# The "*Annoyingly Intractable*" Ratio Estimator

Variance Approximation via Taylor Expansion

> "The distribution of the ratio estimate has proved annoyingly intractable because both y and x vary from sample to sample. The known theoretical results fall short of what we would like to know for practical applications."
>
> – William G. Cochran
>
> *Sampling Techniques*
>
> page 153

When sampling a population, we frequently encounter situations where auxiliary information can substantially improve the accuracy of our estimates. The ratio estimator is one of the oldest and most widely used techniques for exploiting such information. (Scheaffer et al., 2012).

**Definition.** Let $\boldsymbol{Y}$ be a study variable and $\boldsymbol{X}$ an auxiliary variable measured on the same units. The **ratio estimator** of the population mean $\boldsymbol{\mu_y}$ is defined as

$$\boldsymbol{r} = \frac{\bar{\boldsymbol{Y}}}{\bar{\boldsymbol{X}}} \cdot \boldsymbol{\mu_x} = \boldsymbol{R} \cdot \boldsymbol{\mu_x},$$

where $\boldsymbol{R} = \frac{\bar{\boldsymbol{Y}}}{\bar{\boldsymbol{X}}}$ is the sample ratio, $\bar{\boldsymbol{Y}}$ and $\bar{\boldsymbol{X}}$ are sample means, and $\boldsymbol{\mu_x}$ is the known population mean of $\boldsymbol{X}$.

The ratio estimator works best when the study variable $\boldsymbol{Y}$ correlates positively with the auxiliary variable $\boldsymbol{X}$, their relationship passes approximately through the origin, and the population total or mean of $\boldsymbol{X}$ is known from external sources like census data or administrative records. When these conditions hold, the estimator exploits the correlation structure to reduce sampling variability, often achieving substantial efficiency gains over the simple sample mean (Scheaffer et al., 2012).

However, all that glitters is so unfortunately not gold. The ratio estimator's fundamental challenge stems from its nonlinearity: $\boldsymbol{r} = \frac{\bar{\boldsymbol{Y}}}{\bar{\boldsymbol{X}}}$ is a nonlinear function of two random variables, creating two problems (Cochran, 1977). Bias appears because

$$\mathbb{E}[\boldsymbol{r}] \neq \boldsymbol{R} = \frac{\boldsymbol{\mu_y}}{\boldsymbol{\mu_x}} \text{ in finite samples,}$$

breaking the unbiasedness property of the sample mean. And the arguably larger problem that the variance formulas also fail to exist; the distribution of a ratio of random variables is generally intractable, depending on the full joint distribution of $\bar{Y}$ and $\bar{X}$. These issues have driven statisticians to approximation methods since the early 20th century, with the standard approach being Taylor series linearization of the ratio function followed by standard variance formulas (Cochran, 1977; Oehlert, 1992).

The key insight is that while the ratio is nonlinear globally, it behaves approximately linearly in a neighborhood of the population means $(\mu_y, \mu_x)$. For large samples, the sample means $(\bar{Y}, \bar{X})$ concentrate near $(\mu_y, \mu_x)$, so a local linear approximation becomes accurate (Casella and Berger, 2002).

