this for the sums of squares

$$\sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(y_{ij} - \tilde{\bar{Y}})^2 = \sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(\bar{y}_i - \tilde{\bar{Y}})^2 + \sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(y_{ij} - \bar{y}_i)^2$$

or

$$\sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(y_{ij} - \tilde{\bar{Y}})^2 = \bar{N}\sum_{i=1}^{M}(\bar{y}_i - \tilde{\bar{Y}})^2 + \sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(y_{ij} - \bar{y}_i)^2.$$

The sum of squares can then be written as in the ANOVA as

$$\mathrm{SS}_T = \mathrm{SS}_B + \mathrm{SS}_W$$

or the total sums of squares are composed of the between plus the within sums of squares. The mean squares in the sample can be written as

$\mathrm{MS}_B = \dfrac{\mathrm{SS}_B}{m-1}$ and $\mathrm{MS}_W = \dfrac{\mathrm{SS}_W}{m(\bar{N}-1)}$ for between and within mean squares, respectively.

Therefore the cluster estimate of variance of the per element mean can be written as

$$\mathrm{var}(\hat{\mu}_e) = \left(1 - \frac{m}{M}\right)\frac{\mathrm{MS}_B}{m\bar{N}}$$

For an ANOVA decomposition of the previous newspaper example:

Source	df	SS	MS	Mean
Cluster	3	1.075	0.358333	1.875
Error	36	43.30	1.202778	
Total	39	44.375	1.137800	

we can compute $\mathrm{var}(\hat{\mu}_e) = \left(1 - \dfrac{4}{400}\right)\dfrac{0.358333}{4(10)} = 0.0088687$ which is the same as what we had computed earlier. Further, we can now see that $S^2_{1Y} = \bar{N}(\mathrm{MS}_B)$ and $\mathrm{MS}_B = \dfrac{S^2_{1Y}}{\bar{N}}$. Based upon this ANOVA decomposition of our sums of squares, we can get an approximate comparison of cluster sampling with simple random sampling. For simple random sampling, we can approximate the variance by

$$S^2 = \frac{M(\bar{N}-1)(\mathrm{MS}_W) + (M-1)\mathrm{MS}_B}{M\bar{N} - 1}$$

and in the sample

$$s^2 = \frac{m(\bar{N}-1)(\mathrm{MS}_W) + (m-1)\mathrm{MS}_B}{m\bar{N} - 1}$$

Notice that this estimate is the mean square total reported in our table above, which is usually not computed in the ANOVA table. Since sampling is not truly "simple random" with respect to the elements, it is a biased estimate of σ^2 under SRS. However, this does give us an approximate way to compare the sampling plans on the basis of efficiency. The relative efficiency of SRS to cluster sampling (single-stage with equal cluster sizes) is

$$\mathrm{RE}\left(\frac{\hat{\bar{\mu}}_{\mathrm{clu}}}{\bar{\mu}_{\mathrm{SRS}}}\right) = \frac{\mathrm{Var}(\hat{\bar{\mu}}_{\mathrm{SRS}})}{\mathrm{Var}(\hat{\bar{\mu}}_{\mathrm{clu}})} = \frac{S^2}{\mathrm{MS}_B}.$$

For the newspaper example, $\hat{S}^2 \approx \dfrac{\mathrm{SS}_T}{\mathrm{df}_T} = \mathrm{MS}_T = 1.1378$ and, therefore,

$\mathrm{RE}\left(\dfrac{\hat{\bar{\mu}}_{\mathrm{clu}}}{\bar{\mu}_{\mathrm{SRS}}}\right) = \dfrac{1.1378}{0.3583} = 3.18$ which suggests that cluster sampling is superior here to simple random sampling. This result, in general, will not hold for cluster sampling. Typically, clusters are homogeneous within and heterogeneous among, leading to a large variance for the estimate of the population parameters. For the newspaper example, cluster sampling was a good choice of technique.

