Wednesday, Sep 25

Comparison of Two Estimators of a Ratio of Totals

Consider *two* estimators of $R = \tau_y/\tau_x$:

$$r = \bar{y}/\bar{x} \quad \text{and} \quad r' = \bar{y}/\mu_x.$$

The latter would be an option if μ_x is *known*.

Example: Consider the problem of estimating the density of larkspur in a region, where y_i is the number of larkspur in the i-th plot, and x_i is the area of the i-th plot. Here μ_x is the mean area of all plots. We may know this.

Example: Consider the problem of estimating the number of households with televisions, where y_i is the number of households that own a television in the i-th block, and x_i is the number of households in the i-th block. Here τ_x is the total number of households in the city, so $\mu_x = \tau_x/N$. We may know this.

Which estimator should we use? Under simple random sampling the estimated variances of these estimators are

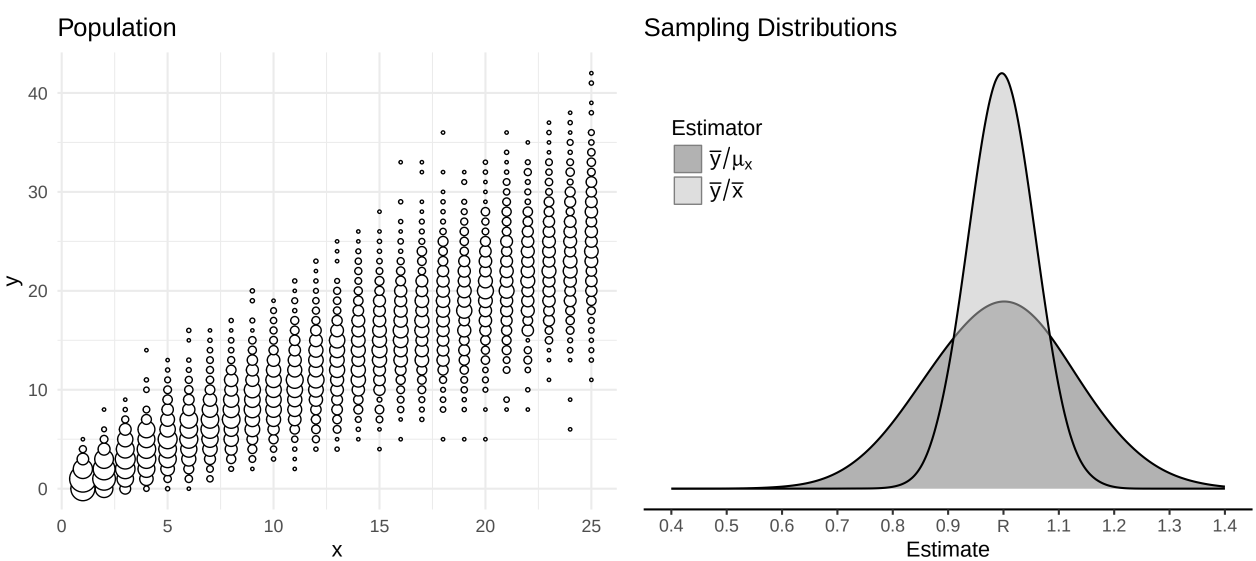
$$\hat{V}(\bar{y}/\bar{x}) = \frac{1}{\mu_x^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} \quad \text{and} \quad \hat{V}(\bar{y}/\mu_x) = \frac{1}{\mu_x^2}\left(1 - \frac{n}{N}\right)\frac{s^2}{n},$$

