Friday, Sep 27

Two Estimators of a Domain Total Revisited

We saw that for a simple random sampling design there are two estimators of τ_d:

$$\hat{\tau}_d = N_d\bar{y}_d \quad \text{and} \quad \hat{\tau}_d = \frac{N}{n}n_d\bar{y}_d.$$

The first has smaller variance, although it requires knowing N_d. How can we explain the difference in variance using what we know about ratio estimators?

Consider that

$$\bar{y}_d = \frac{\sum_{i\in\mathcal{S}} y_i'}{\sum_{i\in\mathcal{S}} x_i},$$

where

$$y_i' = \begin{cases} y_i, & \text{if the } i\text{-th element is from the domain,} \\ 0, & \text{otherwise,} \end{cases}$$

and

$$x_i = \begin{cases} 1, & \text{if the } i\text{-th element is from the domain,} \\ 0, & \text{otherwise.} \end{cases}$$

Also note that $N_d = \tau_x = \sum_{i=1}^N x_i$ and $n_d = \sum_{i\in\mathcal{S}} x_i$. So we can write these estimators as

$$N_d\bar{y}_d = \tau_x \frac{\sum_{i\in\mathcal{S}} y_i'}{\sum_{i\in\mathcal{S}} x_i} = \tau_x \frac{\frac{1}{n}\sum_{i\in\mathcal{S}} y_i'}{\frac{1}{n}\sum_{i\in\mathcal{S}} x_i} = \tau_x \frac{\bar{y}'}{\bar{x}}$$

and

$$\frac{N}{n}n_d\bar{y}_d = \frac{N}{n}n_d\frac{\sum_{i\in\mathcal{S}} y_i'}{\sum_{i\in\mathcal{S}} x_i} = \frac{N}{n}n_d\frac{\sum_{i\in\mathcal{S}} y_i'}{n_d} = \frac{N}{n}\sum_{i\in\mathcal{S}} y_i' = N\bar{y}'.$$

And note that y_i' is "approximately proportional" to x_i since $y_i' = 0$ if $x_i = 0$. So now why does the estimator $N_d\bar{y}_d$ tend to have a smaller variance than the estimator $(N/n)n_d\bar{y}_d$?

Ratio Estimators as Adjusted Estimators

Consider two estimators of μ_y:

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

Writing the ratio estimator as

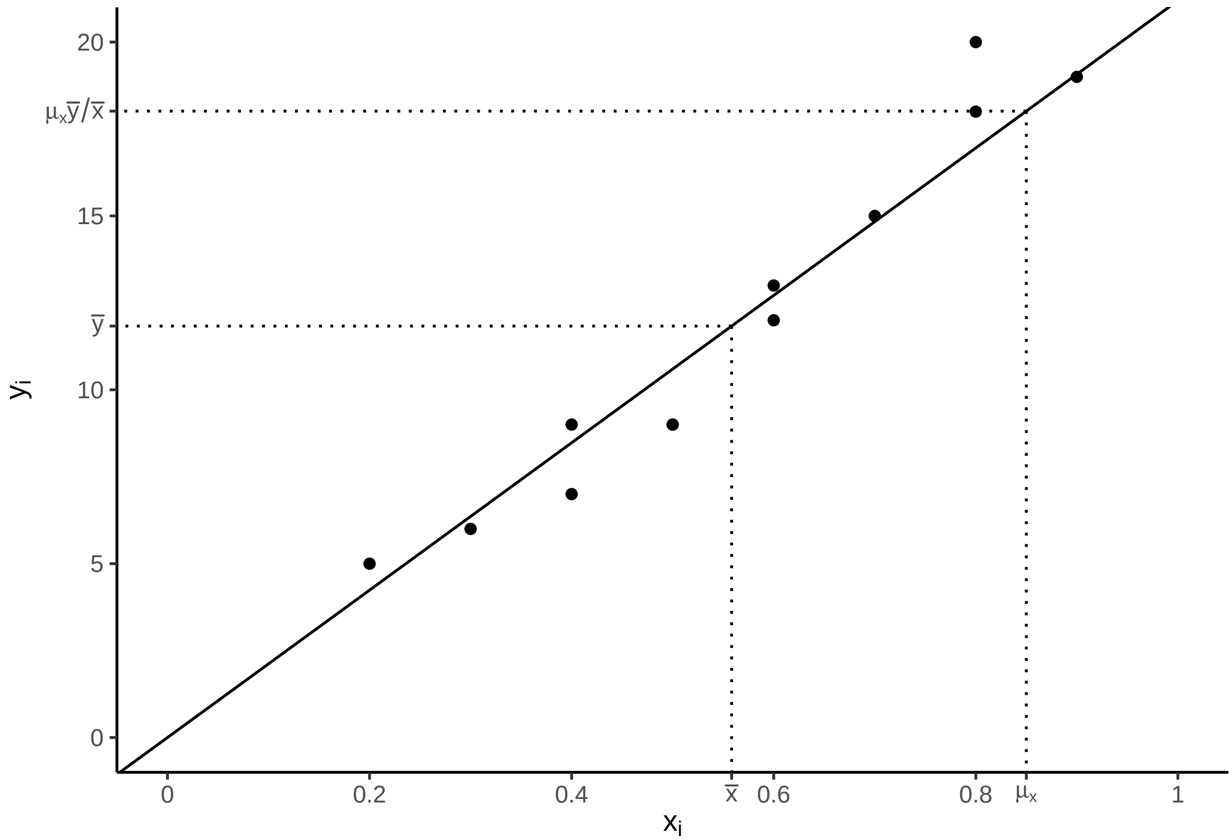
$$\hat{\mu}_y = \frac{\mu_x}{\bar{x}}\bar{y}$$

shows more clearly that the ratio estimator "adjusts" $\bar{y}$ by a factor of $\mu_x/\bar{x}$.

