5 RATIO AND REGRESSION ESTIMATION

5.1 Ratio Estimation

- Suppose the researcher believes an auxiliary variable (or covariate) $x$ is associated with the variable of interest $y$. Examples:
  - **Variable of interest**: the amount of lumber (in board feet) produced by a tree.
    **Auxiliary variable**: the diameter of the tree.
  - **Variable of interest**: the current number of farms per county in the United States.
    **Auxiliary variable**: the number of farms per county taken from the previous census.
  - **Variable of interest**: the income level of a person who is 40 years old.
    **Auxiliary variable**: the number of years of education completed by the person.

- Situation: we have bivariate $(X,Y)$ data and assume there is a positive proportional relationship between $X$ and $Y$. That is, on every sampling unit we take a pair of measurements and assume that $Y \approx BX$ for some constant $B > 0$.

- There are two cases that may be of interest to the researcher:
  1. To estimate the ratio of two population characteristics. The most common case is the population ratio $B$ of means or totals:

     $$B = \frac{\bar{y}_U}{\bar{x}_U}$$

  2. To use the relationship between $X$ and $Y$ to improve estimation of $t_y$ or $\bar{y}_U$.

- The sampling plan will be to take a SRS of $n$ pairs $(x_1, y_1), \ldots, (x_n, y_n)$ from the population of $N$ pairs. We will use the following notation:

  $$\bar{x}_U = \frac{\sum_{i=1}^{N} x_i}{N} \quad t_x = \sum_{i=1}^{N} x_i \quad \bar{y}_U = \frac{\sum_{i=1}^{N} y_i}{N} \quad t_y = \sum_{i=1}^{N} y_i$$

  $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \text{the sample mean of } x\text{'s.} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \text{the sample mean of } y\text{'s.}$$

5.1.1 Estimating $B$, $\bar{y}_U$, and $t_y$

**Case I: $t_x$ and $\bar{x}_U$ are known**

- We will first consider the estimation of $B$ assuming $t_x$ and $\bar{x}_U$ are known. The ratio estimator $\hat{B}$ is the ratio of the sample means and its estimated variance $\hat{V}(\hat{B})$ are

  $$\hat{B} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \quad \hat{V}(\hat{B}) = \left( \frac{N-n}{N \bar{x}^2} \right) \frac{s_e^2}{n}$$

  where

  $$s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{B}x_i)^2 = \frac{1}{n-1} \left( \sum_{i=1}^{n} y_i^2 - \hat{B}^2 \sum_{i=1}^{n} x_i^2 - 2\hat{B} \sum_{i=1}^{n} x_i y_i \right)$$
• If \( Y \approx BX \), then \( y_i \approx \hat{B}x_i \). Thus, \( \hat{B}x_i \) can be considered the predicted value of \( y_i \) from a line through the origin (with intercept=0 and the slope = \( \hat{B} \)).

• The distribution of \( \hat{B} \) is very complicated. For small samples, \( \hat{B} \) is likely to be skewed and is biased for \( B \). For large samples, the bias is negligible (very small) and the distribution of \( \hat{B} \) tends to be approximately normal.

• Multiplication of \( \hat{B} \) and \( \hat{V}(\hat{B}) \) by \( x_U \) and \( x^2_U \), respectively, in (47) yields the estimator \( \hat{y}_r \) for \( y_U \) and its estimated variance:

\[
\hat{y}_r = \hat{V}(\hat{y}_r) = \hat{V}(Bx_U) = \hat{x}_U^2 \hat{V}(\hat{B}) = \left( \frac{N-n}{N} \right) \left( \frac{\bar{x}_U}{\bar{x}} \right)^2 \frac{s^2_e}{n}
\]

• \( \hat{y}_r \) is called the **ratio estimator of the population mean**.

• By multiplying the formulas in (48) by \( N \) and \( N^2 \), respectively, we get the estimator \( \hat{t}_{y r} \) of \( t_y \) and the estimated variance:

\[
\hat{t}_{y r} = \hat{V}(\hat{t}_{y r}) = N(N-n) \left( \frac{\bar{x}_U}{\bar{x}} \right)^2 \frac{s^2_e}{n} = \left( \frac{N-n}{N} \right) \left( \frac{t_x}{\bar{x}} \right)^2 \frac{s^2_e}{n}
\]

• \( \hat{t}_{y r} \) is called the **ratio estimator of the population total**.

• If \( N \) is unknown but we know \( N \) is large relative to \( n \), then the f.p.c. \( (N-n)/N \approx 1 \). Some researchers will replace \( (N-n)/N \) with 1 in the variance formulas in (48) and (49).

• A second ratio estimator is the mean of the sample ratios \( = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i} \). Although this may be appealing, it is generally not used because its bias and mean square error can be large. We will not cover this estimator in this course.

• Later, we will also use bootstrapping techniques to estimate \( V(\hat{y}_r) \) and \( V(\hat{t}_{y r}) \).

**Case II: \( t_x \) and \( x_U \) are unknown**

• If \( t_x \) and \( x_U \) are unknown, it will not affect the estimator \( \hat{B} = \bar{y}/\bar{x} \). It will, however, affect the estimators \( \hat{y}_r \) and \( \hat{t}_{y r} \) that depend on \( t_x \) and \( x_U \).

• In such cases, it is common to replace \( t_x \) with \( N\bar{x} \) or replace \( x_U \) with \( \bar{x} \). This will yield:

\[
\hat{V}(\hat{y}_r) \approx \left( \frac{N-n}{N} \right) \frac{s^2_e}{n} \quad \quad \hat{V}(\hat{t}_{y r}) = N(N-n) \frac{s^2_e}{n}
\]

• When \( \bar{x} \) is larger than \( x_U \), \( \hat{V}(\hat{y}_r) \) and \( \hat{V}(\hat{t}_{y r}) \) tend to be too large as variance estimates. Similarly, when \( \bar{x} \) is smaller than \( x_U \), \( \hat{V}(\hat{y}_r) \) and \( \hat{V}(\hat{t}_{y r}) \) tend to be too small as variance estimates.
Example: Demonstration of Bias in Ratio Estimation

• There are $N = 4$ sampling units in the population:

<table>
<thead>
<tr>
<th>$x_i$ value</th>
<th>67</th>
<th>63</th>
<th>66</th>
<th>69</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$ value</td>
<td>68</td>
<td>62</td>
<td>64</td>
<td>70</td>
</tr>
</tbody>
</table>