where

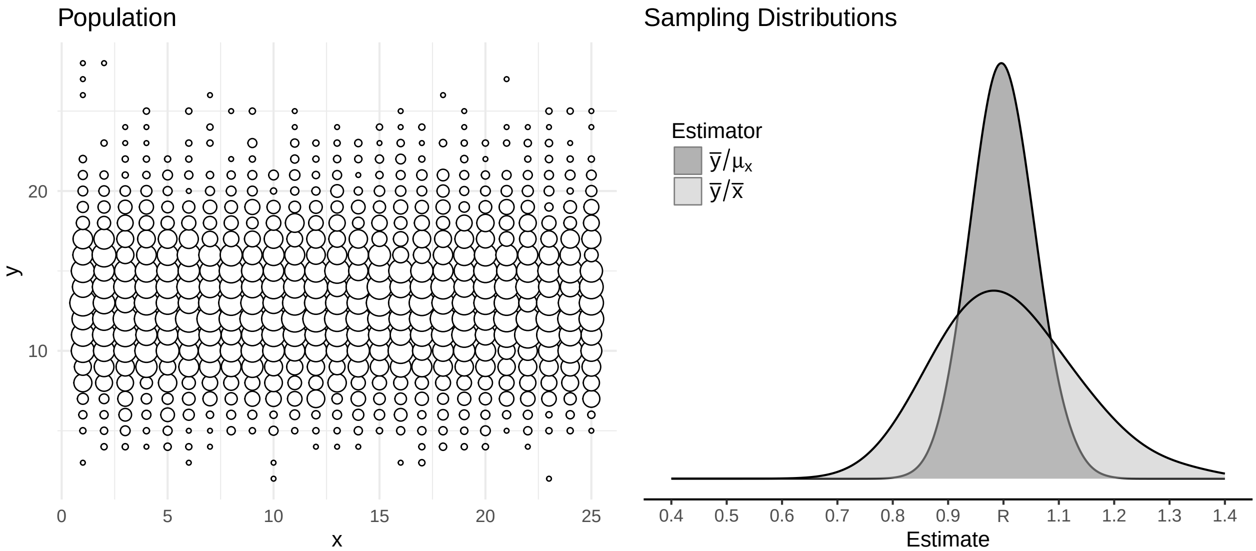
$$s_r^2 = \frac{1}{n-1}\sum_{i \in \mathcal{S}}(y_i - rx_i)^2 \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i \in \mathcal{S}}(y_i - \bar{y})^2,$$

so the (estimated) relative efficiency depends only on s_r^2 and s^2.

Approximately Proportional Variables



Unrelated Variables



Question: Assuming we know μ_x, when should we use the estimator $r = \bar{y}/\bar{x}$ and when should we use the estimator $r' = \bar{y}/\mu_x$?

Auxiliary Variables

An **auxiliary variable** is a variable (x_i) that varies over elements or sampling units. Auxiliary variables can be used in a variety of ways.

Two common uses of auxiliary variables:

1. In specifying a *sampling design*.
2. In defining an *estimator*.

For example, stratified random sampling uses an auxiliary variable for stratification, and post-stratification uses an auxiliary variable (i.e., the strata membership) in the estimator. *Ratio estimators* are an example of using an auxiliary variable *in an estimator*.

Ratio Estimators

Consider that for a simple random sampling design

$$\frac{\mu_y}{\mu_x} \approx \frac{\bar{y}}{\bar{x}} \Rightarrow \mu_y \approx \frac{\bar{y}}{\bar{x}}\mu_x,$$

so we could define an estimator of μ_y as

$$\hat{\mu}_y = \frac{\bar{y}}{\bar{x}}\mu_x.$$

Multiplying this estimator would give us an estimator of τ_y since $\tau_y = N\mu_y$, so that would be

$$\hat{\tau}_y = \frac{\bar{y}}{\bar{x}}\tau_x.$$

The **ratio estimators** of τ_y and μ_y for a simple random sampling design are

$$\hat{\tau}_y = \frac{\bar{y}}{\bar{x}}\tau_x \quad \text{and} \quad \hat{\mu}_y = \frac{\bar{y}}{\bar{x}}\mu_x,$$

respectively, where y denotes the *target variable* and x denotes an *auxiliary variable*. These are alternatives to the estimators,

$$\hat{\tau}_y = N\bar{y} \quad \text{and} \quad \hat{\mu}_y = \bar{y},$$

which do not use an auxiliary variable.

