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:

     $$B \;=$$

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

$$\overline{x}_U = \left(\sum_{i=1}^{N} x_i\right)/N \qquad t_x = \sum_{i=1}^{N} x_i \qquad \overline{y}_U = \left(\sum_{i=1}^{N} y_i\right)/N \qquad t_y = \sum_{i=1}^{N} y_i$$

$$\overline{x} = \sum_{i=1}^{n} x_i/n = \text{the sample mean of } x\text{'s.} \qquad \overline{y} = \sum_{i=1}^{n} y_i/n = \text{the sample mean of } y\text{'s.}$$

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

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

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

$$\widehat{B} \;=\; \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \qquad\qquad \widehat{V}(\widehat{B}) = \left(\frac{N-n}{N\overline{x}^2}\right)\frac{s_e^2}{n} \tag{47}$$

where

$$s_e^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{B}x_i)^2 \;=\; \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + \widehat{B}^2\sum_{i=1}^{n} x_i^2 - 2\widehat{B}\sum_{i=1}^{n} x_i y_i\right)$$

- If $Y \approx BX$, then $y_i \approx \widehat{B}x_i$. Thus, $\widehat{B}x_i$ can be considered the predicted value of y_i from a line through the origin (with intercept=0 and the slope = $\widehat{B}$).

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

- Multiplication of $\widehat{B}$ and $\widehat{V}(\widehat{B})$ by $\overline{x}_U$ and $\overline{x}_U^2$, respectively, in (47) yields the estimator $\widehat{\overline{y}_r}$ for $\overline{y}_U$ and its estimated variance:

$$\begin{aligned} \widehat{\overline{y}_r} &= \\ \widehat{V}(\widehat{\overline{y}_r}) &= \widehat{V}(\widehat{B}\overline{x}_U) = \overline{x}_U^2 \widehat{V}(\widehat{B}) = \left(\frac{N-n}{N}\right)\left(\frac{\overline{x}_U}{\overline{x}}\right)^2 \frac{s_e^2}{n} \end{aligned} \tag{48}$$

- $\widehat{\overline{y}_r}$ 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 $\widehat{t_{yr}}$ of t_y and the estimated variance:

$$\begin{aligned} \widehat{t_{yr}} &= \\ \widehat{V}(\widehat{t_{yr}}) &= N(N-n)\left(\frac{\overline{x}_U}{\overline{x}}\right)^2 \frac{s_e^2}{n} = \left(\frac{N-n}{N}\right)\left(\frac{t_x}{\overline{x}}\right)^2 \frac{s_e^2}{n} \end{aligned} \tag{49}$$

- $\widehat{t_{yr}}$ 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 $= \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \frac{y_i}{x_i}$. 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 $V(\widehat{\overline{y}_r})$ and $V(\widehat{t_{yr}})$.

Case II: t_x and $\overline{x}_U$ are unknown

- If t_x and $\overline{x}_U$ are unknown, it will not affect the estimator $\widehat{B} = \overline{y}/\overline{x}$. It will, however, affect the estimators $\widehat{\overline{y}_r}$ and $\widehat{t_{yr}}$ that depend on t_x and $\overline{x}_U$.

- In such cases, it is common to replace t_x with $N\overline{x}$ or replace $\overline{x}_U$ with $\overline{x}$. This will yield:

$$\widehat{V}(\widehat{\overline{y}_r}) \approx \left(\frac{N-n}{N}\right)\frac{s_e^2}{n} \qquad\qquad \widehat{V}(\widehat{t_{yr}}) = N(N-n)\frac{s_e^2}{n}$$

- When $\overline{x}$ is larger than $\overline{x}_U$, $\widehat{V}(\widehat{\overline{y}_r})$ and $\widehat{V}(\widehat{t_{yr}})$ tend to be too large as variance estimates. Similarly, when $\overline{x}$ is smaller than $\overline{x}_U$, $\widehat{V}(\widehat{\overline{y}_r})$ and $\widehat{V}(\widehat{t_{yr}})$ tend to be too small as variance estimates.

Example: Demonstration of Bias in Ratio Estimation

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

x_i value	67	63	66	69
y_i value	68	62	64	70

- 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, $S_x^2 = 6.250$ and $S_y^2 \approx 13.\overline{3}$.