What else can be learned from the comparison of these methods? First, write the intracluster correlation coefficient in terms of the known variance components. To simplify the formulas write S_b^2 for the MS_B.

$$\rho = \frac{\text{E}((y_{ij}-\mu_e)(y_{ik}-\mu_e))}{\text{E}((y_{ij}-\mu_e)^2)} = \frac{(M-1)\bar{N}S_b^2-(M\bar{N}-1)S^2}{(M\bar{N}-1)(\bar{N}-1)S^2} \approx \frac{S_b^2-S^2}{(\bar{N}-1)S^2}$$

Then the variance of $\hat{\mu}_e$ can be written as

$$\text{Var}(\hat{\mu}_e) = \left(1-\frac{m}{M}\right)\left(\frac{M\bar{N}-1}{\bar{N}^2(M-1)}\right)\frac{S^2}{m}[1+(\bar{N}-1)\rho]$$
$$\approx \left(1-\frac{m}{M}\right)\frac{S^2}{m\bar{N}}[1+(\bar{N}-1)\rho] = \left(1-\frac{m\bar{N}}{M\bar{N}}\right)\frac{S^2}{m\bar{N}}[1+(\bar{N}-1)\rho]$$

This is the variance for a simple random sample of the $m\bar{N}$ units if we make $\rho = 0$. When $\rho < 0$, then cluster sampling is more precise than SRS, while if $\rho > 0$, then it is less precise than SRS. From our previous approximation of ρ, $\rho < 0$ when $S_b^2 - S^2 < 0$ or when $\frac{2}{b} < S^2$. As before, this states that for cluster sampling to be superior to simple random sampling, the between cluster variability must be small relative to the within cluster variability. Note that S^2 is composed of both within and between variabilities.

1.2 Sample Size Considerations

Recall that the single-stage cluster sampling formulas with equal cluster sizes are the simple random sampling formulas encountered earlier in the course. Thus, we can derive sample size formulas directly from our earlier simple random sampling formulas. For a bound B on the error of estimation of the cluster total mean use

$$m = \frac{1}{\frac{1}{m_0}+\frac{1}{M}} \text{ where } m_0 = \frac{Z^2 S_{1Y}^2}{B^2}.$$

For a specified bound B_e on the estimate of the mean per element convert between the variance of the mean per element and the variance of the mean per cluster total. This gives the formula

$$m_0 = \frac{Z^2 S_{1Y}^2}{\bar{N}^2 B_e^2}.$$

2.0 Clusters of unequal size

When a simple random sample of clusters is selected and the clusters have unequal sizes (numbers of elements per cluster), then an inflation estimator or a ratio-to-size estimator are possible. Depending upon whether or not you know the population number of elements, N, and whether interest is in estimating the population mean or total, you may only be able to use one or the other, as you'll see below.

2.0.1 Inflation Estimator

When we have clusters of unequal size, the values N_i will differ from cluster to cluster. Again let

$$y_i = \sum_{j=1}^{N_i} y_{ij} = N_i \bar{y}_i$$

be the ith cluster total. Then

$$\hat{\tau} = \frac{M}{m}\sum_{i=1}^{m} y_i$$

is an unbiased estimator of τ with variance

$$\text{Var}(\hat{\tau}) = M^2\left(1-\frac{m}{M}\right)\frac{1}{m}\left(\sum_{i=1}^{M}(y_i-\bar{Y})^2\right)/(M-1),$$

where $\bar{Y} = \tau/M$ is the population mean per cluster. The estimated variance can be computed as

$$\hat{\text{var}}(\hat{\tau}) = M^2\left(1-\frac{m}{M}\right)\frac{1}{m}\left(\sum_{i=1}^{m}(y_i-\bar{y})^2\right)/(m-1)$$

where $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ is the average of the sample cluster totals. *Note that this estimator does not require knowledge of N.*

To estimate the population mean per element, note that $\hat{\mu}_e = \hat{\tau}/N$ is the unbiased mean per element estimator, though *it does require knowledge of N*.

Although these estimators are unbiased, they may often have large variances since unequal numbers of units among clusters will generate considerable variability among the cluster totals, which are the values upon which the variance depends. If the cluster sizes N_i are strongly related to the cluster totals, which they usually are, then a ratio estimator is often preferred.

2.0.2 Ratio to Size Estimator

When clusters are of unequal size, the ratio-to-size estimator is often a more efficient estimator. To estimate the population mean per element use

$$\hat{\mu}_e = \tilde{\bar{y}} = r = \left(\sum_{i=1}^{m} y_i\right) \Big/ \left(\sum_{i=1}^{m} N_i\right)$$

which is a ratio estimator with N_i as the ancillary variable, and which has variance,

$$\text{var}(\hat{\mu}_{eR}) = \left(\frac{M-m}{M}\right)\frac{1}{\bar{N}^2 m}\sum_{i=1}^{m}(y_i - \tilde{\bar{y}}N_i)^2/(m-1).$$

If $\bar{N}$ is unknown then use the sample based estimator $\hat{\bar{N}} = \bar{n} = \frac{1}{m}\sum_{i=1}^{m} N_i$. Note that although this estimator is biased, *it only requires information obtained from the sample*, while the usual unbiased estimator requires knowledge of N.

