Ratio and Regression Estimation

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

1/29/25

Review: Stratified sampling

1. Identify stratifying variables correlated with the measure(s) of interest
2. Choose "cuts" on the stratifying variables and divide the population into strata
3. S_h, $deff$, and S^2 are known, at least approximately, and known that fpc's within strata can be ignored
4. Compute an stratified sample size $n = n_{SRS} * deff$
5. Determine an allocation for the desired n
6. Adjust n based on expected $deff$ & allocate
7. Select sample and compute estimates taking the stratified sample selection into account

Model-based theory

- In randomization theory, or design-based sampling, the **sampling design determines how sampling variability is estimated**.
- In model-based sampling, the **model determines how variability is estimated**, and the sampling design is irrelevant—as long as the model holds, you could choose any n units you want to from the population.
- The discrepancy is due to the different definitions of variance:
    - In design-based sampling, the variance is the average squared deviation of the estimate from its expected value, averaged over all samples that could be **obtained using a given design**.
    - If we are using a model, the variance is again the average squared deviation of the estimate from its expected value, but here the average is over all possible samples that could be **generated from the population model**.

Model-based theory cont.

- Design-based inferences about finite population quantities using ratio or regression estimation are correct *even if the model does not fit the data well*.
- Model-assisted estimators: A model motivates the form of the estimator, but inference depends on the sampling design.
- If we adopt a model consistent with the reasons we would adopt a certain sampling scheme or method of estimation, the point estimators would be similar. The model-based variance, though, usually differs from the variance from the randomization theory.

Ratio and regression estimation

- Ratio estimation is used to improve the precision of estimates by incorporating auxiliary information correlated with the variable of interest.

- Suppose we want to estimate the total or mean of a population characteristic Y. We have an auxiliary variable X for which the total is known, or its sample mean can be accurately estimated.

- The ratio estimator uses both Y and X to improve estimation.

- Define the ratio $B = \frac{t_y}{t_x} = \frac{\bar{Y}}{\bar{X}}$.

- Ratio and regression estimation both use the correlation of X and Y. The population correlation coefficient of X and Y is

$$R = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)S_x S_y}$$

Why using ratio estimation

- Ratio estimator $\widehat{B} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{\bar{y}}{\bar{x}}$, and $\widehat{\bar{y}}_r = \widehat{B}\bar{X}$
- When the sample size changes as a random variable, we need to use ratio estimation for the mean
- When we want to estimate a population total, but the population size is unknown. We can estimate $\widehat{N}$ by $t_x/\bar{X}$
- Increase the precision of estimated means and totals
- Adjust estimates from the sample to reflect population totals (poststratification)
- Adjust for nonresponse (discussed later)

Examples

- Using the total number of registered births as auxiliary information for estimating the total population in France (Laplace, 1814)
- Using field acreage to estimate field yield of grains
- Using farm acreage in 1987 to estimate farm acreage in 1992
- Using family income to estimate wealth
- Using the number of employees to estimate the amount spent on health insurance in a business

Poststratification: Example

- An SRS of 400 students taken from a school with 4000 students: 240 women and 160 men
- With 84 of the sampled women and 40 of the sampled men planning to follow careers in academia
- Question: How many students planning to work in academia?

1. SRS: $\frac{4000}{400} * 124 = 1240$
2. If we know that the school has 2700 women and 1300 men, another estimate is $\frac{2700}{240} * 84 + \frac{1300}{160} * 40 = 1270$

- We can treat this as a ratio estimation by gender: $\frac{84}{240} * 2700 + \frac{40}{160} * 1300 = 1270$
  - The sample has 60% women, but the population has 67.5%.
  - We adjust the estimated total by the sexual decomposition discrepancy: **Poststratification**

Ratio estimators are usually biased

- If we calculate $\hat{\bar{y}}_r = \bar{y}\frac{\bar{X}}{\bar{x}}$ for all possible SRSs, their average value will be close to but usually not be equal $\bar{Y}$ exactly.