Sample	Units	$\widehat{t_{yr}}$	$\widehat{V}(\widehat{t_{yr}})$	$\widehat{t}_{SRS}$	$\widehat{V}(\widehat{t}_{SRS})$
1	1,2	265	8	260	72
2	1,3	263.0075	18.0903	264	32
3	1,4	268.8971	0.0017	276	8
4	2,3	258.8372	1.7307	252	8
5	2,4	265	8	264	128
6	3,4	263.0370	18.2677	268	72
		$E(\widehat{t_{yr}}) = 263.963$	$E(\widehat{V}(\widehat{t_{yr}})) = 9.0151$		
		$V(\widehat{t_{yr}}) = 10.981$			

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

5.1.2 Bias and MSE of Ratio Estimators

- The ratio estimators are biased. The bias occurs in ratio estimation because $E(\overline{y}/\overline{x}) \neq E(\overline{y})/E(\overline{x})$ (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 $\overline{y}_r$ will be approximately normally distributed.
- In your text, it is shown that $\text{Bias}(\widehat{\overline{y}}_r) = [E(\widehat{\overline{y}}_r) - \overline{y}_U] = -Cov(\widehat{B}, \overline{x})$

  and, from this, it can be shown that $\frac{|\text{Bias}(\widehat{\overline{y}}_r)|}{\sqrt{V(\widehat{\overline{y}}_r)}} \leq \frac{\sqrt{V(\overline{x})}}{\overline{x}_U} = \text{CV}(\overline{x})$ where CV is the *coefficient of variation for* $\overline{x}$.
- It can also be shown that $\text{Bias}(\widehat{\overline{y}}_r) \approx \left(\frac{N-n}{N}\right) \frac{1}{n\overline{x}_U} (BS_x^2 - RS_xS_y)$

  where R is the population correlation coefficient.
- Thus, the bias of $\widehat{\overline{y}}_r$) (as well as $\widehat{t}_r$ and $\widehat{B}$) will be small if
  - the sample size n is large
  - the sampling fraction n/N is large
  - $\overline{x}_U$ 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

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

- 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:

$$\sum_{i=1}^{30} x_i = 3316 \qquad \sum_{i=1}^{30} y_i = 1802 \qquad \sum_{i=1}^{30} x_iy_i = 214,738$$

$$\sum_{i=1}^{30} x_i^2 = 392,440 \qquad \sum_{i=1}^{30} y_i^2 = 118,360 \qquad t_x = \sum_{i=1}^{800} x_i = 89420$$

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, $\overline{y}_U$, and t_y are:

$$\widehat{B} \pm z^* \sqrt{\widehat{V}(\widehat{B})} \qquad \widehat{\overline{y}_r} \pm z^* \sqrt{\widehat{V}(\widehat{\overline{y}_r})} \qquad \widehat{t_{yr}} \pm z^* \sqrt{\widehat{V}(\widehat{t_{yr}})} \tag{50}$$

  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:

$$\widehat{B} \pm t^* \sqrt{\widehat{V}(\widehat{B})} \qquad \widehat{\overline{y}_r} \pm t^* \sqrt{\widehat{V}(\widehat{\overline{y}_r})} \qquad \widehat{t_{yr}} \pm t^* \sqrt{\widehat{V}(\widehat{t_{yr}})} \tag{51}$$

  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 $C_X = \dfrac{S_x}{\overline{x}_U}$ and $C_Y = \dfrac{S_y}{\overline{y}_U}$ are $< .10/\sqrt{n}$.

- **Example**: Find 95% confidence intervals for t_y, $\overline{y}_U$, 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 (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 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 (≥ 1000 acres) in 1992, 1987, and 1982, respectively, for that county.

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

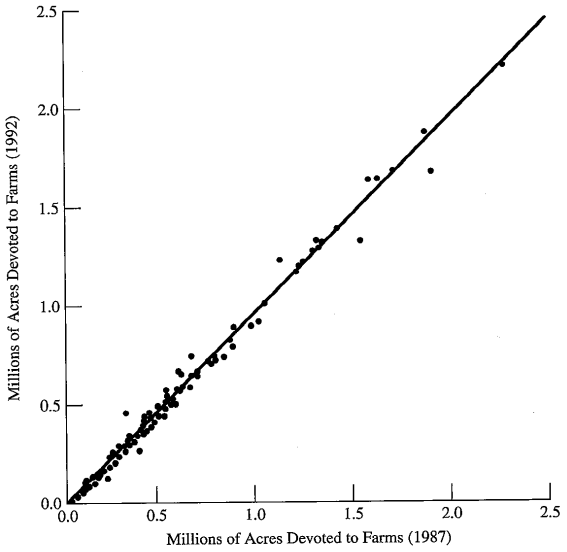
Region represents one of four an assigned geographical regions of the United States (W=West, S=South, NE=Northeast NC=North central).