To estimate the population total, simply inflate the mean per element estimate,

$$\hat{\tau}_R = N\left(\sum_{i=1}^{m} y_i\right) \Big/ \left(\sum_{i=1}^{m} N_i\right) = N\hat{\mu}_e$$

though *this requires knowledge of* N. The variance is computed using the ratio estimator formulas as

$$\text{var}(\hat{\tau}_R) = M^2\left(\frac{M-m}{M}\right)\frac{1}{m}\sum_{i=1}^{m}(y_i - rN_i)^2/(m-1).$$

We can write the variance formula as

$$\text{var}(\hat{\tau}_R) = M^2\left(\frac{M-m}{M}\right)\frac{1}{m}\sum_{i=1}^{m}(y_i - \tilde{\bar{y}}N_i)^2/(m-1)$$

which may be rewritten for hand calculation using the relationship

$$\sum_{i=1}^{m}(y_i - \tilde{\bar{y}}N_i)^2 = \sum_{i=1}^{m} y_i^2 + \tilde{\bar{y}}^2\sum_{i=1}^{m} N_i^2 - 2\tilde{\bar{y}}\sum_{i=1}^{m} y_i N_i$$

or in terms of variances and covariances of the $\{y_i, N_i\}$ where

$$\sum_{i=1}^{m}(y_i - \tilde{\bar{y}}N_i)^2/(m-1) = s^2_{1y} + \tilde{\bar{y}}^2 s^2{}_N - 2\tilde{\bar{y}}s_{1yN},$$

where s^2_{1y} is the variance of the cluster totals, $s^2{}_N$ is the variance of the cluster sizes, and s_{1yN} is the covariance of the cluster totals with cluster sizes.

2.0.3 Unweighted mean of cluster means

One might be tempted to use the unweighted mean of the cluster means as an estimator for μ, i.e., $\hat{\mu}^* = \frac{1}{m}\sum_{i=1}^{m}\bar{y}_i$ ($\bar{y}_i$ is the mean per element in the ith cluster). However, with unequal N_i, not only is this estimator biased, but as $m \to M$, that is, as the sample size approaches sampling the population, $\hat{\mu}^*$ does not get closer to μ, and therefore, it is not a consistent estimator of μ. It is, therefore, not recommended for use.

Example 25: Estimation of current event awareness.

```
*******************************************************;
* events.sas -- Cluster Sampling for Current Events  *;
* Population has 108 Classes.                        *;
* Scheaffer et al., Problem 8.6.                     *;
*******************************************************;
Options PS=55 LS=78 NoDate PageNo=1 NoCenter
        FORMCHAR='|----|+|---+=|-/\<>*';
Title1 "Clusters of Unequal Size: Ratio Estimation";
Data Events;
 Class+1;
 Input Students Score @@;
 RatioWt=1/Students;
 Label Students="Number of Students"
       Score="Total Score"
       Class="Class"
       RatioWt="Ratio Weight";
Datalines;
31 1590 29 1510 25 1490 35 1610 15 800 31 1720 22 1310 27 1427
25 1290 19 860 30 1620 18 710 21 1140 40 1980 38 1990
28 1420 17 900 22 1080 41 2010 32 1740 35 1750 19 890
29 1470 18 910 31 1740
;

Proc Print Data=Events Split=" ";
Run;
```

```
Proc SurveyReg Data=Events N=108;
 Model Score = Students / NoInt CLParm;
 Weight RatioWt;
 Estimate "Mean per Student" Students 1;
Run;
```


Obs	Class	Number of Students	Total Score	Ratio Weight
1	1	31	1590	0.032258
2	2	29	1510	0.034483
3	3	25	1490	0.040000
4	4	35	1610	0.028571
5	5	15	800	0.066667
6	6	31	1720	0.032258
7	7	22	1310	0.045455
8	8	27	1427	0.037037
9	9	25	1290	0.040000
10	10	19	860	0.052632
11	11	30	1620	0.033333
12	12	18	710	0.055556
13	13	21	1140	0.047619
14	14	40	1980	0.025000
15	15	38	1990	0.026316
16	16	28	1420	0.035714
17	17	17	900	0.058824
18	18	22	1080	0.045455
19	19	41	2010	0.024390
20	20	32	1740	0.031250
21	21	35	1750	0.028571
22	22	19	890	0.052632
23	23	29	1470	0.034483
24	24	18	910	0.055556
25	25	31	1740	0.032258