• Let $n = 2$. The total abundances are $t_x = 265$ for $X$ and $t_y = 264$ for $Y$. Therefore, $B = 264/265 = 0.9962$. Also, $S^2_x = 6.250$ and $S^2_y \approx 13.3$.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Units</th>
<th>$\hat{t}_{yr}$</th>
<th>$\hat{V}(\hat{t}_{yr})$</th>
<th>$\hat{t}_{SRS}$</th>
<th>$\hat{V}(\hat{t}_{SRS})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,2</td>
<td>265</td>
<td>8</td>
<td>260</td>
<td>72</td>
</tr>
<tr>
<td>2</td>
<td>1,3</td>
<td>263.0075</td>
<td>18.0903</td>
<td>264</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>1,4</td>
<td>268.8971</td>
<td>0.0017</td>
<td>276</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>2,3</td>
<td>258.8372</td>
<td>1.7307</td>
<td>252</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>2,4</td>
<td>265</td>
<td>8</td>
<td>264</td>
<td>128</td>
</tr>
<tr>
<td>6</td>
<td>3,4</td>
<td>263.0370</td>
<td>18.2677</td>
<td>268</td>
<td>72</td>
</tr>
</tbody>
</table>

$E(\hat{t}_{yr}) = 263.963$  $E(\hat{V}(\hat{t}_{yr})) = 9.0151$

$V(\hat{t}_{yr}) = 10.981$

• Note that $\hat{V}(\hat{t}_{yr}) < \hat{V}(\hat{t}_{SRS})$ for all samples. However, the estimators $\hat{t}_{yr}$ and $\hat{V}(\hat{t}_{yr})$ are biased because $E(\hat{t}_{yr}) \neq t_y$ (263.96 $\neq$ 264), $E(\hat{V}(\hat{t}_{yr})) \neq V(\hat{t}_{yr})$ (9.0151 $\neq$ 10.9081).

5.1.2 Bias and MSE of Ratio Estimators

• The ratio estimators are biased. The bias occurs in ratio estimation because $E(\hat{y}/\hat{x}) \neq E(\bar{y})/E(\bar{x})$ (i.e., the expected value of the ratio $\neq$ the ratio of the expected values).

• When appropriately used, the reduction in variance from using the ratio estimator will offset the presence of bias. Also, for large samples, the estimators $t_{yr}$ and $\bar{y}_r$ will be approximately normally distributed.

• In your text, it is shown that

$$\text{Bias}(\hat{\bar{y}}_r) = [E(\hat{\bar{y}}_r) - \bar{y}_U] = -Cov(\hat{B}, \bar{x})$$

and, from this, it can be shown that

$$\frac{|\text{Bias}(\hat{\bar{y}}_r)|}{\sqrt{\hat{V}(\bar{y}_r)}} \leq \frac{\sqrt{\hat{V}(\bar{x})}}{\bar{x}_U} = \text{CV}(\bar{x})$$

where CV is the coefficient of variation for $\bar{x}$.

• It can also be shown that

$$\text{Bias}(\hat{\bar{y}}_r) \approx \left( \frac{N-n}{N} \right) \frac{1}{n\bar{x}_U} (BS^2_x - RS_xS_y)$$

where $R$ is the population correlation coefficient.

• Thus, the bias of $\hat{\bar{y}}_r$ (as well as $\hat{t}_r$ and $\hat{B}$) will be small if

  – the sample size $n$ is large
  – the sampling fraction $n/N$ is large
  – $\bar{x}_U$ is large
  – $S_x$ is small
  – the correlation coefficient $R$ is close to 1 (see Section 5.3)
• Assuming a small bias, both the variance and MSE can be approximated by

\[
V(\widehat{y}_r) \approx MSE(\widehat{y}_r) = \mathbb{E}[(\widehat{y}_r) - \overline{y}_U)^2] \approx \frac{N - n}{N} \frac{S_y^2 - 2BRyS_x + B^2S_x^2}{n} 
\]

• Thus, the MSE and variance of \( \widehat{B} \) will be small if
  
  – the sample size \( n \) is large
  – the sampling fraction \( n/N \) is large
  – \( \overline{x}_U \) is large
  – the deviations \( y_i - Bx_i \) are small
  – the correlation coefficient \( R \) is close to 1

**Example of Ratio Estimation:** A manager at a mill wants to estimate the total weight of dry wood (\( t_y \)) for a certain number of truckloads of 5-foot bundles of pulpwood (wood from recently cut trees). The process begins by

1. Weighing the total amount of pulpwood (\( t_x \)).
2. Randomly selecting a sample of \( n \) bundles from the trucks.
3. Recording the weight of the pulpwood (\( x_i \)) for each of the \( n \) bundles.
4. Removing the bark and drying the wood from the \( n \) bundles.
5. Recording the weight of the dry wood (\( y_i \)) for each of the \( n \) bundles.

Suppose a sample of \( n = 30 \) bundles was taken from a total of \( N = 800 \) bundles.

The unit of measurement is pounds. Here are summary statistics:

\[
\begin{align*}
\sum_{i=1}^{30} x_i &= 3316 \\
\sum_{i=1}^{30} y_i &= 1802 \\
\sum_{i=1}^{30} x_i y_i &= 214,738 \\
\sum_{i=1}^{30} x_i^2 &= 392,440 \\
\sum_{i=1}^{30} y_i^2 &= 118,360 \\
\end{align*}
\]

\( t_x = \sum_{i=1}^{800} x_i = 89420 \)

Use ratio estimation to estimate the total amount of dry wood (\( t_y \)), the mean amount of drywood per bundle (\( \overline{y}_U \)). Also calculate the standard errors of these estimates.
5.1.3 Confidence Intervals for $B$, $t_y$, and $t_y$

- For large samples, approximate $100(1 - \alpha)\%$ confidence intervals for $B$, $\bar{y}_U$, and $t_y$ are:

\[
\hat{B} \pm z^* \sqrt{\hat{V}(\hat{B})} \quad \hat{\bar{y}}_r \pm z^* \sqrt{\hat{V}(\hat{\bar{y}}_r)} \quad \hat{t}_{yr} \pm z^* \sqrt{\hat{V}(\hat{t}_{yr})}
\]  

(50)

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

- For smaller samples, approximate $100(1 - \alpha)\%$ confidence intervals for $B$ and $t_y$ are:

\[
\hat{B} \pm t^* \sqrt{\hat{V}(\hat{B})} \quad \hat{\bar{y}}_r \pm t^* \sqrt{\hat{V}(\hat{\bar{y}}_r)} \quad \hat{t}_{yr} \pm t^* \sqrt{\hat{V}(\hat{t}_{yr})}
\]  

(51)

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

- General rule: a normal approximation can be used if (i) $n \geq 30$, (ii) the sampling fraction $n/N \leq .25$, and (iii) the coefficients of variation $C_X = \frac{S_X}{\bar{x}_U}$ and $C_Y = \frac{S_Y}{\bar{y}_U}$ are $< .10/\sqrt{n}$.

