

Ratio Estimation (Chapter 7)

This handout covers the basic idea behind ratio estimation, gives the forms and properties of the relevant estimators, compares ratio estimation to other estimation methods studied, and provides some examples which use ratio estimation.

Example: Reconsider the farm example where we were interested in estimating:

1. the population total $\tau =$ the total # of workers, and
2. the population mean $\mu =$ average # of workers per farm.

In the last section, we used “size” as an auxiliary variable in the design phase of the study where we used PPS sampling to select farms. This was not only convenient (because we didn’t have a list of all the farms from which to draw an SRS), but was advantageous because the number of workers was positively correlated with size. The Hansen-Hurwitz estimator based on a PPS sample has smaller variance than the estimator based on an SRS if there’s a strong positive relationship between the size variable and the response variable (see separate handout on how PPS sampling does relative to SRS for the farm data).

Another way we could use the auxiliary variable “size” is in the estimation stage, after we have collected the data from an SRS. We can do this is through a ratio estimator. Like PPS sampling, a ratio estimator is advantageous only if there is a strong positive relationship between the auxiliary variable and the response variable. Specifically, ratio estimation is optimal when there is a linear relationship through the origin between the two variables.

It’s important to note that in order to use an auxiliary variable x in a ratio estimator to estimate τ or μ for the y variable, then we need to know τ_x , the total value of x for the whole population. Ratio estimation is therefore commonly used when the auxiliary variable is a variable which is easily measured on the whole population while the response variable is harder to measure and is obtained from only an SRS of the population. Some situations where ratio estimation might be beneficial are:

- Let $x =$ the girth of a tree, and $y =$ the volume of the tree
- Let $x =$ the total # of animals on a plot of land, and $y =$ the # of females.
- Let $x =$ total volume of a haul of fish, and $y =$ the number fish in the haul .
- Let $x =$ a visual estimate of the % of some ground cover, and $y =$ the actual % of some ground cover.

Example 1: Consider the second situation above, where the population consists of $N = 20$ plots of land, and we take an SRS of $n = 7$ plots, counting the number of animals and number of females on these 7 plots. In addition, we also count the number of animals on all

20 plots, without knowing what sex they are because it may be easy to count the number of animals on a plot, but hard to identify which are females. Assume all the plots are equal in size. Primary interest here is in estimating either:

$$\begin{aligned}\mu_y &= \text{the average number of females per plot of land, or} \\ \tau_y &= \text{the total number of females} = N\mu.\end{aligned}$$

Response Variable: y_i = the number of female animals on plot i , $i = 1, \dots, N$,

Auxiliary Variable: x_i = the total number of animals on plot i , $i = 1, \dots, N$.

	x_i	y_i	y_i/x_i
The data for the SRS of size 7 are given to the right:	10	7	.7
	18	12	.67
	10	4	.4
1. <u>First, estimate μ_y, τ_y without using the auxiliary variable:</u>	12	6	.5
	25	19	.76
$\bar{y} = \frac{60}{7} = \underline{8.57}$, $s = 5.26$, $\widehat{SE}(\bar{y}) = \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}} = \underline{1.60}$,	15	7	.467
$\hat{\tau}_y = N\bar{y} = 20(8.57) = \underline{171.4}$, $SE(\hat{\tau}_y) = NSE(\bar{y}) = \underline{32.0}$.	10	5	.5

- These estimates will later be compared to those obtained through ratio estimation.
2. Estimate the overall proportion of females in all 20 plots: Since y_i is the number of females in plot i and x_i is the number of animals in plot i , then to estimate the proportion of females, we look at the ratio y_i/x_i .

- We might consider then taking the average of these ratios from the sample. Any problem with this?

- It is better to take the ratio of the means, than the mean of the ratios.

Definition: The population ratio R and sample ratio r are given respectively as:

$$R = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}, \quad r = \frac{\sum y_i}{\sum x_i} = \frac{\bar{y}}{\bar{x}}.$$

- Expected value of r : $E(r) \neq \frac{E(\bar{y})}{E(\bar{x})} = \frac{\mu_y}{\mu_x} = R$, so in general, r is not an unbiased estimator of R . For most cases, however, the bias is small.

- Variance of r : The variance is approximated by:

$$\text{Var}(r) \approx \left(\frac{N-n}{N}\right) \frac{1}{\mu_x^2} \cdot \frac{\sigma_r^2}{n} \quad \text{where} \quad \sigma_r^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - Rx_i)^2,$$

with an estimated variance given by:

$$\widehat{\text{Var}}(r) = \left(\frac{N-n}{N}\right) \frac{1}{\mu_x^2} \cdot \frac{s_r^2}{n} \quad \text{where} \quad s_r^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2.$$

The sample mean \bar{x} can be used in place of μ_x in the above expression if μ_x is not known.

- This approximation is based on a Taylor series expansion of the ratio \bar{y}/\bar{x} and will be discussed later.
- When will this variance be small?