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) = -Cov(\hat{B}, \bar{x})$$

Ratio estimators are usually biased

- If we calculate $\hat{\bar{y}}_r = \bar{y}\frac{\bar{X}}{\bar{x}}$ for all possible SRSs, their average value will be close to but usually not be equal $\bar{Y}$ exactly.

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) = -Cov(\hat{B}, \bar{x})$$

- We have

$$\frac{|Bias(\hat{\bar{y}}_r)|}{\sqrt{Var(\hat{\bar{y}}_r)}} = \frac{|Cov(\hat{B}, \bar{x})|}{\bar{X}\sqrt{Var(\hat{B})}} \leq \frac{\sqrt{Var(\bar{x})}}{\bar{X}} = CV(\bar{x})$$

Ratio estimators are usually biased

- If we calculate $\hat{\bar{y}}_r = \bar{y}\frac{\bar{X}}{\bar{x}}$ for all possible SRSs, their average value will be close to but usually not be equal $\bar{Y}$ exactly.

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) = -Cov(\hat{B}, \bar{x})$$

- We have

$$\frac{|Bias(\hat{\bar{y}}_r)|}{\sqrt{Var(\hat{\bar{y}}_r)}} = \frac{|Cov(\hat{B}, \bar{x})|}{\bar{X}\sqrt{Var(\hat{B})}} \leq \frac{\sqrt{Var(\bar{x})}}{\bar{X}} = CV(\bar{x})$$

- The absolute value of the bias is small relative to the standard deviation if $CV(\bar{x})$ is small.

Ratio estimators are usually biased

- If we calculate $\hat{\bar{y}}_r = \bar{y}\frac{\bar{X}}{\bar{x}}$ for all possible SRSs, their average value will be close to but usually not be equal $\bar{Y}$ exactly.

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) = -Cov(\hat{B}, \bar{x})$$

- We have

$$\frac{|Bias(\hat{\bar{y}}_r)|}{\sqrt{Var(\hat{\bar{y}}_r)}} = \frac{|Cov(\hat{B}, \bar{x})|}{\bar{X}\sqrt{Var(\hat{B})}} \leq \frac{\sqrt{Var(\bar{x})}}{\bar{X}} = CV(\bar{x})$$

- The absolute value of the bias is small relative to the standard deviation if $CV(\bar{x})$ is small.
- A small $CV(\bar{x})$ means that $\bar{x}$ is stable from sample to sample

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

- We can show

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) \approx (1 - \frac{n}{N})\frac{1}{n\bar{X}(BS_x^2 - RS_xS_y)} \quad (1)$$

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

- We can show

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) \approx (1 - \frac{n}{N})\frac{1}{n\bar{X}(BS_x^2 - RS_xS_y)} \quad (1)$$

- The bias is small if

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

- We can show

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) \approx (1 - \frac{n}{N})\frac{1}{n\bar{X}(BS_x^2 - RS_xS_y)} \qquad (1)$$

- The bias is small if
  - n is large

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

- We can show

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) \approx (1 - \frac{n}{N})\frac{1}{n\bar{X}(BS_x^2 - RS_xS_y)} \qquad (1)$$

- The bias is small if
  - n is large
  - n/N is large

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

- We can show

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) \approx (1 - \frac{n}{N})\frac{1}{n\bar{X}(BS_x^2 - RS_xS_y)} \qquad (1)$$

- The bias is small if
  - n is large
  - n/N is large
  - $\bar{X}$ is large

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

- We can show

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) \approx (1 - \frac{n}{N})\frac{1}{n\bar{X}(BS_x^2 - RS_xS_y)} \qquad (1)$$

- The bias is small if
  - n is large
  - n/N is large
  - $\bar{X}$ is large
  - S_x is small

Approximation for the bias

- Linearization approach to approximating variances

$$\hat{\bar{y}}_r - \bar{Y} = (\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})$$