Note: There is a connection between a *ratio estimator* and the *estimator of a ratio* ($r = \bar{y}/\bar{x}$) since we can write the ratio estimators as $\hat{\mu}_y = r\mu_x$ and $\hat{\tau}_y = r\tau_x$.

Example: Consider a survey to estimate the proportion of households with televisions in Des Moines, Iowa, in 1951. The sampling units are *blocks* of households, selected using simple random sampling from a population of $N = 9460$ blocks.

block	tv	households
1	1	3
2	0	8
3	1	8
4	0	7
5	4	7
$\vdots$	$\vdots$	$\vdots$
132	2	4

Let y_i and x_i denote the number of households with televisions and the number of households, respectively, for the i-th block. We can compute $\bar{y} \approx 1.8$ and $\bar{x} \approx 5.23$. We also know that the total number of households for all 9460 blocks is $\tau_x = 46296$. What are our estimates of the total number of households with televisions in Des Moines using the two estimators we have?

Example: The area of each of 744 leaves from a shining gum (*Eucalyptus nitens*) was crudely approximated by multiplying leaf length and width. The area of each leaf of a simple random sample of 20 leaves was measured accurately.

i	Area	Length $\times$ Width
1	80.7	113.08
2	69.7	98.4
3	66.1	97.17
4	124.6	198.4
5	72.6	103.2
$\vdots$	$\vdots$	$\vdots$
20	29.4	40

The mean crude area approximation of all 744 leaves was 76.97 square cm. For the sample of 20 leaves, the mean of the accurate area measurements was 52.12 square cm, and the mean of the crude area measurements was 74.39 square cm. What is the estimate of the mean area of all the leaves using the two estimators?

The (Estimated) Variance of Ratio Estimators

We have two estimators of μ_y and two estimators of τ_y under simple random sampling. How do they compare?

Let $\mu_y = \mu_x \bar{y}/\bar{x}$ be the ratio estimator of μ_y. Under simple random sampling, the estimated variance of $\hat{\mu}_y$ is

$$\hat{V}(\hat{\mu}_y) = \left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} \quad \text{where} \quad s_r^2 = \frac{\sum_{i \in \mathcal{S}}(y_i - rx_i)^2}{n-1},$$

where $r = \bar{y}/\bar{x}$. Note that the estimated variance of μ_y is very close related to that of r. This isn't surprising since we can write the ratio estimator as $\hat{\mu}_y = r\mu_x$. The μ_x cancels out the term $1/\mu_x^2$ in the variance of r to create the variance of μ_y.

Example: Recall the survey of shining gum leaves. We have that $\sum_{i \in \mathcal{S}}(y_i - rx_i)^2 \approx 509.33$. What is the bound on the error of estimation of the ratio estimator of the mean leaf area?

Let $\hat{\tau}_y = \tau_x \bar{x}/\bar{y}$ be the ratio estimator of τ_y under simple random sampling. The estimated variance of $\hat{\tau}_y$ is

$$\hat{V}(\hat{\tau}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_r^2}{n} \quad \text{where} \quad s_r^2 = \frac{\sum_{i \in \mathcal{S}} (y_i - r x_i)^2}{n-1},$$

where $r = \bar{y}/\bar{x}$.

Example: Recall the television survey. There $\sum_{i \in \mathcal{S}} (y_i - r x_i)^2 \approx 210.63$. What is the bound on the error of estimation for the ratio estimator for the total number of households with televisions?

Relative Efficiency of Ratio Estimators

Under simple random sampling the estimator $\hat{\mu}_y = \bar{y}$ has the (estimated) variance

$$\hat{V}(\hat{\mu}_y) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \quad \text{where} \quad s^2 = \frac{\sum_{i \in \mathcal{S}}(y_i - \bar{y})^2}{n-1},$$

while the ratio estimator $\hat{\mu}_y = \mu_x \bar{y}/\bar{x}$ has an (estimated) variance of

$$\hat{V}(\hat{\mu}_y) = \left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} \quad \text{where} \quad s_r^2 = \frac{\sum_{i \in \mathcal{S}}(y_i - rx_i)^2}{n-1}.$$