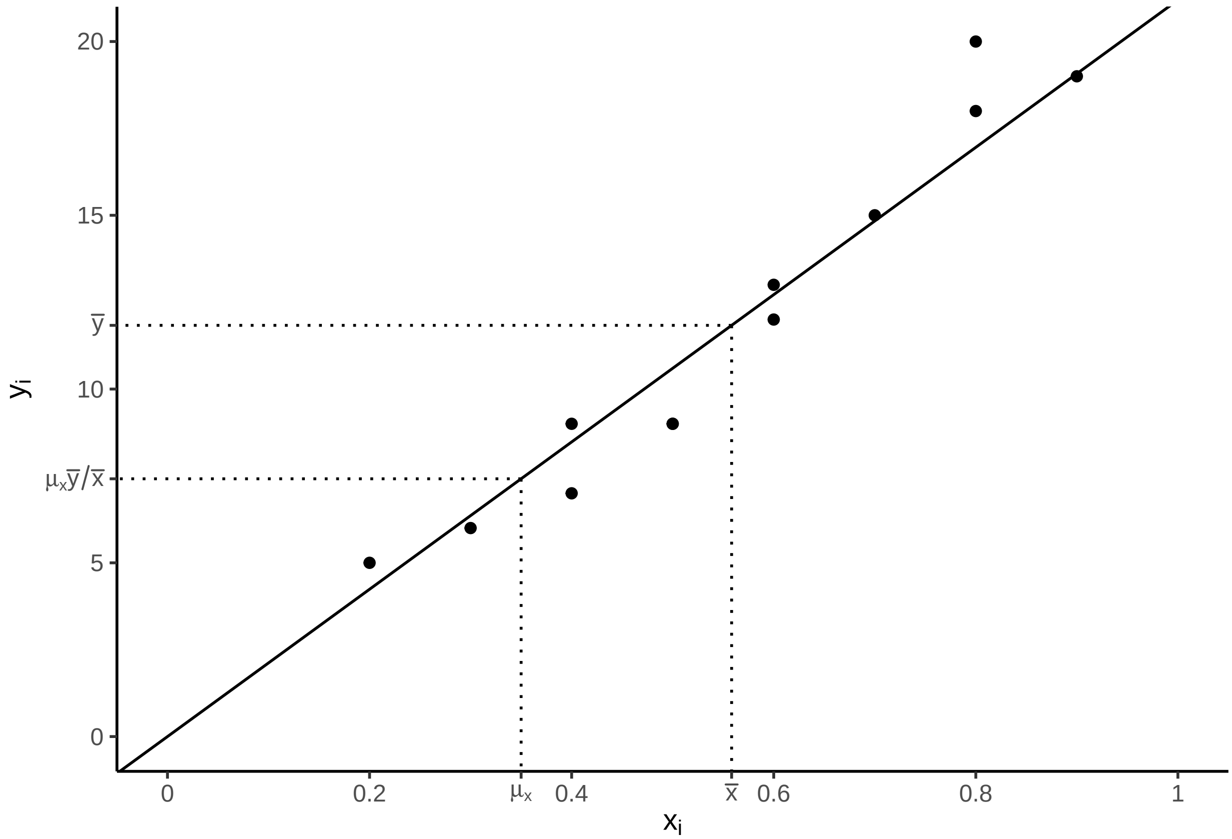
$$\begin{aligned}
\bar{x} < \mu_x &\Rightarrow \frac{\mu_x}{\bar{x}} > 1 \Rightarrow \frac{\mu_x}{\bar{x}}\bar{y} > \bar{y} \quad \text{(i.e., adjust estimate up)} \\
\bar{x} = \mu_x &\Rightarrow \frac{\mu_x}{\bar{x}} = 1 \Rightarrow \frac{\mu_x}{\bar{x}}\bar{y} = \bar{y} \quad \text{(i.e., no adjustment)} \\
\bar{x} > \mu_x &\Rightarrow \frac{\mu_x}{\bar{x}} < 1 \Rightarrow \frac{\mu_x}{\bar{x}}\bar{y} < \bar{y} \quad \text{(i.e., adjust estimate down)}
\end{aligned}$$

The factor of $\mu_x/\bar{x}$ tells us if μ_x is *underestimated* or *overestimated* by $\bar{x}$. This gives us some idea that *might* have underestimated or overestimated μ_y as well, so we might then adjust our estimate.

Example: Here μ_x is *underestimated* by $\bar{x}$.



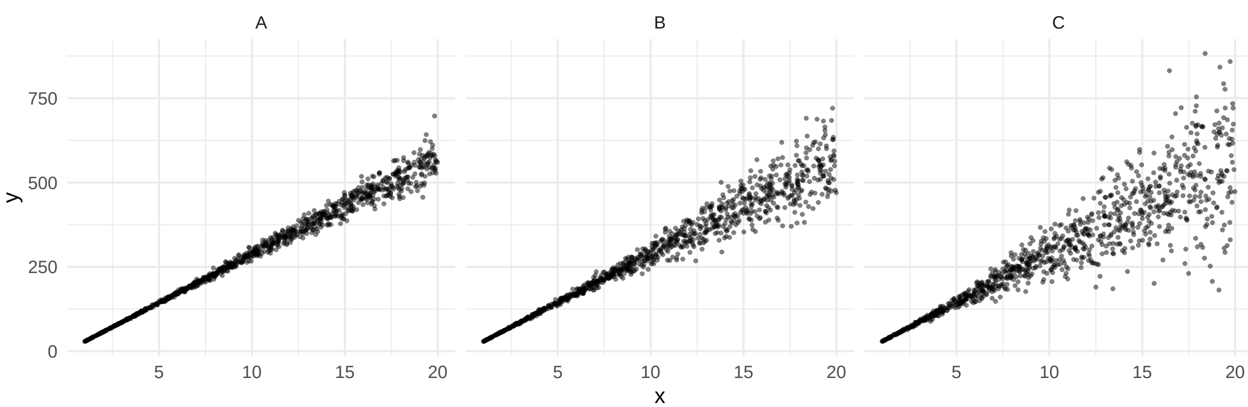
Example: Here μ_x is *overestimated* by $\bar{x}$.