- **Example:** Find 95% confidence intervals for $t_y$, $\bar{y}_U$, and $B$ for the pulpwood and drywood example.

### 5.2 Software to Perform Ratio Estimation

**EXAMPLE:** This example comes from a data set from the textbook by Lohr (2010).

The United States government conducts a Census of Agriculture every 5 years. Data is collected on all farms. In this census, a farm is defined to be any place from which US$1000 or more of agricultural products were produced and sold. The study population is restricted to the 50 American states (and excludes territories like Puerto Rico and Guam). The data recorded in the census includes information on farm size and yield of different crops (corn, wheat, etc.).

The data on the handout contains summary data from a random sample of $n = 300$ counties in the United States for the years 1982, 1987, and 1992. There are $N = 3078$ counties in the 50 states in the United States. By chance, this sample does not contain any counties from Alaska, Arizona, Connecticut, Delaware, Hawaii, Rhode Island, Utah, or Wyoming.

**Obs** is the label for the sample unit ($\text{Obs} = 1, 2, \ldots, 300$).

**ACRES92, ACRES87, and ACRES82** are the numbers of acres devoted to farms in 1992, 1987, and 1982 for that county. ($1 \text{ acre} \approx 4040 \text{ m}^2$)

**F92, F87, and F82** are the numbers of farms in 1992, 1987, and 1982 for that county.

**LF92, LF87, and LF82** are the numbers of large farms ($\geq 1000$ acres) in 1992, 1987, and 1982, respectively, for that county.

**SF92, SF87, and SF82** are the numbers of small farms ($\leq 9$ acres) in 1992, 1987, and 1982, respectively, for that county.

**Region** represents one of four an assigned geographical regions of the United States (W=West, S=South, NE=Northeast NC=North central).

- The following table is a summary of the number of counties ($N_i$) in each state ($i$).
The graph is a scatterplot of Acres92 ($y$) vs Acres87 ($x$). The plot suggests a proportional relationship between $y$ and $x$ (a positive linear relationship with 0 intercept). Therefore, ratio estimation should be a useful procedure for estimating $\bar{y}_U$ or $t_y$.

It is known that $t_x = 964, 470, 625$ total farm acres in the United States in the year 1987. Therefore, $\bar{x}_U = t_x/3078 \approx 313343.283$ farm acres per county.

### 5.2.1 Ratio Estimation Using R

- **CASE 1**: Estimating $\hat{B}$.
- **CASE 2**: Estimating $\bar{y}_U$ when $t_x$ and $\bar{x}_U$ are known.
- **CASE 3**: Estimating $t_y$ when $t_x$ and $\bar{x}_U$ are known.
- **CASE 4**: Estimating $\bar{y}_U$ when $t_x$ and $\bar{x}_U$ are unknown.
- **CASE 5**: Estimating $t_y$ when $t_x$ and $\bar{x}_U$ are unknown.
R code for ratio estimation

library(survey)
source("c:/courses/st446/rcode/confintt.r")
# In Excel, save your spreadsheet as a text tab-delimited file
# If variable names are in row 1, then use header=T)
ratio <- read.table("c://courses/st446/Rcode/agsrs.txt",header=T)
N=3078     # population size
n=300      # sample size
tx=964470625    # X population total if known: for Case 3
mux = tx/N    # X population mean if known: for Case 2
ratio_ttl <- tx*ratio$ACRES92
ratio_mn <- ratio_ttl/N
mnhx = mean(ratio$ACRES87)  # estimated X pop. mean if unknown: Case 4
thatx = N*mnhx          # estimated X pop. total if unknown: Case 5
ratio_umn = mnhx*ratio$ACRES92
ratio_uttl = N*ratio_umn
fpc <- c(rep(N,n))

ratio <- cbind(ratio,fpc,ratio_ttl,ratio_mn,ratio_uttl,ratio_umn)
ratio <- data.frame(ratio)
# Create the sampling design
agdsgn <- svydesign(data=ratio,id=~1,fpc=~fpc)

# Estimation of the ratio (Case 1)
agratio <- svyratio(~ACRES92,~ACRES87,design=agdsgn)
confint.t(agratio,tdf=n-1,level=.95)

# Estimation of the y population mean (when tx is known): Case 2
agratio_mean <- svyratio(~ratio_mn,~ACRES87,design=agdsgn)
confint.t(agratio_mean,tdf=n-1,level=.95)

# Estimation of the y population total (when tx is known): Case 3
agratio_total <- svyratio(~ratio_ttl,~ACRES87,design=agdsgn)
confint.t(agratio_total,tdf=n-1,level=.95)

# Estimation of the y population mean (when tx is unknown): Case 4
agratio_umean <- svyratio(~ratio_umn,~ACRES87,design=agdsgn)
confint.t(agratio_umean,tdf=n-1,level=.95)

# Estimation of the y population total (when tx is unknown): Case 5
agratio_utotal <- svyratio(~ratio_uttl,~ACRES87,design=agdsgn)
confint.t(agratio_utotal,tdf=n-1,level=.95)
R output for ratio estimation