- The following table is a summary of the number of counties (N_i) in each state (i).

State		N_i	Region
Alaska	AK	5	W
Alabama	AL	67	S
Arkansas	AR	75	S
Arizona	AZ	15	W
California	CA	58	W
Colorado	CO	63	W
Connecticut	CT	8	NE
Delaware	DE	3	NE
Florida	FL	67	S
Georgia	GA	159	S
Hawaii	HI	4	W
Iowa	IA	99	NC
Idaho	ID	44	W
Illinois	IL	102	NC
Indiana	IN	92	NC
Kansas	KS	105	NC
Kentucky	KY	120	S

State		N_i	Region
Louisiana	LA	64	S
Massachusetts	MA	14	NE
Maryland	MD	23	S
Maine	ME	16	NE
Michigan	MI	83	NC
Minnisota	MN	87	NC
Missouri	MO	114	NC
Mississippi	MS	82	S
Montana	MT	56	W
North Carolina	NC	100	S
North Dakota	ND	53	NC
Nebraska	NE	93	NC
New Hampshire	NH	10	NE
New Jersey	NJ	21	NE
New Mexico	NM	33	W
Nevada	NV	17	W
New York	NY	62	NE

State		N_i	Region
Ohio	OH	88	NC
Oklahoma	OK	77	S
Oregon	OR	36	W
Pennsylvania	PA	67	NE
Rhode Island	RI	5	NE
South Carolina	SC	46	S
South Dakota	SD	66	NC
Tennessee	TN	95	S
Texas	TX	254	S
Utah	UT	29	W
Virginia	VA	98	S
Vermont	VT	14	NE
Washington	WA	39	W
Wisconsin	WI	72	NC
West Virginia	WV	55	S
Wyoming	WY	23	W
Total	**50**	**3078**	



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

5.2.1 Ratio Estimation Using R

- **CASE 1**: Estimating $\widehat{B}$.
- **CASE 2**: Estimating $\overline{y}_U$ when t_x and $\overline{x}_U$ are known.
- **CASE 3**: Estimating t_y when t_x and $\overline{x}_U$ are known.
- **CASE 4**: Estimating $\overline{y}_U$ when t_x and $\overline{x}_U$ are unknown.
- **CASE 5**: Estimating t_y when t_x and $\overline{x}_U$ are unknown.

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

mnhatx  = mean(ratio$ACRES87)  # estimated X pop. mean  if unknown: Case 4
thatx = N*mnhatx               # estimated X pop. total if unknown: Case 5
ratio_umn  = mnhatx*ratio$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                                        CASE 1

-------------------------------------------------------------------
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

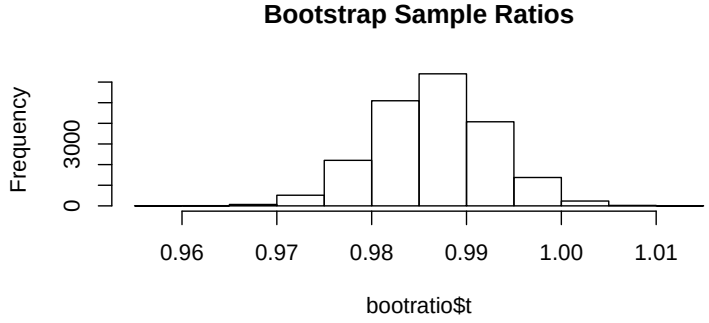
- Bootstrapping can only be used to estimate t_y and $\overline{y}_U$ when t_x and $\overline{x}_U$ are known.
- Why? Replacing t_x and $\overline{x}_U$ with $N\overline{x}$ and $\overline{x}$ in each bootstrap sample is equivalent to the SRS bootstrap of $\overline{y}$ and $N\overline{y}$. Thus, bootstrap estimates of t_y and $\overline{y}_U$ ignore all information about x.
- The follow 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.