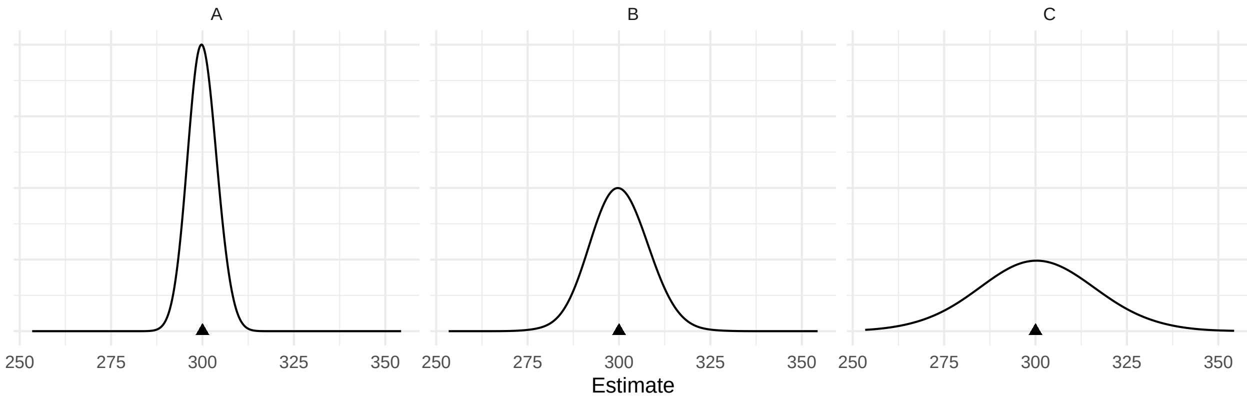
Performance of Ratio Estimators

How does the relationship between the target and auxiliary variable affect the ratio estimator?

Example: In each of the following populations $N = 1000$ and $\mu_y = 300$.



Consider the sampling distributions of the ratio estimator $\hat{\mu}_y = \mu_x \bar{y}/\bar{x}$ with $n = 25$.