> # Estimation of the ratio  
>  
> mean( ACRES92/ACRES87 ) = 0.98657  
> SE( ACRES92/ACRES87 ) = 0.00575  
>  
> Two-Tailed CI for ACRES92/ACRES87 where alpha = 0.05 with 299 df  
> 2.5 %  97.5 %  
> 0.97525  0.99788  
>  
> > # Estimation of the y population mean (when tx is known) CASE 2  
>  
> mean( ratio_mn/ACRES87 ) = 309133.59028  
> SE( ratio_mn/ACRES87 ) = 1801.87199  
>  
> Two-Tailed CI for ratio_mn/ACRES87 where alpha = 0.05 with 299 df  
> 2.5 %  97.5 %  
> 305587.63292  312679.54763  
>  
> > # Estimation of the y population total (when tx is known) CASE 3  
>  
> mean( ratio_ttl/ACRES87 ) = 951513190.87092  
> SE( ratio_ttl/ACRES87 ) = 5546161.99416  
>  
> Two-Tailed CI for ratio_ttl/ACRES87 where alpha = 0.05 with 299 df  
> 2.5 %  97.5 %  
> 940598734.13317  962427647.60867  
>  
> > # Estimation of the y population mean (when tx is unknown) CASE 4  
>  
> mean( ratio_umn/ACRES87 ) = 297897.04667  
> SE( ratio_umn/ACRES87 ) = 1736.37664  
>  
> Two-Tailed CI for ratio_umn/ACRES87 where alpha = 0.05 with 299 df  
> 2.5 %  97.5 %  
> 294479.97956  301314.11378  
>  
> > # Estimation of the y population total (when tx is unknown) CASE 5  
>  
> mean( ratio_uttl/ACRES87 ) = 916927109.64000  
> SE( ratio_uttl/ACRES87 ) = 5344567.30153  
>  
> Two-Tailed CI for ratio_uttl/ACRES87 where alpha = 0.05 with 299 df  
> 2.5 %  97.5 %  
> 906409377.07901  927444842.20099
5.2.2 Bootstrapping Ratio Estimates Using R

- Bootstrapping can only be used to estimate $t_y$ and $\overline{y}_U$ when $t_x$ and $\overline{x}_U$ are known.

- Why? Replacing $t_x$ and $\overline{x}_U$ with $N\overline{x}$ and $\overline{x}$ in each bootstrap sample is equivalent to the SRS bootstrap of $\overline{y}$ and $N\overline{y}$. Thus, bootstrap estimates of $t_y$ and $\overline{y}_U$ ignore all information about $x$.

- The following R code is for estimation of $B$ and for estimation of $t_y$ and $\overline{y}_U$ when $t_x$ and $\overline{x}_U$ are known.

R code for bootstrapping ratio estimates

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

indata <- read.table("c://courses/st446/Rcode/agsrs.txt",header=T)

y <- indata$ACRES92
x <- indata$ACRES87
ratio <- cbind(x,y)
ratio <- data.frame(ratio)

N=3068 # population size
Brep = 20000

tx=964470625 # X population total if known
mux = tx/N # X population mean if known

# Bootstrap the sample ratio
sampratio <- function(ratio,i) mean(y[i]/mean(x[i]))
bootratio <- boot(data=ratio,statistic=sampratio,R=Brep)
boot.ci(bootratio,conf=.95,type=c("norm","perc"))
par(mfrow=c(2,1))
hist(bootratio$t,main="Bootstrap Sample Ratios")
plot(ecdf(bootratio$t),main="Empirical CDF of Bootstrap Ratios")

# Bootstrap the estimates of the y population mean (tx known)
sampmean <- function(ratio,i) mux*mean(y[i])/mean(x[i])
bootmean <- boot(data=ratio,statistic=sampmean,R=Brep)
boot.ci(bootmean,conf=.95,type=c("norm","perc"))
par(mfrow=c(2,1))
hist(bootmean$t,main="Bootstrap y Population Mean Estimates")
plot(ecdf(bootmean$t),main="Empirical CDF of Bootstrap Mean Estimates")

# Bootstrap the estimates of the y population total (tx known)
samptotal <- function(ratio,i) tx*mean(y[i])/mean(x[i])
boottotal <- boot(data=ratio,statistic=samptotal,R=Brep)
boot.ci(boottotal,conf=.95,type=c("norm","perc"))
par(mfrow=c(2,1))
hist(boottotal$t,main="Bootstrap y Population Total Estimates")
plot(ecdf(boottotal$t),main="Empirical CDF of Bootstrap Total Estimates")
```
R output for bootstrapping ratio estimates

Bootstrap Statistics:
- original bias std. error
- t1* 0.9865652 -2.797441e-06 0.005980109 <-- for ratio B

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 20000 bootstrap replicates

Intervals:
- Level Normal Percentile
- 95% (0.9748, 0.9983) (0.9746, 0.9981)

Bootstrap Statistics:
- original bias std. error
- t1* 310141.2 -4.233312 1908.739 <-- for y mean

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 20000 bootstrap replicates

Intervals:
- Level Normal Percentile
- 95% (306404, 313886) (306354, 313862)

Bootstrap Statistics:
- original bias std. error
- t1* 951513191 54457.72 5772867 <-- for y total

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 20000 bootstrap replicates

Intervals:
- Level Normal Percentile
- 95% (940144121, 962773345) (940039313, 962538574)

Bootstrap Sample Ratios

Empirical CDF of Bootstrap Ratios
5.2.3 Ratio Estimation Using SAS Proc Surveymeans (Supplemental)

- **CASE 1**: Because \( \hat{B} \) and the s.e.(\( \hat{B} \)) do not depend on knowing any population values, the default output for estimating the population ratio \( B \) is correct in the output. The analysis is produced by the first block of code.

- To get estimates \( \hat{y}_r \) and \( \hat{t}_y \) with the correct standard errors and \( t \)-based confidence intervals, the simplest way I found to do this is just to scale the \( y \) values (which will allow us to avoid using the "weight" command).

- **CASE 2**: Estimating \( \overline{y}_U \) when \( t_x \) and \( \overline{x}_U \) are known.
  - Replace the sampled \( y \)-values with \( (\overline{x}_U y) \)-values. Then consider the ratio \( (\overline{x}_U \overline{y})/\overline{x} = \hat{B}_{\overline{x}_U} \) which is the ratio estimator of \( \overline{y}_U \).

- **CASE 3**: Estimating \( t_y \) when \( t_x \) and \( \overline{x}_U \) are known.
  - Replace the sampled \( y \)-values with \( (t_x \overline{y}) \)-values. Then consider the ratio \( (t_x \overline{y})/\overline{x} = \hat{B}_{t_x} \) which is the ratio estimator of \( t_y \).

- **CASE 4**: Estimating \( \overline{y}_U \) when \( t_x \) and \( \overline{x}_U \) are unknown.
  - Replace the sampled \( y \)-values with \( (\overline{x} y) \)-values. Then consider the ratio \( (\overline{x} \overline{y})/\overline{x} = \hat{B}_{t_x} \) which is the ratio estimator of \( \overline{y}_U \) when \( t_x \) and \( \overline{x}_U \) are unknown.

- **CASE 5**: Estimating \( t_y \) when \( t_x \) and \( \overline{x}_U \) are unknown.
  - Replace the sampled \( y \)-values with \( (N \overline{x} y) \)-values. Then consider the ratio \( (N \overline{x} \overline{y})/\overline{x} = \hat{B}_{t_x} \) which is the ratio estimator of \( t_y \) when \( t_x \) and \( \overline{x}_U \) are unknown.

- For Case 1, the `ratio clm alpha=.05` options produce the ratio estimate \( \hat{B} \), its standard error, and a 95% confidence interval for \( B \).

- For Cases 2 and 3, the `ratio clm alpha=.05` options produce ratio-based estimates for \( \overline{y}_U \) and \( t_y \), their standard errors, and 95% confidence intervals when \( t_x \) and \( \overline{x}_U \) are known.

- For Cases 4 and 5, the `ratio clm alpha=.05` options produce ratio-based estimates for \( \overline{y}_U \) or \( t_y \), their standard errors, and 95% confidence intervals when \( t_x \) and \( \overline{x}_U \) are unknown.

### The SAS Proc Surveymeans code for ratio estimation