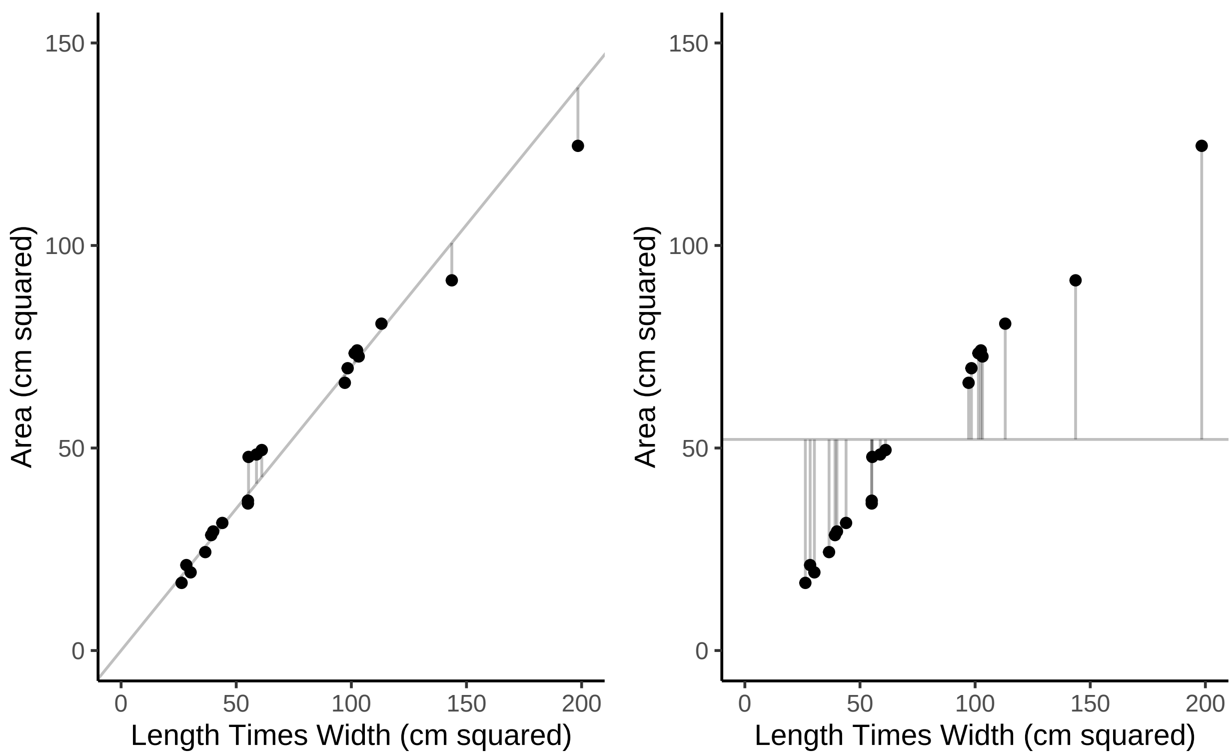
The (estimated) *relative efficiency* of the ratio estimator when compared to $\bar{y}$ is

$$\frac{\hat{V}(\bar{y})}{\hat{V}(\mu_x \bar{y}/\bar{x})} = \frac{\sum_{i \in \mathcal{S}}(y_i - \bar{y})^2}{\sum_{i \in \mathcal{S}}(y_i - rx_i)^2}.$$