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Score

Data Summary

Number of Observations	25
Sum of Weights	0.99632
Weighted Mean of Score	1291.8
Weighted Sum of Score	1287.0

Fit Summary

R-square	0.9942
Root MSE	104.80
Denominator DF	24

ANOVA for Dependent Variable Score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1802348	1802348	4118.05	<.0001
Error	24	10504	438		
Uncorrected Total	25	1812852			

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	5884.36	<.0001
Students	1	5884.36	<.0001

NOTE: The denominator degrees of freedom for the F tests is 24.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval	
Students	51.5589971	0.67213222	76.71	<.0001	50.1717843	52.9462098

NOTE: The denominator degrees of freedom for the t tests is 24.

Analysis of Estimable Functions

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Mean per Student	51.5589971	0.67213222	76.71	<.0001

Parameter	95% Confidence Interval	
Mean per Student	50.1717843	52.9462098

NOTE: The denominator degrees of freedom for the t tests is 24.

Example 26: Sick-leave accounting.

A firm has 80 retail stores in Florida and 140 in California. The firm wishes to estimate the average sick-leave time per employee and has decided to stratify upon state. Stores can be viewed as clusters with total sick-leave time determined from store records. A simple random sample of 10 stores was taken in California and 8 stores in Florida with results shown below. Estimate the average sick-leave time per employee and provide a 95% confidence interval estimate (problem 8.19, Scheaffer et al. 1996:328).

California	
Number of Employees N_i	**Total Days Sick-leave** y_i
16	51
8	32
4	11
3	10
12	33
17	39
24	61
30	37
21	40
9	41

Florida	
Number of Employees N_i	**Total Days Sick-leave** y_i
12	40
20	52
8	30
14	36
24	71
15	48
10	39
6	21

First, estimate the average sick-leave days per employee for each state separately.

California:

- Summary Statistics:

$$m = 10,\ \sum_{i=1}^{10} N_i = 144,\ \sum_{i=1}^{10} N_i^2 = 2776,\ \sum_{i=1}^{10} y_i = 355,\ \sum_{i=1}^{10} y_i^2 = 14,827,\ \sum_{i=1}^{10} N_i y_i = 5988,$$

$$M = 140,\ \bar{n} = \frac{144}{10} = 14.4.$$

- Population mean per element estimate (average number of days per employee):

$$\hat{\mu}_e = r = \frac{355}{144} = 2.4652778 \text{ days/empolyee.}$$

- Estimated variance for our estimate of the population mean per element:

$$\text{var}(\hat{\mu}_e) = \left(\frac{M-m}{M}\right)\frac{1}{\bar{N}^2 m}\sum_{i=1}^{m}(y_i - \tilde{\bar{y}}N_i)^2/(m-1)$$

$$= \left(\frac{140-10}{140}\right)\frac{1}{(14.4)^2 10}\frac{(14827 - 2(2.465)(5988) + (2.465)^2(2776))}{9} = 0.10818$$

$$\text{se}(\hat{\mu}_e) = 0.3289$$

- Approximate 95% confidence interval for the mean sick-leave days per employee:

$$B = 2\sqrt{\text{var}(\hat{\mu}_e)} = 2\sqrt{0.10818} = 0.6578$$

and so our interval is

$$\hat{\mu}_e \pm B \text{ or } (1.81, 3.12) \text{ days/employee.}$$

Florida:

- Summary Statistics:

$$m = 8,\ \sum_{i=1}^{8} N_i = 109,\ \sum_{i=1}^{8} N_i^2 = 1741,\ \sum_{i=1}^{8} y_i = 337,\ \sum_{i=1}^{8} y_i^2 = 15,807,\ \sum_{i=1}^{8} N_i y_i = 5204,$$

$$M = 80,\ \bar{n} = \frac{109}{8} = 13.625.$$

- Population mean per element estimate (average sick-leave days per employee):

$$\hat{\mu}_e = r = \frac{337}{109} = 3.0917 \text{ days/employee.}$$

- Estimated variance for our estimate of the population mean per element:

$$\text{var}(\hat{\mu}_e) = \frac{\left(1 - \frac{8}{80}\right)}{(13.625)^2 8}\frac{(15807 - 2(3.0917)(5204) + (3.0917)^2(1741))}{7} = 0.02339$$

- Approximate 95% confidence interval for the mean sick-leave days per employee:

$$B = 2\sqrt{\text{var}(\hat{\mu}_e)} = 2\sqrt{0.02339} = 0.3058$$

and so our interval is

$$\hat{\mu}_e \pm B \text{ or } (2.79, 3.40) \text{ days/employee.}$$

1. Use **separate ratios** estimators to estimate the population sick-leave days per employee. In order to use this technique, we need to know the total number of employees for all stores in each state. As an example, let's assume that these figures are easily available, say from budget records in the main office. These data (hypothetical) are added to the table below corresponding with the row for N_h.