```sas
DATA ratioest;
  INFILE 'C:\COURSES\st446\SASsurv\agsrs.dat';
  FORMAT county $char14.;
  INPUT i county $ st $ acres92 acres87 acres82 F92 F87 F82
    LF92 LF87 LF82 SF92 SF87 SF82 region $ @@;
  KEEP acres92 acres87 acres82;
  *** pick the variables you want for the ratio ;

DATA ratioest; SET ratioest;
  y = acres92;
  x = acres87;
  *** enter population information if it is known;
  N = 3078;            ** enter the population size ;
  nn= 300;             ** enter the sample size ;
  taux = 964470625;    ** enter tau_x if tau_x is known ;
  mux = taux/N;        ** enter mu_x if mu_x is known ;
  flag = 1;
```

---

109
*** calculate sum of sample x values (if taux and mux are unknown) ***;

PROC MEANS DATA = ratioest MEAN NOPRINT;
   VAR x; OUTPUT OUT = sset MEAN = xbar;
DATA sset; SET sset; flag=1; KEEP flag xbar;

DATA ratioest; MERGE ratioest sset; BY flag;

*** scale the y values for estimation of a total or mean of y ***;
   y_total = acres92*taux; *** known case ***;
   y_mean = acres92*mux;
   y_utotal = acres92*N*xbar; *** unknown case ***;
   y_umean = acres92*xbar;

PROC SURVEYMEANS data=ratioest total=3078 ratio clm alpha=.05; +-------
   var y x; | Case 1
   ratio y / x; +-------
title 'Ratio Estimation of the Ratio B';

PROC SURVEYMEANS data=ratioest total=3078 ratio clm alpha=.05; +-------
   var x y; | Case 2
   ratio y_mean / x; +-------
title 'Ratio Estimation of the y Mean --- known x mean and total';

PROC SURVEYMEANS data=ratioest total=3078 ratio clm alpha=.05; +-------
   var x y; | Case 3
   ratio y_total / x; +-------
title 'Ratio Estimation of the y Total --- known x mean and total';

PROC SURVEYMEANS data=ratioest total=3078 ratio clm alpha=.05; +-------
   var x y; | Case 4
   ratio y_umean / x; +-------
title 'Ratio Estimation of the y Mean --- unknown x mean and total';

PROC SURVEYMEANS data=ratioest total=3078 ratio clm alpha=.05; +-------
   var x y; | Case 5
   ratio y_utotal / x; +-------
title 'Ratio Estimation of the y Total --- unknown x mean and total';

RUN;

The SAS Proc Surveymeans output for ratio estimation

(OUTPUT FOR CASE 1)

Ratio Estimation of the Ratio B

The SURVEYMEANS Procedure
Data Summary

Number of Observations 300

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>297897</td>
<td>18898</td>
<td>260706.257 335087.836</td>
</tr>
<tr>
<td>x</td>
<td>301954</td>
<td>18914</td>
<td>264732.959 339174.488</td>
</tr>
</tbody>
</table>

110
### Ratio Analysis

<table>
<thead>
<tr>
<th>Numerator</th>
<th>Denominator</th>
<th>Ratio</th>
<th>Std Err</th>
<th>95% CL for Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>x</td>
<td>0.986565</td>
<td>0.005750</td>
<td>0.97524871 - 0.99788176</td>
</tr>
</tbody>
</table>

(OUTPUT FOR CASE 2)

Ratio Estimation of the y Mean --- known x mean and total

The SURVEYMEANS Procedure

Data Summary

Number of Observations 300

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>301954</td>
<td>18914</td>
<td>264733 - 339174</td>
</tr>
<tr>
<td>y</td>
<td>297897</td>
<td>18898</td>
<td>260706 - 335088</td>
</tr>
<tr>
<td>y_mean</td>
<td>93344038591</td>
<td>5921697487</td>
<td>8.16906E10 - 1.04998E11</td>
</tr>
</tbody>
</table>

### Ratio Analysis

<table>
<thead>
<tr>
<th>Numerator</th>
<th>Denominator</th>
<th>Ratio</th>
<th>Std Err</th>
<th>95% CL for Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>y_mean</td>
<td>x</td>
<td>309134</td>
<td>1801.871993</td>
<td>305587.633 - 312679.548</td>
</tr>
</tbody>
</table>

(OUTPUT FOR CASE 3)

Ratio Estimation of the y Total --- known x mean and total

The SURVEYMEANS Procedure

Data Summary

Number of Observations 300

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>301954</td>
<td>18914</td>
<td>264733 - 339174</td>
</tr>
<tr>
<td>y</td>
<td>297897</td>
<td>18898</td>
<td>260706 - 335088</td>
</tr>
<tr>
<td>y_total</td>
<td>2.8731295E14</td>
<td>1.8226985E13</td>
<td>2.51444E14 - 3.23182E14</td>
</tr>
</tbody>
</table>

### Ratio Analysis

<table>
<thead>
<tr>
<th>Numerator</th>
<th>Denominator</th>
<th>Ratio</th>
<th>Std Err</th>
<th>95% CL for Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>y_total</td>
<td>x</td>
<td>951513191</td>
<td>5546162</td>
<td>940598734 - 962427648</td>
</tr>
</tbody>
</table>
Ratio Estimation of the y Mean --- unknown x mean and total

The SURVEYMEANS Procedure

Data Summary

Number of Observations 300

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>301954</td>
<td>18914</td>
<td>264733 - 339174</td>
</tr>
<tr>
<td>y</td>
<td>297897</td>
<td>18898</td>
<td>260706 - 335088</td>
</tr>
<tr>
<td>y_umean</td>
<td>89951122411</td>
<td>5706452641</td>
<td>7.87212E10 - 1.01181E11</td>
</tr>
</tbody>
</table>

Ratio Analysis

<table>
<thead>
<tr>
<th>Numerator</th>
<th>Denominator</th>
<th>Ratio</th>
<th>Std Err</th>
<th>95% CL for Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>y_umean</td>
<td>x</td>
<td>297897</td>
<td>1736.376641</td>
<td>294479.980 - 301314.114</td>