- We can show

$$Bias(\hat{\bar{y}}_r) = E(\hat{\bar{y}}_r - \bar{Y}) \approx (1 - \frac{n}{N})\frac{1}{n\bar{X}(BS_x^2 - RS_xS_y)} \quad (1)$$

- The bias is small if
  - n is large
  - n/N is large
  - $\bar{X}$ is large
  - S_x is small
  - R is close to 1

Ratio estimation with weights

- With the sampling weight $w_i = 1/\pi_i$, $\bar{y}_w = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$

Ratio estimation with weights

- With the sampling weight $w_i = 1/\pi_i$, $\bar{y}_w = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$
- Note that $\hat{t}_{yr} = \frac{t_x}{\hat{t}_x} \hat{t}_{yw} = \frac{t_x}{\hat{t}_x} \sum_{i \in s} w_i y_i$, we can think of the modification used in ratio estimation as an adjustment to each weight.

Ratio estimation with weights

- With the sampling weight $w_i = 1/\pi_i$, $\bar{y}_w = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$
- Note that $\hat{t}_{yr} = \frac{t_x}{\hat{t}_x} \hat{t}_{yw} = \frac{t_x}{\hat{t}_x} \sum_{i \in s} w_i y_i$, we can think of the modification used in ratio estimation as an adjustment to each weight.
- Define $g_i = \frac{t_x}{\hat{t}_x}$, then $\hat{t}_{yr} = \sum_{i \in s} w_i g_i y_i$ as a weighted sum of the observations, with new weights

Ratio estimation with weights

- With the sampling weight $w_i = 1/\pi_i$, $\bar{y}_w = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$
- Note that $\hat{t}_{yr} = \frac{t_x}{\hat{t}_x} \hat{t}_{yw} = \frac{t_x}{\hat{t}_x} \sum_{i \in s} w_i y_i$, we can think of the modification used in ratio estimation as an adjustment to each weight.
- Define $g_i = \frac{t_x}{\hat{t}_x}$, then $\hat{t}_{yr} = \sum_{i \in s} w_i g_i y_i$ as a weighted sum of the observations, with new weights
- The weight adjustments g_i calibrates the estimates based on x: $\sum_{i \in s} w_i g_i x_i = t_x$.

Mean squared error (MSE)

- We have

$$E[(\hat{\bar{y}}_r - \bar{Y})^2] = E[\{(\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})\}^2] \approx E[(\bar{y} - B\bar{x})^2]$$

Mean squared error (MSE)

- We have

$$E[(\hat{\bar{y}}_r - \bar{Y})^2] = E[\{(\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})\}^2] \approx E[(\bar{y} - B\bar{x})^2]$$

- The **variance and MSE are both approximated** by

$$MSE(\hat{\bar{y}}_r) = E[(\bar{y} - B\bar{x})^2] = (1/n - 1/N)(S_y^2 - 2BRS_x + B^2S_x^2)$$

Mean squared error (MSE)

- We have

$$E[(\hat{\bar{y}}_r - \bar{Y})^2] = E[\{(\bar{y} - B\bar{x})(1 - \frac{\bar{x} - \bar{X}}{\bar{x}})\}^2] \approx E[(\bar{y} - B\bar{x})^2]$$

- The **variance and MSE are both approximated** by

$$MSE(\hat{\bar{y}}_r) = E[(\bar{y} - B\bar{x})^2] = (1/n - 1/N)(S_y^2 - 2BRS_x + B^2 S_x^2)$$

- The MSE is small if n is large, n/N is large, the deviations $d_i = y_i - Bx_i$ is small, and R is close to 1.

Variance estimator based on regression residuals

- Since $E[(\bar{y} - B\bar{x})^2] = Var(\bar{d})$, $\bar{d}_U = 0$, define a new variable $e_i = y_i - \hat{B}x_i$ as the i-th residual from fitting the line $y = \hat{B}x$.