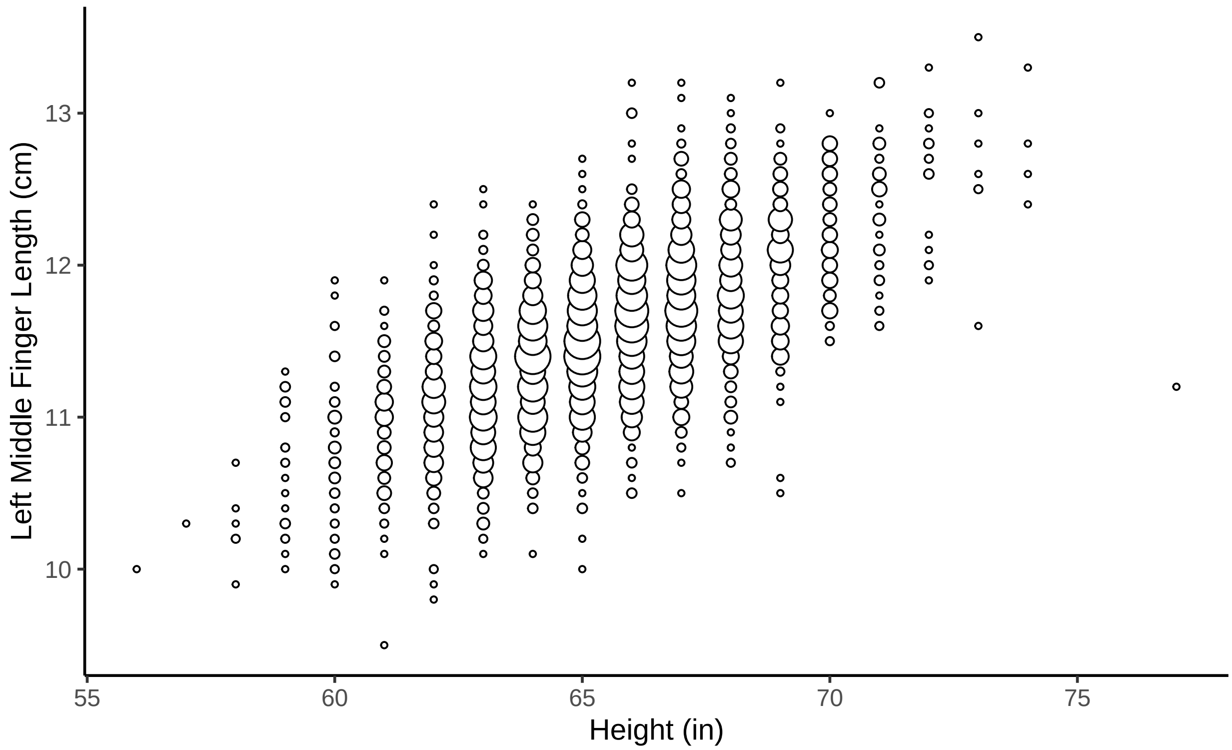
So a relative efficiency greater than one favors the ratio estimator, whereas a relative efficiency less than one favors $\bar{y}$.

Note: We can also compute an *effective sample size* of the ratio estimator (relative to $\bar{y}$) as the sample size times the relative efficiency.