</tr>
</tbody>
</table>

(OUTPUT FOR CASE 5)

Ratio Estimation of the y Total --- unknown x mean and total

The SURVEYMEANS Procedure

Data Summary

Number of Observations 300

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>301954</td>
<td>18914</td>
<td>264733 - 339174</td>
</tr>
<tr>
<td>y</td>
<td>297897</td>
<td>18898</td>
<td>260706 - 335088</td>
</tr>
<tr>
<td>y_utotal</td>
<td>2.7686955E14</td>
<td>1.7564461E13</td>
<td>2.42304E14 - 3.11435E14</td>
</tr>
</tbody>
</table>

Ratio Analysis

<table>
<thead>
<tr>
<th>Numerator</th>
<th>Denominator</th>
<th>Ratio</th>
<th>Std Err</th>
<th>95% CL for Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>y_utotal</td>
<td>x</td>
<td>916927110</td>
<td>5344567</td>
<td>906409377 - 927444842</td>
</tr>
</tbody>
</table>

5.3 \( \hat{y}_U \) vs \( \hat{y}_r \) or \( \hat{t} \) vs \( \hat{t}_{yr} \) Which is better? SRS or Ratio Estimation?

- Let \( S_x \) and \( S_y \) be the population standard deviations of \( X \) and \( Y \). Let \( S_{xy} \) be the population covariance between \( X \) and \( Y \). The population correlation coefficient

\[
R = \frac{S_{xy}}{S_y S_x}
\]

where

\[
S_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{N - 1}
\]
• It can be shown that approximations for the true population variances and MSEs of \( \hat{t}_{yr} \) and \( \hat{y}_r \) are

\[
\begin{align*}
MSE(\hat{t}_{yr}) &\approx V(\hat{t}_{yr}) \approx \frac{N(N-n)}{n} (S_y^2 - 2BRS_yS_x + B^2S_x^2) \\
MSE(\hat{y}_r) &\approx V(\hat{y}_r) \approx \frac{N-n}{Nn} (S_y^2 - 2BRS_yS_x + B^2S_x^2)
\end{align*}
\]

• Thus, these variances will be smaller as \( R \) approaches 1. Or, equivalently, the stronger the positive correlation, the smaller the variance.

• If the researcher wants to estimate \( t_y \) or \( \bar{y}_U \), the main sampling question is ‘When is worth the additional effort and expense to collect information about \( X \) instead of just using a SRS estimator \( \hat{y}_U \) or \( \hat{t} \) which does not require knowledge about \( X \)?’

• The answer requires looking at the coefficient of variation for both \( X \) and \( Y \).

\[
C_X = \quad C_Y =
\]

• It can be shown that if \( R > \frac{1}{2} \frac{C_X}{C_Y} \), then the variance of the ratio estimator is smaller than the variance of the SRS estimator.

• Because the maximum value of \( R \) is 1, if we have \( C_X > 2C_Y \), then the variance of the ratio estimator must be larger than the variance of the SRS estimator. Thus, when \( C_X > 2C_Y \), the SRS estimator is better (more efficient) than the ratio estimator.

• Because \( C_X \) and \( C_Y \) are unknown, we would calculate the sample (Pearson) correlation coefficient \( r \) and the sample coefficients of variation (\( \hat{C}_x \) and \( \hat{C}_y \)) to check if these conditions are met. The formulas are:

\[
\hat{C}_x = \quad \hat{C}_y = \quad r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{(n-1)s_x s_y}
\]

where \( s_x \) and \( s_y \) are the sample standard deviations of the \( x \) and \( y \) observations.

• In summary, if the following conditions hold, using ratio estimators can provide a substantial improvement over the SRS estimators:

1. You must be able to simultaneously observe \( X \) and \( Y \) values that are ‘roughly proportional’ to each other. That is, there is a strong positive linear relationship between \( Y \) and \( X \) that passes through the origin (zero intercept).

2. The coefficient of variation for \( X \) should not be substantially larger than the coefficient of variation for \( Y \).

3. The population total \( t_x \) or population mean \( \bar{x}_U \) should be known.

• If there is a linear relationship between \( Y \) and \( X \) and the intercept is not zero or the correlation between \( X \) and \( Y \) is negative, then a regression estimator should be considered.
5.4 Estimation in Domains (or Subpopulations)

- It is common to want estimates of a mean or total for subpopulations. The subpopulations are called **domains**.

- For example, in the previous example, we may want estimates for each of the four regions (W, S, NE, and NC). Each region is an example of a domain (or subpopulation).

- Let $U_d$ be the set of population units in domain $d$ and let $N_d$ be the number of population units in domain $d$. The domain total and domain mean for domain $d$ are

\[ t_{yd} = \sum_{i \in U_d} y_i \quad \bar{y}_{U_d} = t_{yd} / N_d = \left( \sum_{i \in U_d} y_i \right) / N_d \]

- Let $S_d$ be the set of sample units in domain $d$ and let $n_d$ be the number of sample units in domain $d$. Natural estimators for the domain mean $\bar{y}_{U_d}$ and the domain total $t_{yd}$ are

\[ \hat{y}_{U_d} = \frac{\sum_{i \in S_d} y_i}{n_d} = \bar{y}_d \quad \hat{t}_{yd} = \frac{N_d}{n_d} \left( \sum_{i \in S_d} y_i \right) = N_d \bar{y}_d \]