Variance estimator based on regression residuals

- Since $E[(\bar{y} - B\bar{x})^2] = Var(\bar{d})$, $\bar{d}_U = 0$, define a new variable $e_i = y_i - \hat{B}x_i$ as the i-th residual from fitting the line $y = \hat{B}x$.
- Estimate $var(\hat{\bar{y}}_r) = (1 - \frac{n}{N})(\frac{\bar{X}}{\bar{x}})^2 \frac{s_e^2}{n}$ with $s_e^2 = \frac{1}{n-1} \sum_{i \in S} e_i^2$

Variance estimator based on regression residuals

- Since $E[(\bar{y} - B\bar{x})^2] = Var(\bar{d})$, $\bar{d}_U = 0$, define a new variable $e_i = y_i - \hat{B}x_i$ as the i-th residual from fitting the line $y = \hat{B}x$.
- Estimate $var(\hat{\bar{y}}_r) = (1 - \frac{n}{N})(\frac{\bar{X}}{\bar{x}})^2 \frac{s_e^2}{n}$ with $s_e^2 = \frac{1}{n-1} \sum_{i \in S} e_i^2$
- With weights $g_i = \frac{\bar{X}}{\bar{x}}$ and $u_i = g_i e_i$, for an SRS, $var(\bar{u}) = (1 - \frac{n}{N}) \frac{1}{n(n-1)} \sum_{i \in S} (u_i - \bar{u})^2 = (1 - \frac{n}{N})(\frac{\bar{X}}{\bar{x}})^2 \frac{s_e^2}{n} = var(\hat{\bar{y}}_r)$

Variance estimator based on regression residuals

- Since $E[(\bar{y} - B\bar{x})^2] = Var(\bar{d})$, $\bar{d}_U = 0$, define a new variable $e_i = y_i - \hat{B}x_i$ as the i-th residual from fitting the line $y = \hat{B}x$.
- Estimate $var(\hat{\bar{y}}_r) = (1 - \frac{n}{N})(\frac{\bar{X}}{\bar{x}})^2 \frac{s_e^2}{n}$ with $s_e^2 = \frac{1}{n-1} \sum_{i \in S} e_i^2$
- With weights $g_i = \frac{\bar{X}}{\bar{x}}$ and $u_i = g_i e_i$, for an SRS, $var(\bar{u}) = (1 - \frac{n}{N}) \frac{1}{n(n-1)} \sum_{i \in S} (u_i - \bar{u})^2 = (1 - \frac{n}{N})(\frac{\bar{X}}{\bar{x}})^2 \frac{s_e^2}{n} = var(\hat{\bar{y}}_r)$
- Similarly

$$var(\hat{B}) = (1 - \frac{n}{N}) \frac{s_e^2}{n\bar{x}^2} \tag{2}$$

R code: Examples 4.2 in Lohr

```
agsrs$sampwt <- rep(3078/n,n)
agdsrs <- svydesign(id = ~1, weights=~sampwt, fpc=rep(3078,300), data = agsrs)
# estimate the ratio acres92/acres87
sratio<-svyratio(numerator = ~acres92, denominator = ~acres87,design = agdsrs); sratio
```

```
Ratio estimator: svyratio.survey.design2(numerator = ~acres92, denominator = ~acres87,
    design = agdsrs)
Ratios=
          acres87
acres92 0.9865652
SEs=
            acres87
acres92 0.005750473
```

```
confint(sratio, df=degf(agdsrs))
```

```
                    2.5 %    97.5 %
acres92/acres87 0.9752487 0.9978818
```

```
# provide the population total of x
xpoptotal <- 964470625
# Ratio estimate of population total
predict(sratio,total=xpoptotal)
```

```
$total
          acres87
acres92 951513191

$se
        acres87
acres92 5546162
```

```
# Ratio estimate of population mean
predict(sratio,total=xpoptotal/3078)
```

Regression estimation in SRS

- Ratio estimation works best in the data are well fit by a straight line through the origin.