From the previous analyses we have

	California	Florida
N_h	2000	1100
M_h	140	80
$\hat{\mu}_{eh}$	2.465	3.0917
$\text{var}(\hat{\mu}_{eh})$	0.10818	0.02339

Using formulas from stratified random sampling we find

$$\hat{\mu}_e = \frac{(2000)(2.465) + (1100)(3.0917)}{2000 + 1100} = 2.6873774$$

with

$$\text{var}(\hat{\mu}_e) = \frac{(2000)^2(0.10818) + (1100)^2(0.02339)}{(3100)^2} = 0.04797314$$

This yields a 95% bound on the error of estimation of

$$B = 2\sqrt{0.04797314} = 0.43805544$$

with 95% confidence interval of $(2.25, 3.13)$.

2. Use the **combined stratified population ratio** estimator to estimate the population mean sick-leave days per employee in the firm.

- Find the combined ratio estimator $r_C = \frac{\hat{\tau}_y}{\hat{\tau}_N}$ where $\hat{\tau}_y = \sum_{h=1}^{2} M_h \hat{\mu}_h$ and $\hat{\tau}_N = \sum_{h=1}^{2} M_h \hat{\mu}_N$.

$$\hat{\tau}_y = 80\left(\frac{337}{8}\right) + 140\left(\frac{355}{10}\right) = 8340 \text{ days}$$

and

$$\hat{\tau}_N = 80\left(\frac{109}{8}\right) + 140\left(\frac{144}{10}\right) = 3106 \text{ employees}$$

so that

$$r_C = \hat{\mu}_{eC} = \frac{8340}{3106} = 2.685 \text{ days per employee.}$$

- Compute a variance for our estimate.
  First we need the variances and covariances of y and N from each state in addition to several other summary statistics. These are listed in the table below.

Statistic	California	Florida
M_h	140	80
m_h	10	8
f_h	0.07142857	0.1
$\bar{n}$	14.4	13.625
s^2_{yh}	247.1667	230.125
s^2_{Nh}	78.0444	36.5536
s_{yNh}	97.3333	87.4821
$\hat{\rho}_h$	0.7008	0.9538

$$\text{var}(\hat{\mu}_{eC}) = \text{var}(r_C) = \sum_{h=1}^{L} \left(\frac{M_h}{M}\right)^2 \frac{(1-f_h)}{m_h \mu_{\hat{N}}^2} (s_{yh}^2 + r^2 s_{\hat{N}h}^2 - 2r s_{yNh})$$

$$= \left(\frac{140}{80+140}\right)^2 \frac{\left(1-\frac{10}{140}\right)}{10(14.118182)^2}(247.1667 + (2.685)^2(78.0444) - 2(2.685)(97.3333))$$

$$+ \left(\frac{80}{80+140}\right)^2 \frac{\left(1-\frac{8}{80}\right)}{8(14.118182)^2}(230.125 + (2.685)^2(36.5536) - (2)(2.685)(87.4821))$$

$$= 0.05416787 + 0.00178143 = 0.0559493$$

where $\mu_N \approx \frac{\hat{\tau}_N}{M} = \frac{3106}{220} = 14.118182$.

- An approximate 95% confidence interval on the population sick-leave days per employee would be

$$\hat{\mu}_{eC} \pm 2\sqrt{\text{var}(\hat{\mu})} \text{ or } 2.685 \pm 2\sqrt{0.0559493} \text{ or } 2.685 \pm 0.4730721$$

which yields the interval

$$(2.21, 3.16) \text{ days/employee.}$$

2.0.4 Sample Size Under Ratio Estimation

As before, treat the cluster totals as the data of interest, then the sample size formulas are the same as for ratio estimation under simple random sampling. Modify the formulas slightly for specifying a bound on the error of estimation of the mean per element rather than on the cluster mean.