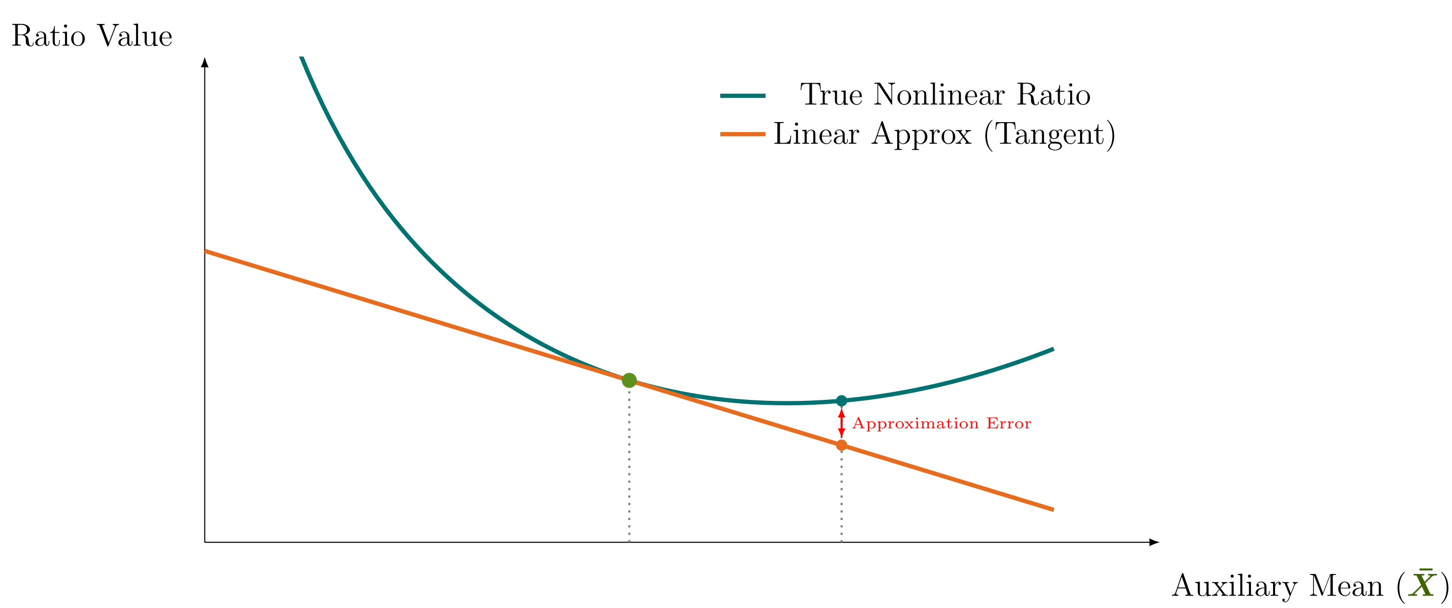
**Core Idea.** Expand the function $g(\bar{Y}, \bar{X}) = \frac{\bar{Y}}{\bar{X}}$ in a Taylor series around the point $(\mu_y, \mu_x)$. The first-order (linear) approximation converts the nonlinear problem into a linear one, for which variance calculations are straightforward.

Linearization is a general technique that works by using a linear function $L$ to approximate a more complex function, $r$, usually a ratio or non-linear statistic. For a linear function in the form

$$L = a + b\bar{Y} + c\bar{X}$$

its variance can be calculated exactly using the formula

$$\mathrm{Var}(L) = b^2\,\mathrm{Var}(\bar{Y}) + c^2\,\mathrm{Var}(\bar{X}) + 2bc\,\mathrm{Cov}(\bar{Y}, \bar{X}).$$




Taylor series expansion is the tool that we will use in order to construct this linear approximation. To linearize our ratio with a sufficient degree of accuracy requires only the first-order Taylor polynomial, but quantifying the bias of our estimator demands that we compute the second-order Taylor polynomial.

Our ratio can be expressed as a function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows:

$$\text{Let } g(\boldsymbol{y}, \boldsymbol{x}) = \frac{\boldsymbol{y}}{\boldsymbol{x}}$$

To derive the linear approximation of $g$ we begin by setting up the Taylor expansion. We expand $g$ around the population means $(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})$ using a first-order Taylor polynomial where subscripts denote partial derivatives.:

$$g(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) \approx g(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) + g_{\boldsymbol{y}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}) + g_{\boldsymbol{x}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}),$$