- Normally, since the estimate r of the population ratio R is biased, we would use the $\text{MSE}(r) = \text{Var}(r) + \text{Bias}^2(r)$ to compare the ratio estimator to other estimators. However, the squared bias is generally very small, so it is often ignored.

Recall the table of animal counts given earlier, and consider the augmented table below to compute the estimated ratio and its variance:

x_i (# animals)	y_i (# females)	rx_i	$(y_i - rx_i)^2$
10	7		
18	12		
10	4		
12	6		
25	19		
15	7		
10	5		

- Instead of $\widehat{\text{Var}}(r)$, why wouldn't we just use the sample proportion $\hat{p} = 60/100 = 0.6$ and its standard error based on the SRS formula?

$$\text{SE}(\hat{p}) = \sqrt{\left(\frac{N-n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\left(\frac{N-100}{N}\right) \frac{.60(1-.60)}{100-1}} \approx .0416 \text{ (with } N = 350\text{)}.$$

- This assumes we have taken an SRS of 100 animals, which is not true. What type of sample of animals have we taken?

- In looking at the form of the standard error of the estimator of r , it should be clear that the estimator will be “good” when the $y_i - rx_i$ are “small.” If $y_i = rx_i$, then there is a linear relationship between y & x through the origin.
- This seems reasonable for the animals/females example, as for 0 animals there would be 0 females. And as the number of animals increases, we would expect the number of females to increase linearly with it.
- We can, and should, examine the relationship between x and y for our sample with a scatterplot.

3. Ratio Estimator of μ_y

- As was done in the first part of this example, \bar{y} is the (SRS) naive estimate of μ .
- Suppose we are given:

$$\begin{aligned} \bar{x} &= \text{mean \# of animals per plot in the sample} \\ \mu_x &= \text{mean \# of animals per plot in entire population} \\ \bar{y} &= \text{mean number of females per plot in the sample} \end{aligned}$$

- The idea of a ratio estimator is to “adjust” the naive estimator \bar{y} using the relationship between y & x . Recall that $R = \mu_y/\mu_x$ so

$$\mu_y = \left(\frac{\mu_y}{\mu_x}\right) \mu_x = R\mu_x.$$

We replace R by its estimator $r = \bar{y}/\bar{x}$ to give an estimator of the population mean μ_y (mean # of females per plot):

$$\hat{\mu}_r = r \cdot \mu_x = \bar{y} \cdot \frac{\mu_x}{\bar{x}},$$

with corresponding variance and estimated variance given by:

$$\text{Var}(\hat{\mu}_r) =$$

=

$$\widehat{\text{Var}}(\hat{\mu}_r) = \left(\frac{N-n}{N}\right) \frac{s_r^2}{n}, \text{ where } s_r^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2.$$

- Note that we need to know μ_x to use ratio estimation to improve the estimate of μ_y .
- An alternative estimator of variance is given as equation (7) on page 69 of the text. This alternative estimator is more robust to the value of \bar{x} than the variance estimator given above.
- Note that the estimated variance of $\hat{\mu}_r$ given above has the same form as the estimated variance of \bar{y} (the conventional estimator of the mean), except that s^2 is replaced by s_r^2 . This implies that whenever s_r^2 is smaller than s^2 , the ratio estimator will be superior to the conventional SRS-based estimator. When will this be true?

4. Ratio Estimator of τ_y

Since the population total $\tau_y = N\mu_y = NR\mu_x = R\tau_x$, an estimator of the population total (total # of females) is given by:

$$\hat{\tau}_r = r \cdot \tau_x = \frac{\bar{y}}{\bar{x}} \tau_x \quad (\text{where } \tau_x \text{ is assumed known}),$$

with corresponding variance and estimated variance given by:

$$\text{Var}(\hat{\tau}_r) = N^2 \text{Var}(\hat{\mu}_r) = N(N-n) \frac{\sigma_r^2}{n}, \quad \widehat{\text{Var}}(\hat{\tau}_r) = N(N-n) \frac{s_r^2}{n},$$

where s_r^2 was given earlier. Suppose $\tau_x = 350$ (total # animals). Then:

$$\begin{aligned} \hat{\tau}_r &= r \cdot \tau_x = (.6)(350) = \underline{210 \text{ females}}. \quad (\text{With SRS, } \hat{\tau} = 171.4). \\ \widehat{\text{Var}}(\hat{\tau}_r) &= N(N-n) \frac{s_r^2}{n} = 20(20-7) \frac{4.813}{7} = 178.77, \quad \text{so that:} \\ \text{SE}(\hat{\tau}_r) &= \sqrt{178.77} = \underline{13.37}. \quad (\text{With SRS, } \text{SE}(\hat{\tau}) = 32.0). \end{aligned}$$