Regression estimation in SRS

- Ratio estimation works best in the data are well fit by a straight line through the origin.
- General regression model $y = B_0 + B_1 x$ with ordinary least squares regression coefficient estimate

$$\hat{B}_1 = \frac{\sum_{i \in s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in s}(x_i - \bar{x})^2} = \frac{r s_y}{s_x}$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$$

Regression estimation in SRS

- Ratio estimation works best in the data are well fit by a straight line through the origin.
- General regression model $y = B_0 + B_1 x$ with ordinary least squares regression coefficient estimate

$$\hat{B}_1 = \frac{\sum_{i \in s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in s}(x_i - \bar{x})^2} = \frac{r s_y}{s_x}$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$$

- If we know $\bar{X}$, we can obtain the regression estimator $\hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{X} = \bar{y} + \hat{B}_1(\bar{X} - \bar{x})$

Regression estimator is biased

- Let population slope $B_1 = \frac{RS_y}{S_x}$, we have
  $E(\hat{\bar{y}}_{reg} - \bar{Y}) = -Cov(\hat{B}_1, \bar{x})$

Regression estimator is biased

- Let population slope $B_1 = \frac{RS_y}{S_x}$, we have $E(\hat{\bar{y}}_{reg} - \bar{Y}) = -Cov(\hat{B}_1, \bar{x})$
- For large SRSs the MSE for regression estimation is similar to the variance. Let $d_i = y_i - [\bar{Y} + B_1(x_i - \bar{X})]$, then

$$MSE(\hat{\bar{y}}_{reg}) = E[\{\bar{y}\bar{Y} + B_1(x_i - \bar{X})\}^2] \approx V(\bar{d}) = (1 - \frac{n}{N})\frac{S_d^2}{n}$$

Regression estimator is biased

- Let population slope $B_1 = \frac{RS_y}{S_x}$, we have $E(\hat{\bar{y}}_{reg} - \bar{Y}) = -Cov(\hat{B}_1, \bar{x})$
- For large SRSs the MSE for regression estimation is similar to the variance. Let $d_i = y_i - [\bar{Y} + B_1(x_i - \bar{X})]$, then

$$MSE(\hat{\bar{y}}_{reg}) = E[\{\bar{y}\bar{Y} + B_1(x_i - \bar{X})\}^2] \approx V(\bar{d}) = (1 - \frac{n}{N})\frac{S_d^2}{n}$$

- We can estimate S_d^2 by using residuals $e_i = y_i - (\hat{B}_0 + \hat{B}_1 x_i)$ and $s_e^2 = \sum_i e_i^2/(n-2)$

$$SE(\hat{\bar{y}}_{reg}) = \sqrt{(1 - \frac{n}{N})\frac{1}{n}s_y^2(1 - r^2)} \tag{3}$$

R code: Lohr Example 4.7

```
data(deadtrees); #head(deadtrees) nrow(deadtrees) # 25
# Fit with survey regression
dtree<- svydesign(id = ~1, weight=rep(4,25), fpc=rep(100,25), data = deadtrees)
myfit1 <- svyglm(field~photo, design=dtree)
summary(myfit1) # displays regression coefficients
```

```
Call:
svyglm(formula = field ~ photo, design = dtree)

Survey design:
svydesign(id = ~1, weight = rep(4, 25), fpc = rep(100, 25), data = deadtrees)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0593     1.3930   3.632   0.0014 **
photo         0.6133     0.1259   4.870 6.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5.548341)

Number of Fisher Scoring iterations: 2
```

```
confint(myfit1,df=23) # df = 25-2
```

```
                2.5 %   97.5 %
(Intercept) 2.1777362 7.940848
photo       0.3527717 0.873777
```