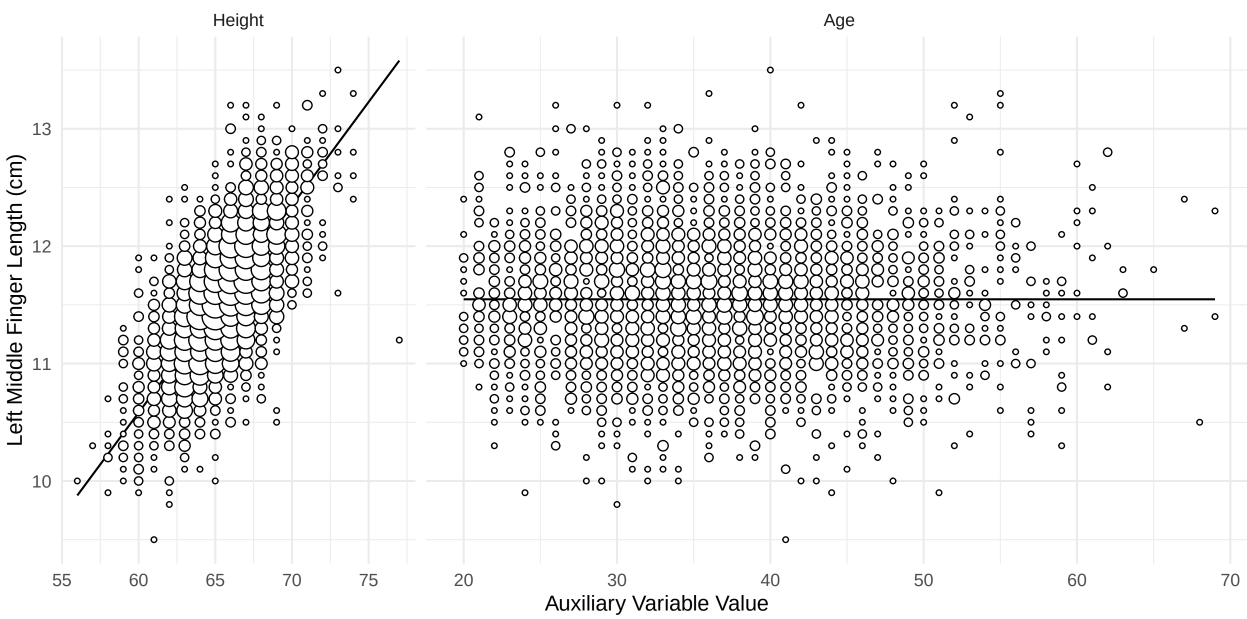


How does the relationship between the target and auxiliary variable affect the ratio estimator, and how does this compare to using the "non-ratio" estimator? Is a ratio estimator always better than a "non-ratio" estimator? Can a ratio estimator be *worse*?

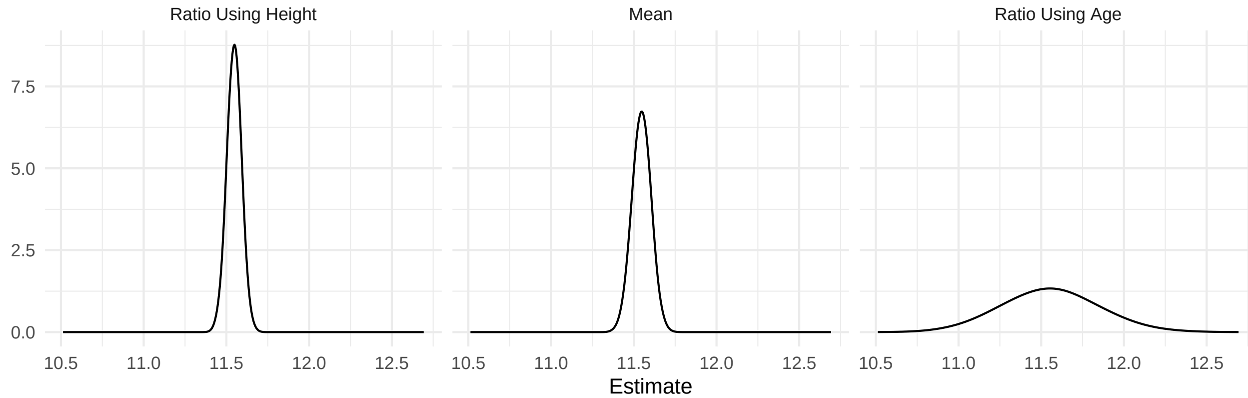
Example: Consider a population of $N = 3000$ elements (prisoners) where the target variable is finger length, and three estimators of μ_y:

1. $\hat{\mu}_y = \bar{y}$ (i.e., the sample mean)
2. $\hat{\mu}_y = \mu_h \bar{y}/\bar{h}$ (i.e., a ratio estimator using *height* as the auxiliary variable)
3. $\hat{\mu}_y = \mu_a \bar{y}/\bar{a}$ (i.e., a ratio estimator using *age* as the auxiliary variable)

The plots below show the distribution of finger length with height and with age in the *population*.



The plots below show the *sampling distributions* of the three estimators based on a simple random sampling design with $n = 25$.



estimator	variance	B
Ratio Using Height	0.00174	0.08
Mean	0.00292	0.11
Ratio Using Age	0.07795	0.56

Sources of Auxiliary Variables for Ratio Estimators

1. What is *necessary* for a variable to be used as an auxiliary variable?
2. What is *desirable* for a variable to be used as an auxiliary variable?

What are some sources of auxiliary variables?

1. *Rough approximations* to the target variable.
2. Some measure of sampling unit *size*.
3. Prior observations of the target variable from a *census*.