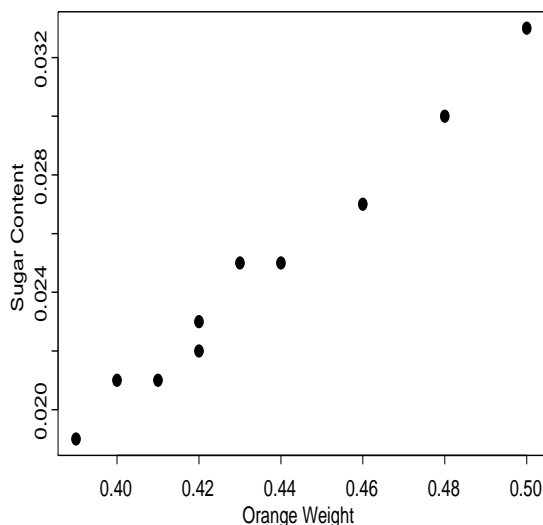
- The comparison of standard errors for \bar{y} and the ratio estimate emphasizes the gains to be had with ratio estimation when the response and auxiliary variables are linearly related through the origin.

Example 2: In a study to estimate the total sugar content of a truckload of oranges, a random sample of $n = 10$ oranges was juiced and weighed. The data for the 10 oranges are given in the table below and displayed in a plot of sugar content versus weight. The total

Orange	Sugar Content (in pounds)	Weight of Orange (in pounds)
1	.021	.40
2	.030	.48
3	.025	.43
4	.022	.42
5	.033	.50
6	.027	.46
7	.019	.39
8	.021	.41
9	.023	.42
10	.025	.44

$$\sum_{i=1}^{10} y_i = .246 \qquad \sum_{i=1}^{10} x_i = 4.35$$

Sugar Content vs. Weight - Orange Example



weight of all the oranges, obtained by first weighing the truck loaded and then unloaded, was found to be 1800 pounds. Estimate τ_y , the total sugar content for the oranges, and place a bound on the error of estimation. In this example, the sugar content of an orange (y) is the response and the weight of an orange (x) is the auxiliary variable.

- Note that if we ignore the auxiliary variable weight here, we cannot estimate the total sugar content τ_y as requested using basic SRS ideas, because we don't know the population size $N = \text{total } \# \text{ of oranges}$. (i.e.: the usual estimator of τ_y is: $\hat{\tau} = N\bar{y}$, but here we don't know N).
- Here then, is a case where we must use a ratio estimator.
- What do we know?

- The estimated variance of $\hat{\tau}_r$ is: $\widehat{\text{Var}}(\hat{\tau}_r) = N(N - n)\frac{s_r^2}{n}$. Any problem here? What do we do?

Computing:

$$r = \frac{\bar{y}}{\bar{x}} = \frac{\sum y_i}{\sum x_i} = \frac{.246}{4.35} = \underline{.05655},$$

$$\begin{aligned} s_r^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2 = \frac{1}{9} [(.021 - .05655(.40))^2 + \cdots + (.025 - .05655(.44))^2] \\ &= \underline{(.00241)^2}, \end{aligned}$$

$$\begin{aligned} \widehat{\text{Var}}(\hat{\tau}_r) &= \tau_x^2 \widehat{\text{Var}}(r) \approx (1800)^2 \frac{1}{(.435)^2} \frac{(.00241)^2}{10} \\ &= \underline{9.949} \implies \text{SE}(\hat{\tau}_r) = \underline{3.15 \text{ pounds}}. \end{aligned}$$

- An approximate 95% confidence interval for the total sugar content τ_y is:

$$\hat{\tau}_r \pm t_9(.975) \cdot \text{SE}(\hat{\tau}_r) = 101.79 \pm (2.262)(3.15) = \underline{(94.66, 108.92) \text{ pounds}}.$$

```

> # Example 1: ratio estimation of number and proportion of females
> x <- c(10,18,10,12,25,15,10)
> y <- c(7,12,4,6,19,7,5)
> n <- 7
> N <- 20
> plot(x,y,pch=16,xlab="Number of animals",ylab="Number of females")

> # Estimation of proportion of females
> r <- mean(y)/mean(x)
> r
[1] 0.6
> sr2 <- (1/(n-1))*sum((y-r*x)^2)
> sr2
[1] 4.813333
> SE.r <- sqrt((1-n/N)*sr2/(mean(x)^2*n)) # assumes mu_x not known
> SE.r
[1] 0.04679815
> c(r - qt(.975,n-1)*SE.r,r + qt(.975,n-1)*SE.r)
[1] 0.4854891 0.7145109

> # Estimation of total number of females
> tau.x <- 350 # total number of animals on all 20 plots (given)
> tau.hat <- r*tau.x
> SE.tau <- sqrt(N*(N-n)*sr2/n)
> c(tau.hat,SE.tau)
[1] 210.0000 13.3709
> c(tau.hat - qt(.975,n-1)*SE.tau,tau.hat + qt(.975,n-1)*SE.tau)
[1] 177.2826 242.7174

```