- $\bar{y}_d$ looks like the estimator $\bar{y}_U = \bar{y}$ for a SRS of size $n_d$, and $N_d \bar{y}_d$ looks like the estimator $t = N \bar{y}$ for a SRS of size $n_d$. This suggests that we should be able to apply the variance formulas for SRS from Section 2 of the notes. But we cannot! Why?

- For a SRS in Section 2, the sample size $n$ is fixed. For a domain (or subpopulation), the sample size $n_d$ is a random variable. That is, if we took different random samples of size $n$, we would get different values for $n_d$.

- Because $n_d$ is a random variable, we cannot use the SRS variance formulas from Section 2. To find the variance $V(\bar{y}_d)$, we need to see that $\bar{y}_d$ is a ratio estimator.

- Let $u_i = \begin{cases} y_i & \text{if } i \in U_d \\ 0 & \text{if } i \notin U_d \end{cases}$, $x_i = \begin{cases} 1 & \text{if } i \in U_d \\ 0 & \text{if } i \notin U_d \end{cases}$ for $i = 1, 2, \ldots, N$.

- Notation: Let $\bar{y}_{U_d}, \bar{x}_{U_d},$ and $\bar{x}_{U_d}$ be the domain means of the $y$, $x$, and $u$ values, respectively. Then,

\[ \bar{x}_{U_d} = \left( \sum_{i=1}^{N} x_i \right) / N = \frac{\sum_{i \in U_d} x_i + \sum_{i \notin U_d} x_i}{N} = \frac{\sum_{i \in U_d} 1 + \sum_{i \notin U_d} 0}{N} = \frac{N_d}{N} \]

Let $B_d = \frac{\bar{u}_{U_d}}{\bar{x}_{U_d}} = \frac{(\sum_{i=1}^{N} u_i) / N}{(\sum_{i=1}^{N} x_i) / N} = \frac{\sum_{i=1}^{N} u_i}{\sum_{i=1}^{N} x_i}$

\[ = \frac{\sum_{i \in U_d} u_i + \sum_{i \notin U_d} u_i}{\sum_{i \in U_d} x_i + \sum_{i \notin U_d} x_i} = \frac{\sum_{i \in U_d} u_i + \sum_{i \notin U_d} 0}{\sum_{i \in U_d} 1 + \sum_{i \notin U_d} 0} = \frac{\sum_{i \in U_d} u_i}{\sum_{i \in U_d} 1} = \frac{\sum_{i \in U_d} y_i}{N_d} = \bar{y}_{U_d} \]
• Let $S$ be the set of $n$ SRS units, and let $S_d \subset S$ be the set of $n_d$ SRS units in domain $d$. We can estimate domain ratio $B_d = \bar{y}_{Ud}$ with the ratio of domain sample means:

$$\hat{B}_d = \frac{\overline{u}_d}{\overline{x}_d} = \frac{\left(\sum_{i \in S} u_i / n\right)}{\left(\sum_{i \in S} x_i / n\right)} = \frac{\sum_{i \in S_d} u_i}{\sum_{i \in S_d} x_i}$$

$$= \frac{\sum_{i \in S_d} y_i + \sum_{i \notin S_d} 0}{\sum_{i \in S_d} 1 + \sum_{i \notin S_d} 0}$$

$$= \frac{\sum_{i \in S_d} y_i}{\sum_{i \in S_d} 1} = \frac{\sum_{i \in S_d} y_i}{n_d} = \bar{y}_d$$

• $\hat{B}_d = \bar{y}_d$ is the ratio estimator of $\bar{y}_{Ud}$. That is, $\hat{\bar{y}}_{Ud} = \bar{y}_d$. We now use the variance formula in (47) for ratio estimation:

$$\hat{V}(\bar{y}_d) = \hat{V}(B_d) = \left(\frac{N - n}{N \bar{x}_d^2}\right) \frac{s_u^2}{n} = \left(\frac{N - n}{N(N_d/N)^2}\right) \frac{s_u^2}{n} = \left(\frac{N - n}{N}\right) \left(\frac{N_d}{N}\right)^2 \frac{s_u^2}{n}$$

(52)

where $s_u^2 = \frac{1}{n-1} \sum_{i \in S} (u_i - \hat{B}_d x_i)^2 = \frac{1}{n-1} \left(\sum_{i \in S} u_i^2 + \hat{B}_d^2 \sum_{i \in S} x_i^2 - 2 \hat{B}_d \sum_{i \in S} x_i u_i\right)$

$$= \frac{1}{n-1} \left(\sum_{i \in S_d} u_i^2 + \hat{B}_d^2 \sum_{i \in S_d} x_i^2 - 2 \hat{B}_d \sum_{i \in S_d} x_i u_i\right)$$

$$= \frac{1}{n-1} \left(\sum_{i \in S_d} y_i^2 + \hat{B}_d^2 \sum_{i \in S_d} (1) - 2 \hat{B}_d \sum_{i \in S_d} (1) y_i\right)$$

$$= \frac{1}{n-1} \left(\sum_{i \in S_d} y_i^2 + n_d \hat{B}_d^2 - 2 \hat{B}_d \sum_{i \in S_d} y_i\right)$$

• When $N_d$ is unknown, replace $N_d$ with estimate $\hat{N}_d = N n_d / n$ (because $n_d / n \approx N_d / N$). This is what is done by default in SAS Proc SurveyMeans.

**EXAMPLE of Domain Estimation:** Suppose we are interested in estimating the mean acres per farm for the states in each region. The regions are the domains (or subpopulations). The table contains summary values for the proportion of the sample ($\bar{x}_d = n_d/300$) from domain $d$ and the proportion of population units ($\bar{y}_{Ud} = N_d/3078$) in domain $d$:

<table>
<thead>
<tr>
<th>$d$</th>
<th>$n_d$</th>
<th>$\sum_{i \in S_d} y_i$</th>
<th>$\bar{y}_d$</th>
<th>$\bar{x}_d$</th>
<th>$\hat{N}_d$</th>
<th>$\bar{y}_{Ud}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>107</td>
<td>37,481,245</td>
<td>.356</td>
<td>.3424</td>
<td>1054</td>
<td>.3424</td>
</tr>
<tr>
<td>NE</td>
<td>24</td>
<td>1,727,300</td>
<td>.08</td>
<td>.0715</td>
<td>220</td>
<td>.0715</td>
</tr>
<tr>
<td>S</td>
<td>130</td>
<td>26,812,026</td>
<td>.43</td>
<td>.4490</td>
<td>1382</td>
<td>.4490</td>
</tr>
<tr>
<td>W</td>
<td>39</td>
<td>23,348,543</td>
<td>.13</td>
<td>.1371</td>
<td>422</td>
<td>.1371</td>
</tr>
<tr>
<td>Total</td>
<td>$n = 300$</td>
<td>$\sum_{i \in S_d} y_i$</td>
<td>$\bar{y}_d$</td>
<td>$\bar{x}_d$</td>
<td>$\hat{N}_d$</td>
<td>$\bar{y}_{Ud}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>268,951,872</td>
<td>.379</td>
<td>.356</td>
<td>1054</td>
<td>.3424</td>
</tr>
</tbody>
</table>

Note that the proportion of the sample from domain $d$ is close to the actual proportion of population units in domain $d$ ($\bar{x}_d \approx \bar{y}_{Ud}$).
5.4.1 Using R to Perform a Domain Analysis

R code for Domain Analysis

library(survey)