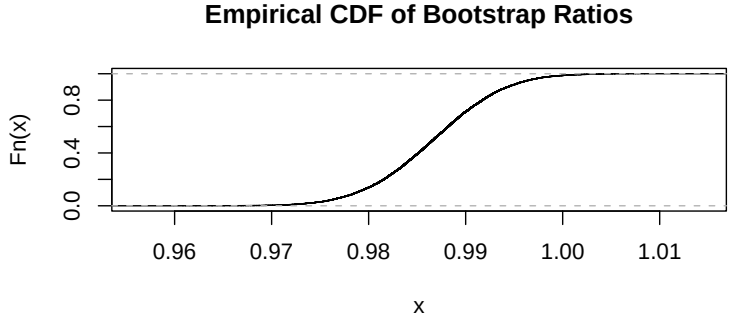
R code for bootstrapping ratio estimates

```
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)
bootratio
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)
bootmean
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)
boottotal
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 )
```





5.2.3 Ratio Estimation Using *SAS* Proc Surveymeans (Supplemental)

- **CASE 1**: Because $\widehat{B}$ and the s.e.($\widehat{B}$) 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 $\widehat{\overline{y}_r}$ and $\widehat{t_{yr}}$ 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 $(\overline{x}_U\overline{y})/\overline{x} = \widehat{B}\overline{x}_U$ 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 $(t_x\, y)$-values. Then consider the ratio $(t_x\overline{y})/\overline{x} = \widehat{B}t_x$ 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 $(\overline{xy})/\overline{x} = \widehat{B}\overline{x}$ 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 $(N\overline{x}\,\overline{y})/\overline{x} = \widehat{B}N\overline{x}$ 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 $\widehat{B}$, its standard error, and a 95% confidence interval for B.
- For Cases 2 and 3, the 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 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

```
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;
```

```
*** 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

                                    Std Error
     Variable            Mean         of Mean       95% CL for Mean
     ---------------------------------------------------------------
     y                 297897           18898    260706.257 335087.836
     x                 301954           18914    264732.959 339174.488
     ---------------------------------------------------------------
```

Ratio Analysis

Numerator	Denominator	Ratio	Std Err	95% CL for Ratio	
y	x	0.986565	0.005750	0.97524871	0.99788176 <--

(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

Variable	Mean	Std Error of Mean	95% CL for Mean	
x	301954	18914	264733	339174
y	297897	18898	260706	335088
y_mean	93344038591	5921697487	8.16906E10	1.04998E11

Ratio Analysis

Numerator	Denominator	Ratio	Std Err	95% CL for Ratio	
y_mean	x	309134	1801.871993	305587.633	312679.548 <--

(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

Variable	Mean	Std Error of Mean	95% CL for Mean	
x	301954	18914	264733	339174
y	297897	18898	260706	335088
y_total	2.8731295E14	1.8226985E13	2.51444E14	3.23182E14

Ratio Analysis

Numerator	Denominator	Ratio	Std Err	95% CL for Ratio	
y_total	x	951513191	5546162	940598734	962427648 <--

```
(OUTPUT FOR CASE 4)

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

                        The SURVEYMEANS Procedure
                              Data Summary

                     Number of Observations           300

                              Statistics

                                   Std Error
     Variable            Mean        of Mean       95% CL for Mean
     ---------------------------------------------------------------
     x                 301954          18914        264733     339174
     y                 297897          18898        260706     335088
     y_umean      89951122411     5706452641    7.87212E10 1.01181E11
     ---------------------------------------------------------------

                             Ratio Analysis

Numerator Denominator     Ratio        Std Err       95% CL for Ratio
-------------------------------------------------------------------------
y_umean   x               297897    1736.376641   294479.980   301314.114 <--
-------------------------------------------------------------------------


(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

                                   Std Error
     Variable            Mean        of Mean       95% CL for Mean
     ---------------------------------------------------------------
     x                 301954          18914        264733     339174
     y                 297897          18898        260706     335088
     y_utotal    2.7686955E14    1.7564461E13    2.42304E14 3.11435E14
     ---------------------------------------------------------------

                             Ratio Analysis

Numerator Denominator      Ratio      Std Err       95% CL for Ratio
-----------------------------------------------------------------------
y_utotal  x            916927110      5344567    906409377    927444842 <--
-----------------------------------------------------------------------
```

5.3 $\widehat{\overline{y}_U}$ vs $\widehat{\overline{y}_r}$ or $\widehat{t}$ vs $\widehat{t_{yr}}$ 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**

$$R = \frac{S_{xy}}{S_y S_x} \quad \text{where} \quad S_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x}_U)(y_i - \overline{y}_U)}{N-1} .$$

- It can be shown that approximations for the true population variances and MSEs of $\widehat{t_{yr}}$ and $\widehat{\overline{y}}_r$ are

$$MSE(\widehat{t_{yr}}) \approx V(\widehat{t_{yr}}) \approx \frac{N(N-n)}{n}(S_y^2 - 2BRS_yS_x + B^2S_x^2)$$

$$MSE(\widehat{\overline{y}}_r) \approx V(\widehat{\overline{y}}_r) \approx \frac{N-n}{Nn}(S_y^2 - 2BRS_yS_x + B^2S_x^2)$$

- 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 $\overline{y}_U$, 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 $\widehat{\overline{y}_U}$ or $\widehat{t}$ which does not require knowledge about X?'