R code: Lohr Example 4.7 cont.

```
# Regression estimate of population mean field trees
newdata <- data.frame(photo=11.3)
predict(myfit1, newdata)
```

```
    link    SE
1 11.989 0.418
```

```
confint(predict(myfit1, newdata),df=23)
```

```
     2.5 %   97.5 %
1 11.12455 12.85404
```

```
# Estimate total field tree, add population size in total= argument
newdata2 <- data.frame(photo=1130)
predict(myfit1, newdata2, total=100)
```

```
    link     SE
1 1198.9 41.802
```

```
confint(predict(myfit1, newdata2,total=100),df=23)
```

```
     2.5 %   97.5 %
1 1112.455 1285.404
```

Subdomain estimation

- Often we want separate estimates for subpopulations; the subpopulations are called domains or subdomains.
- The number of persons in an SRS who fall into each domain n_d is a random variable, so estimating domain means is a special case of ratio estimation.
- Let $x_i = 1$ and $u_i = x_i y_i = y_i$ if $i \in s_d$ and $x_i = u_i = 0$ otherwise
- Domain mean $\bar{y}_d = \frac{\hat{t}_u}{\hat{t}_x} = \frac{\bar{u}}{\bar{x}} = \hat{B}$

Subdomain estimation cont.

- Using the variance of $\hat{B}$ in Equation (2),

$$SE(\bar{y}_d) = \sqrt{(1 - \frac{n}{N})\frac{n}{n_d^2}\frac{(n_d - 1)s_{yd}^2}{n - 1}},$$

where $s_{yd}^2 = \sum_{i \in s_d}(y_i - \bar{y}_d)^2/(n_d - 1)$

Subdomain estimation cont.

- Using the variance of $\hat{B}$ in Equation (2),

$$SE(\bar{y}_d) = \sqrt{(1 - \frac{n}{N}) \frac{n}{n_d^2} \frac{(n_d - 1)s_{yd}^2}{n - 1}},$$

where $s_{yd}^2 = \sum_{i \in s_d} (y_i - \bar{y}_d)^2 / (n_d - 1)$

- If $E(n_d)$ is large, $SE(\bar{y}_d) \approx \sqrt{(1 - \frac{n}{N}) \frac{s_{yd}^2}{n_d}}$

Subdomain estimation cont.

- Using the variance of $\hat{B}$ in Equation (2),

$$SE(\bar{y}_d) = \sqrt{(1 - \frac{n}{N})\frac{n}{n_d^2}\frac{(n_d - 1)s_{yd}^2}{n-1}},$$

where $s_{yd}^2 = \sum_{i \in s_d}(y_i - \bar{y}_d)^2/(n_d - 1)$

- If $E(n_d)$ is large, $SE(\bar{y}_d) \approx \sqrt{(1 - \frac{n}{N})\frac{s_{yd}^2}{n_d}}$
- The results for now only apply to SRSs, and approximations depends on having a sufficiently large sample so that $E(n_d)$ is large

Subdomain estimation cont.

- Using the variance of $\hat{B}$ in Equation (2),

$$SE(\bar{y}_d) = \sqrt{(1 - \frac{n}{N})\frac{n}{n_d^2}\frac{(n_d - 1)s_{yd}^2}{n-1}},$$

where $s_{yd}^2 = \sum_{i \in s_d}(y_i - \bar{y}_d)^2/(n_d - 1)$

- If $E(n_d)$ is large, $SE(\bar{y}_d) \approx \sqrt{(1 - \frac{n}{N})\frac{s_{yd}^2}{n_d}}$
- The results for now only apply to SRSs, and approximations depends on having a sufficiently large sample so that $E(n_d)$ is large
- More generally, we use small area estimation approaches that rely on models