source("c:/courses/st446/rcode/confintt.r")
domain <- read.table("c://courses/st446/Rcode/agsrs.txt",header=T)

N=3078 # population size
n=300 # sample size
fpc <- c(rep(N,n))

domaindt <- cbind(domain,fpc)
domaindat <- data.frame(domaindt)
#domaindat

# Create the sampling design
domain_dsgn <- svydesign(data=domaindat, id=~1, fpc="fpc")
domain_dsgn

# Estimation of domain totals

# Domain = NC
esttotal <- svytotal(~ACRES92,subset(domain_dsgn,REGION=="NC"))
esttotal
cconfint(esttotal,df=n-1)

# Domain = NE
nesttotal <- svytotal(~ACRES92,subset(domain_dsgn,REGION=="NE"))
esttotal
cconfint(esttotal,df=n-1)

# Domain = S
esttotal <- svytotal(~ACRES92,subset(domain_dsgn,REGION=="S"))
esttotal
cconfint(esttotal,df=n-1)

# Domain = W
esttotal <- svytotal(~ACRES92,subset(domain_dsgn,REGION=="W"))
esttotal
cconfint(esttotal,df=n-1)

# Estimation of domain means

# Domain = NC
estmean <- svymean(~ACRES92,subset(domain_dsgn,REGION=="NC"))
estmean
cconfint(estmean,df=n-1)

# Domain = NE
estmean <- svymean(~ACRES92,subset(domain_dsgn,REGION=="NE"))
estmean
cconfint(estmean,df=n-1)

# Domain = S
estmean <- svymean(~ACRES92,subset(domain_dsgn,REGION=="S"))
estmean
cconfint(estmean,df=n-1)

# Domain = W
estmean <- svymean(~ACRES92,subset(domain_dsgn,REGION=="W"))
estmean
cconfint(estmean,df=n-1)
> # Estimation of domain totals
> # Domain = NC
  total  SE   2.5 %  97.5 %
  ACRES92 384557574  41022160  ACRES92 303828848  465286299
> # Domain = NE
  total  SE   2.5 %  97.5 %
  ACRES92 17722098  4490614   ACRES92 8884885  26559311
> # Domain = S
  total  SE   2.5 %  97.5 %
  ACRES92 275091387 35287421  ACRES92 205648224 344534549
> # Domain = W
  total  SE   2.5 %  97.5 %
  ACRES92 239556051 46090457  ACRES92 148853274 330258829

> # Estimation of domain means
> # Domain = NC
  mean  SE   2.5 %  97.5 %
  ACRES92 350292  26985  297186.7  403397.3
> # Domain = NE
  mean  SE   2.5 %  97.5 %
  ACRES92 71971  12360  47646.95  96294.71
> # Domain = S
  mean  SE   2.5 %  97.5 %
  ACRES92 206246  23066  160854.6  251638.1
> # Domain = W
  mean  SE   2.5 %  97.5 %
  ACRES92 598681  77637  445897.3  751464

5.4.2 Using SAS to Perform a Domain Analysis (Supplemental)

- In the code, you must include a ‘domain’ statement that includes the domain variable. For this example, the domain variable is ‘region’.
- Unlike R, SAS is not case-sensitive. There is no difference between ‘REGION’, ‘region’, and ‘Region’ in SAS.
- The top section of the output ‘Statistics’ contains the SRS analysis of variable $y = \text{acres92}$ for estimating $\overline{y}_U$ and $t_y$.

SAS code for Domain Analysis

DATA agsrs;
  INFILE 'C:\COURSES\THAI\SASPSM\agsrs.dat';
  FORMAT county $char14.;
  INPUT i county $ st $ acres92 acres87 acres82 F92 F87 F82
           LF92 LF87 LF82 SF92 SF87 SF82 region $ @@;
DATA agsrs; SET agsrs;
*** enter population and sample sizes;
    N = 3078;
    nn= 300;

*** estimate weights when domain sizes Nd are unknown;
    utwgt = N/nn;    *** utwgt = N/n;

PROC SURVEYMEANS data=agsrs total=3078 nobs mean clm sum clsum df;
    var acres92;
    weight utwgt;
    domain region;
    title1 'Domain Estimation of ybar_Ud and t_d -- Acreage 1992 --- Nd unknown';
RUN;

SAS output for Domain Analysis
Domain Estimation of ybar_Ud and t_d -- Acreage 1992 --- Nd unknown

The SURVEYMEANS Procedure

Data Summary

Number of Observations 300
Sum of Weights  3078

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>DF</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>acres92</td>
<td>300</td>
<td>299</td>
<td>297897</td>
<td>18898</td>
<td>260706.257</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sum</th>
<th>Std Dev</th>
<th>95% CL for Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>acres92</td>
<td>916927110</td>
<td>58169381</td>
<td>802453859</td>
</tr>
</tbody>
</table>

Domain Analysis: region

<table>
<thead>
<tr>
<th>region</th>
<th>Variable</th>
<th>N</th>
<th>DF</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>acres92</td>
<td>107</td>
<td>299</td>
<td>350292</td>
<td>26985</td>
<td>297186.692</td>
</tr>
<tr>
<td>NE</td>
<td>acres92</td>
<td>24</td>
<td>299</td>
<td>71971</td>
<td>12360</td>
<td>47646.954</td>
</tr>
<tr>
<td>S</td>
<td>acres92</td>
<td>130</td>
<td>299</td>
<td>206246</td>
<td>23066</td>
<td>160854.596</td>
</tr>
<tr>
<td>W</td>
<td>acres92</td>
<td>39</td>
<td>299</td>
<td>598681</td>
<td>77637</td>
<td>445897.252</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>region</th>
<th>Variable</th>
<th>Sum</th>
<th>Std Dev</th>
<th>95% CL for Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>acres92</td>
<td>384557574</td>
<td>41022160</td>
<td>303828848</td>
</tr>
<tr>
<td>NE</td>
<td>acres92</td>
<td>17722098</td>
<td>4490614</td>
<td>8884885</td>
</tr>
<tr>
<td>S</td>
<td>acres92</td>
<td>275091387</td>
<td>35287421</td>
<td>205648224</td>
</tr>
<tr>
<td>W</td>
<td>acres92</td>
<td>239556051</td>
<td>46090457</td>
<td>148853274</td>
</tr>
</tbody>
</table>