$$\text{For } g(\boldsymbol{y}, \boldsymbol{x}) = \frac{\boldsymbol{y}}{\boldsymbol{x}} \Rightarrow \begin{cases} g(\boldsymbol{y}, \boldsymbol{x}) = \dfrac{\boldsymbol{y}}{\boldsymbol{x}} & \Rightarrow \quad g(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) = \dfrac{\boldsymbol{\mu_y}}{\boldsymbol{\mu_x}} = \boldsymbol{R} \\ g_{\boldsymbol{y}}(\boldsymbol{y}, \boldsymbol{x}) = \dfrac{\partial}{\partial \boldsymbol{y}}\left(\dfrac{\boldsymbol{y}}{\boldsymbol{x}}\right) = \dfrac{1}{\boldsymbol{x}} & \Rightarrow \quad g_{\boldsymbol{y}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) = \dfrac{1}{\boldsymbol{\mu_x}} \\ g_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{x}) = \dfrac{\partial}{\partial \boldsymbol{x}}\left(\dfrac{\boldsymbol{y}}{\boldsymbol{x}}\right) = -\dfrac{\boldsymbol{y}}{\boldsymbol{x}^2} & \Rightarrow \quad g_{\boldsymbol{x}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) = -\dfrac{\boldsymbol{\mu_y}}{\boldsymbol{\mu_x}^2} = -\dfrac{\boldsymbol{R}}{\boldsymbol{\mu_x}} \end{cases}$$

Substituting these derivatives into the Taylor expansion yields the first-order approximation:

$$\boldsymbol{r} = \frac{\bar{\boldsymbol{Y}}}{\bar{\boldsymbol{X}}} \approx \boldsymbol{R} + \frac{1}{\boldsymbol{\mu_x}}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}) - \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})$$

This can be rewritten more compactly as:

$$\boldsymbol{r} \approx \boldsymbol{R} + \frac{1}{\boldsymbol{\mu_x}}\Big[(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}) - \boldsymbol{R}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})\Big].$$

The linearized form expresses the deviation of the sample ratio from the population ratio as a weighted combination of the deviations of the sample means from their population values.

Now that we have a linear estimator we can turn to calculating its variance:

$$\text{Var}(\boldsymbol{r}) \approx \text{Var}\left(\boldsymbol{R} + \frac{1}{\boldsymbol{\mu_x}}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}) - \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})\right).$$

We first distribute the terms to separate the random variables $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})$ from the population constants.

Since the variance of a constant is zero, the additive constant terms drop out:

$$\mathrm{Var}(\boldsymbol{r}) \approx \mathrm{Var}\left(\left[\boldsymbol{R} - \frac{\boldsymbol{\mu_y}}{\boldsymbol{\mu_x}} + \boldsymbol{R}\right] + \frac{1}{\boldsymbol{\mu_x}}\bar{\boldsymbol{Y}} - \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}\bar{\boldsymbol{X}}\right) = \mathrm{Var}\left(\frac{1}{\boldsymbol{\mu_x}}\bar{\boldsymbol{Y}} - \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}\bar{\boldsymbol{X}}\right).$$

Next, we apply the following property of the variance:

$$\mathrm{Var}(aY - bX) = a^2\,\mathrm{Var}(Y) + b^2\,\mathrm{Var}(X) - 2ab\,\mathrm{Cov}(Y, X)$$

$$\begin{aligned}\mathrm{Var}(\boldsymbol{r}) &\approx \mathrm{Var}\left(\frac{1}{\boldsymbol{\mu_x}}\bar{\boldsymbol{Y}}\right) + \mathrm{Var}\left(\frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}\bar{\boldsymbol{X}}\right) - 2\,\mathrm{Cov}\left(\frac{1}{\boldsymbol{\mu_x}}\bar{\boldsymbol{Y}}, \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}\bar{\boldsymbol{X}}\right) \\ &= \left(\frac{1}{\boldsymbol{\mu_x}}\right)^2 \mathrm{Var}(\bar{\boldsymbol{Y}}) + \left(\frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}\right)^2 \mathrm{Var}(\bar{\boldsymbol{X}}) - 2\left(\frac{1}{\boldsymbol{\mu_x}}\right)\left(\frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}\right)\mathrm{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}).\end{aligned}$$