- The answer requires looking at the **coefficient of variation** for both X and Y.

$$C_X = \qquad\qquad C_Y =$$

- 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 ($\widehat{C}_x$ and $\widehat{C}_y$) to check if these conditions are met. The formulas are:

$$\widehat{C}_x = \qquad\qquad \widehat{C}_y =$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) = \frac{\sum_{i=1}^n x_iy_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{(n-1)s_xs_y}$$

  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 $\overline{x}_U$ 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

$$t_{yd} = \sum_{i \in U_d} y_i \qquad \overline{y}_{Ud} = t_{yd}/N_d = \left(\sum_{i \in U_d} y_i \right) /N_d$$

- 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 $\overline{y}_{Ud}$ and the domain total t_{yd} are

$$\widehat{\overline{y}_{Ud}} = \left(\sum_{i \in S_d} y_i \right) /n_d = \overline{y}_d \qquad \widehat{t_{yd}} = \frac{N_d}{n_d} \left(\sum_{i \in S_d} y_i \right) = N_d \overline{y}_d$$

- $\overline{y}_d$ looks like the estimator $\widehat{\overline{y}_U} = \overline{y}$ for a SRS of size n_d, and $N_d \overline{y}_d$ looks like the estimator $\widehat{t} = N\overline{y}$ 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 $V(\overline{y}_d)$, we need to see that $\overline{y}_d$ is a ratio estimator.
- Let $u_i = \left\{ \begin{array}{ll} y_i & \text{if } i \in U_d \\ 0 & \text{if } i \notin U_d \end{array} \right\} \quad x_i = \left\{ \begin{array}{ll} 1 & \text{if } i \in U_d \\ 0 & \text{if } i \notin U_d \end{array} \right\} \quad \text{for} \quad i = 1, 2, \ldots, N.$
- Notation: Let $\overline{y}_{Ud}$, $\overline{x}_{Ud}$, and $\overline{x}_{Ud}$ be the domain means of the y, x, and u values, respectively. Then,

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

$$\begin{aligned} \text{Let } B_d &= \frac{\overline{u}_{Ud}}{\overline{x}_{Ud}} = \frac{(\sum_{i=1}^{N} u_i)/N}{(\sum_{i=1}^{N} x_i)/N} = \frac{\sum_{i=1}^{N} u_i}{\sum_{i=1}^{N} x_i} \\ &= \frac{\sum_{i \in U_d} u_i + \sum_{i \notin U_d} u_i}{\sum_{i \in U_d} x_i + \sum_{i \notin U_d} x_i} \\ &= \frac{\sum_{i \in U_d} u_i + \sum_{i \notin U_d} 0}{\sum_{i \in U_d} 1 + \sum_{i \notin U_d} 0} \\ &= \frac{\sum_{i \in U_d} u_i}{\sum_{i \in U_d} 1} = \frac{\sum_{i \in U_d} y_i}{N_d} = \overline{y}_{Ud} \end{aligned}$$

- 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 $B_d = \overline{y}_{Ud}$ with the ratio of domain sample means:

$$\begin{aligned}
\widehat{B}_d &= \frac{\overline{u}_d}{\overline{x}_d} = \frac{(\sum_{i\in S} u_i)/n}{(\sum_{i\in S} x_i)/n} = \frac{\sum_{i\in S} u_i}{\sum_{i\in S} x_i} \\
&= \frac{\sum_{i\in S_d} u_i + \sum_{i\notin S_d} u_i}{\sum_{i\in S_d} x_i + \sum_{i\notin S_d} x_i} \\
&= \frac{\sum_{i\in S_d} y_i + \sum_{i\notin S_d} 0}{\sum_{i\in S_d} 1 + \sum_{i\notin S_d} 0} \\
&= \frac{\sum_{i\in S_d} y_i}{\sum_{i\in S_d} 1} = \frac{\sum_{i\in S_d} y_i}{n_d} = \overline{y}_d
\end{aligned}$$