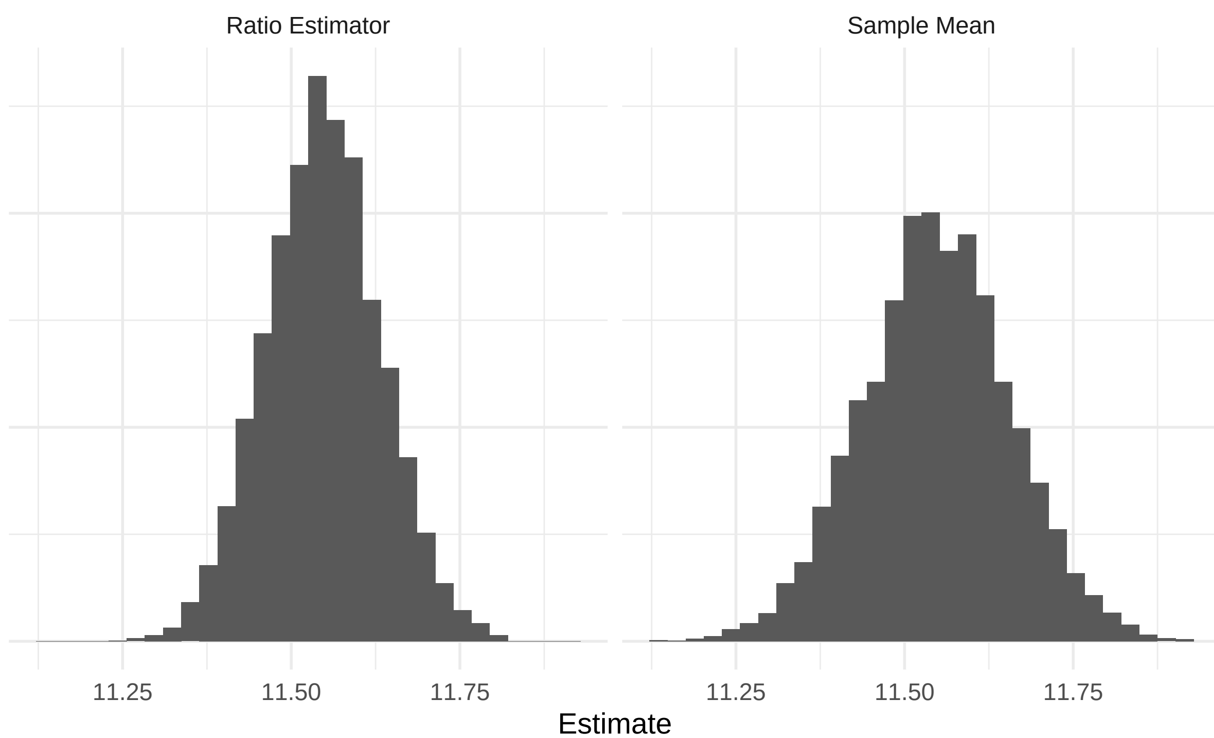
Example: Recall the survey of leaf area. We have $\sum_{i \in \mathcal{S}}(y_i - rx_i)^2 \approx 509.33$ and $\sum_{i \in \mathcal{S}}(y_i - \bar{y})^2 \approx 15558.43$. This gives us a relative efficiency of approximately 30.55.



Example: Consider the observations of a target variable (left middle finger length) and an auxiliary variable (height) for a population of 3000 elements.