R code: Lohr Example 4.8

```
agsrsnew<-agsrs
agsrsnew$farmcat<-rep("large",n)
agsrsnew$farmcat[agsrsnew$farms92 < 600] <- "small" #head(agsrsnew)
dsrsnew <- svydesign(id = ~1, weights=~sampwt, fpc=rep(3078,300), data=agsrsnew)
# domain estimation for large farmcat with subset statement
dsub1<-subset(dsrsnew,farmcat=='large')   # design info for domain large farmcat
smean1<-svymean(~acres92,design=dsub1); smean1
```

```
          mean    SE
acres92 316566 21553
```

```
df1<-sum(agsrsnew$farmcat=='large')-1; df1 #calculate domain df if desired
```

```
[1] 128
```

```
confint(smean1, level=.95,df=df1) # CI
```

```
           2.5 %   97.5 %
acres92 273918.9 359212.4
```

```
# use svyby function
bothmeans<-svyby(~acres92,by=~factor(farmcat),design=dsrsnew,svymean); bothmeans
```

```
      factor(farmcat)  acres92       se
large           large 316565.7 21553.21
small           small 283813.7 28852.24
```

```
confint(bothmeans,level=.95)
```

```
         2.5 %   97.5 %
large 274322.1 358809.2
small 227264.4 340363.1
```

Poststratification

- Since poststrata are formed after data collection, the sample domain sizes are random quantities.

Poststratification

- Since poststrata are formed after data collection, the sample domain sizes are random quantities.
- The poststratification estimator of $\bar{Y}$ is $\bar{y}_{post} = \sum_h N_h/N\bar{y}_h$ is ratio estimation

Poststratification

- Since poststrata are formed after data collection, the sample domain sizes are random quantities.
- The poststratification estimator of $\bar{Y}$ is $\bar{y}_{post} = \sum_h N_h/N\bar{y}_h$ is ratio estimation
- If n_h is reasonably large, we can use an approximate variance estimator $var(\bar{y}_{post}) \approx (1 - \frac{n}{N}) \sum N_h/N s_h^2/n$

Poststratification

- Since poststrata are formed after data collection, the sample domain sizes are random quantities.
- The poststratification estimator of $\bar{Y}$ is $\bar{y}_{post} = \sum_h N_h/N\bar{y}_h$ is ratio estimation
- If n_h is reasonably large, we can use an approximate variance estimator $var(\bar{y}_{post}) \approx (1 - \frac{n}{N}) \sum N_h/N s_h^2/n$
- Difference between stratification (design) and poststratification (estimation): n_h fixed or random?

R code: Lohr Example 4.9

```
data(agsrs)
dsrs <- svydesign(id = ~1, weights=rep(3078/300,300), fpc=rep(3078,300),data = agsrs)
# Create a data frame that gives the population totals for the poststrata
pop.region <- data.frame(region=c("NC","NE","S","W"), Freq=c(1054,220,1382,422))
# create design information with poststratification
dsrsp<-postStratify(dsrs, ~region, pop.region); summary(dsrsp)
```

```
Independent Sampling design
postStratify(dsrs, ~region, pop.region)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.09242 0.09407 0.09407 0.09771 0.10152 0.10909
Population size (PSUs): 3078
Data variables:
 [1] "county"   "state"    "acres92"  "acres87"  "acres82"  "farms92"
 [7] "farms87"  "farms82"  "largef92" "largef87" "largef82" "smallf92"
[13] "smallf87" "smallf82" "region"
```

```
1/unique(dsrsp$prob)  # See the poststratified weight for each region
```

```
[1] 10.630769 10.820513  9.850467  9.166667
```

```
svymean(~acres92, dsrsp)
```

```
          mean    SE
acres92 299778 17513
```

```
svytotal(~acres92, dsrsp)
```

```
            total       SE
acres92 922717031 53906392
```

Summary

- Ratio and regression estimation use an auxiliary variable that is highly correlated with the variable of interest to reduce the MSE of estimated population means or totals.
- The estimators in ratio and regression estimation come from models that we hope describe the data, but the randomization-theory properties of the estimators do not depend on these models.
- Ratio estimation is especially useful in cluster sampling as we shall see in the next sessions