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

$$\widehat{V}(\overline{y}_d) = \widehat{V}(B_d) = \left(\frac{N-n}{N\overline{x}^2_{Ud)}}\right)\frac{s_u^2}{n} = \left(\frac{N-n}{N(N_d/N)^2}\right)\frac{s_u^2}{n} = \left(\frac{N-n}{N}\right)\left(\frac{N}{N_d}\right)^2\frac{s_u^2}{n} \qquad (52)$$

$$\begin{aligned}
\text{where } s_u^2 &= \frac{1}{n-1}\sum_{i\in S}(u_i - \widehat{B}_d x_i)^2 = \frac{1}{n-1}\left(\sum_{i\in S} u_i^2 + \widehat{B}_d^2\sum_{i\in S} x_i^2 - 2r_d\sum_{i\in S} x_i u_i\right) \\
&= \frac{1}{n-1}\left(\sum_{i\in S_d} u_i^2 + \widehat{B}_d^2\sum_{i\in S_d} x_i^2 - 2\widehat{B}_d\sum_{i\in S_d} x_i u_i\right) \\
&= \frac{1}{n-1}\left(\sum_{i\in S_d} y_i^2 + \widehat{B}_d^2\sum_{i\in S_d}(1) - 2\widehat{B}_d\sum_{i\in S_d}(1) y_i\right) \\
&= \frac{1}{n-1}\left(\sum_{i\in S_d} y_i^2 + n_d\widehat{B}_d^2 - 2\widehat{B}_d\sum_{i\in S_d} y_i\right)
\end{aligned}$$

- When N_d is unknown, replace N_d with estimate $\widehat{N}_d = Nn_d/n$ (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 ($\overline{x}_d = n_d/300$) from domain d and the proportion of population units ($\overline{x}_{Ud} = N_d/3078$) in domain d:

d	n_d	$\sum_{i\in S_d} y_i$	$\overline{y}_d$	$\overline{x}_d$	N_d	$\overline{x}_{Ud}$
NC	107	37,481,245	350292	$.35\overline{6}$	1054	.3424
NE	24	1,727,300	71971	.08	220	.0715
S	130	26,812,026	206246	.43	1382	.4490
W	39	23,348,543	598681	.13	422	.1371
Total	$n = 300$				$N = 3078$	

Note that the proportion of the sample from domain d is close to the actual proportion of population units in domain d ($\overline{x}_d \approx \overline{x}_{Ud}$).

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
confint(esttotal,df=n-1)

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

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

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

# Estimation of domain means

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

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

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

# Domain = W
estmean <- svymean(~ACRES92,subset(domain_dsgn,REGION=="W"))
estmean
confint(estmean,df=n-1)
```

R output for Domain Analysis

```
> # 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       ACRES92 297186.7 403397.3

> # Domain = NE
         mean    SE                  2.5 %   97.5 %
ACRES92 71971 12360        ACRES92 47646.95 96294.71

> # Domain = S
          mean    SE                  2.5 %   97.5 %
ACRES92 206246 23066       ACRES92 160854.6 251638.1

> # Domain = W
          mean    SE                  2.5 % 97.5 %
ACRES92 598681 77637       ACRES92 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 $y =$**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
                                    Std Error
Variable     N     DF      Mean     of Mean       95% CL for Mean
-----------------------------------------------------------------------
acres92    300    299    297897       18898    260706.257 335087.836
-----------------------------------------------------------------------

Variable             Sum         Std Dev        95% CL for Sum
-------------------------------------------------------------------
acres92        916927110        58169381     802453859 1031400361
-------------------------------------------------------------------

                Domain Analysis: region

                                              Std Error
region   Variable     N   DF      Mean     of Mean     95% CL for Mean
-----------------------------------------------------------------------------
NC       acres92    107  299    350292       26985  297186.692 403397.326
NE       acres92     24  299     71971       12360   47646.954  96294.713
S        acres92    130  299    206246       23066  160854.596 251638.111
W        acres92     39  299    598681       77637  445897.252 751463.927
-----------------------------------------------------------------------------

region   Variable           Sum         Std Dev        95% CL for Sum
-----------------------------------------------------------------------------
NC       acres92      384557574        41022160     303828848  465286299
NE       acres92       17722098         4490614       8884885   26559311
S        acres92      275091387        35287421     205648224  344534549
W        acres92      239556051        46090457     148853274  330258829
-----------------------------------------------------------------------------
```