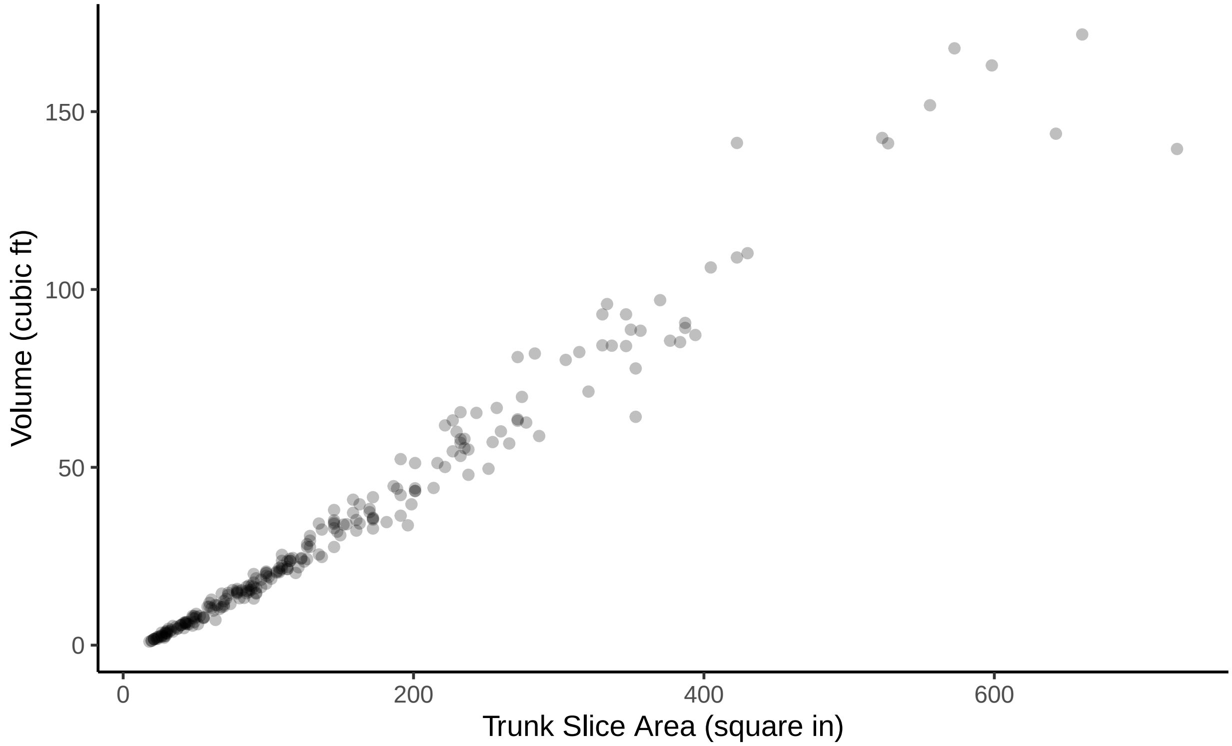


Let's simulate the sampling distributions of two estimators of μ_y under simple random sampling with $n = 25$: the sample mean $\bar{y}$ and the ratio estimator $\mu_x \bar{y}/\bar{x}$. Note that the value of μ_y is approximately 11.55 cm.

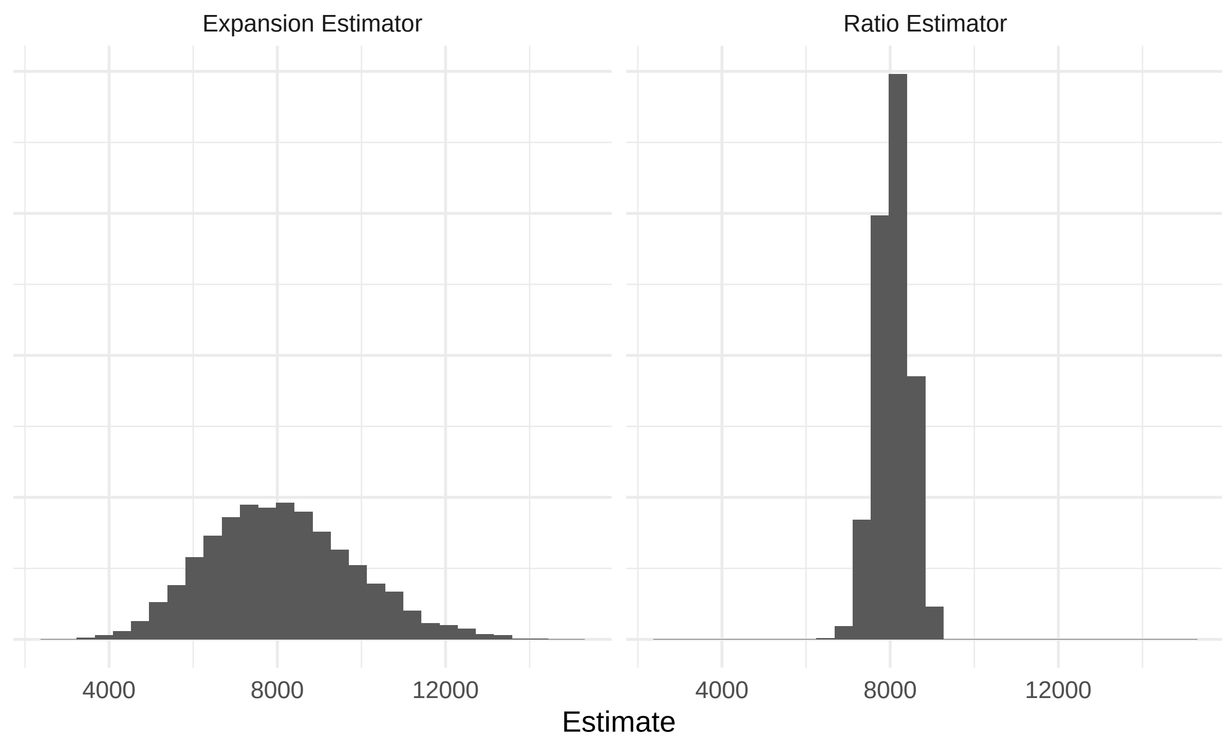


We have $\hat{V}(\bar{y}) \approx 0.012$ and $\hat{V}(\mu_x \bar{y}/\bar{x}) \approx 0.007$.

Example: Consider a population of red oaks.



Let's simulate the sampling distributions of two estimators of τ_y under simple random sampling with $n = 20$: the estimator $N\bar{y}$ (i.e., the "expansion estimator") and the ratio estimator $\tau_x \bar{y}/\bar{x}$. Note that the value of τ_y is 8124.9 cubic feet.



We have $\hat{V}(N\bar{y}) \approx 3113433$ and $\hat{V}(\tau_x \bar{y}/\bar{x}) \approx 160054.2$.

Question: Could the ratio estimator for τ_y or μ_y have a *larger* variance than the "non-ratio" estimators $\hat{\tau}_y = N\bar{y}$ and $\hat{\mu}_y = \bar{y}$, respectively, and thus *lower* relative efficiency? *When* would this happen?

Unknown N

One estimator of τ_y under simple random sampling is

$$\hat{\tau}_y = N\bar{y},$$

which is clearly useless if N is *unknown*. But we could use the ratio estimator

$$\hat{\tau}_y = \frac{\bar{y}}{\bar{x}}\tau_x,$$

if we know τ_x. Recall that $\tau_x = \sum_{i=1}^N x_i$. So one way to compute τ_x is to get x_i for *all* elements in the population, but if we could do that we'd know N. But in some cases there are other ways to find τ_x.

Example: Suppose we have a load of an unknown number of oranges. Let y_i be the sugar content of the i-th orange and let x_i be the weight of the i-th orange. The ratio estimator of the total sugar content of the load of oranges (assuming a simple random sampling design) is

$$\hat{\tau}_y = \frac{\bar{y}}{\bar{x}}\tau_x,$$

where τ_x is the total weight of all the oranges in the load.

Example: The following data are from hauls of Ventura Marsh, Iowa, in October of 1953 for Northern Pike (*Esox lucius*) after 1146 pike had been caught, tagged, and released.

Haul	Catch	Tagged
1	493	11
2	1584	57
3	1275	80
4	1488	60
5	1070	32
⋮	⋮	⋮
16	331	10

The mean catch size was 949.69 pike and the mean number of tagged pike per haul was 37.25. If we regard these hauls as a sample from a simple random sampling design, then an estimate of the total number of Northern Pike is

$$\hat{\tau}_y = \frac{\bar{y}}{\bar{x}}\tau_x,$$

where $\bar{y}$ and $\bar{x}$ are the average number of caught and tagged fish in the sample of hauls, respectively, and τ_x is the total number of tagged fish.[1]

[1] If N is unknown we cannot compute the estimated variance of the ratio estimator, but if we can assume that $1 - n/N \approx 1$ (or we can omit the term $1 - n/N$ due to sampling *with replacement*), then we can estimate N with $\tau_x/\bar{x}$ (since τ_x can be estimated as $N\bar{x}$ and so $\tau_x \approx N\bar{x} \Rightarrow N \approx \tau_x/\bar{x}$) and the estimated variance

$$\hat{V}(\hat{\tau}_y) = N^2\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} \quad \text{where} \quad s_r^2 = \frac{\sum_{i\in\mathcal{S}}(y_i - rx_i)^2}{n-1},$$

becomes

$$\hat{V}(\hat{\tau}_y) = \frac{\tau_x^2 s_r^2}{\bar{x}^2 n}.$$

Advantages of Ratio Estimators

In the context of a simple random sampling design, what *two* advantages do we see for ratio estimators relative to the estimators that do not use an auxiliary variable?