Simplifying the coefficients yields the final approximation:

$$\mathrm{Var}(\boldsymbol{r}) \approx \frac{1}{\boldsymbol{\mu_x}^2}\,\mathrm{Var}(\bar{\boldsymbol{Y}}) + \frac{\boldsymbol{R}^2}{\boldsymbol{\mu_x}^2}\,\mathrm{Var}(\bar{\boldsymbol{X}}) - \frac{2\boldsymbol{R}}{\boldsymbol{\mu_x}^2}\,\mathrm{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}).$$

Factoring out $\frac{1}{\boldsymbol{\mu_x}^2}$, we obtain the celebrated 'delta-method' variance formula for the ratio estimator after linearizing it via Taylor expansion (Scheaffer et al., 2012):

$$\mathrm{Var}(\boldsymbol{r}) \approx \frac{1}{\boldsymbol{\mu_x}^2}\left[\mathrm{Var}(\bar{\boldsymbol{Y}}) - 2\boldsymbol{R}\,\mathrm{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) + \boldsymbol{R}^2\,\mathrm{Var}(\bar{\boldsymbol{X}})\right]$$

It should be noted that this variance formula can be expressed in several equivalent forms. Defining the "residual" $\boldsymbol{e_i} = \boldsymbol{y_i} - \boldsymbol{R}\boldsymbol{x_i}$, we can write:

$$\mathrm{Var}(\boldsymbol{r}) \approx \frac{1}{\boldsymbol{\mu_x}^2}\,\mathrm{Var}(\bar{e}), \quad \text{where} \quad \bar{e} = \bar{\boldsymbol{Y}} - \boldsymbol{R}\bar{\boldsymbol{X}} \text{ is the sample mean of the residuals.}$$

For simple random sampling without replacement (SRSWOR) from a finite population of size $N$:

$$\mathrm{Var}(\boldsymbol{r}) \approx \left(1 - \frac{n}{N}\right)\frac{\boldsymbol{S_r}^2}{n\boldsymbol{\mu_x}^2} \quad \text{where} \quad \boldsymbol{S_r}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{y_i} - \boldsymbol{R}\boldsymbol{x_i})^2$$

$\boldsymbol{S_r}^2$ is the population variance of the residuals (Scheaffer et al., 2012).

The first-order Taylor expansion shows that $\mathbb{E}[\boldsymbol{r}] \approx \boldsymbol{R}$, suggesting the ratio estimator is approximately unbiased. However, this approximation ignores higher-order terms that contribute to bias.

To quantify the bias, we extend to a second-order Taylor expansion (Cochran, 1977):

$$g(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) \approx g(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) + g_{\boldsymbol{y}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}) + g_{\boldsymbol{x}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}) + \frac{1}{2}\Big[g_{\boldsymbol{yy}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y})^2 + 2g_{\boldsymbol{yx}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y})(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}) + g_{\boldsymbol{xx}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})^2\Big].$$

Computing these partial derivatives and substituting them in yields:

$$g(\boldsymbol{y}, \boldsymbol{x}) = \frac{\boldsymbol{y}}{\boldsymbol{x}} \quad \begin{cases} g_{\boldsymbol{y}}(\boldsymbol{y}, \boldsymbol{x}) = \dfrac{1}{\boldsymbol{x}} \\ g_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{x}) = -\dfrac{\boldsymbol{y}}{\boldsymbol{x}^2} \end{cases} \begin{cases} g_{\boldsymbol{yy}}(\boldsymbol{y}, \boldsymbol{x}) = 0 & \Rightarrow \quad g_{\boldsymbol{yy}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) = 0 \\ g_{\boldsymbol{yx}}(\boldsymbol{y}, \boldsymbol{x}) = -\dfrac{1}{\boldsymbol{x}^2} & \Rightarrow \quad g_{\boldsymbol{yx}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) = -\dfrac{1}{\boldsymbol{\mu_x}^2} \\ g_{\boldsymbol{xx}}(\boldsymbol{y}, \boldsymbol{x}) = \dfrac{2\boldsymbol{y}}{\boldsymbol{x}^3} & \Rightarrow \quad g_{\boldsymbol{xx}}(\boldsymbol{\mu_y}, \boldsymbol{\mu_x}) = \dfrac{2\boldsymbol{\mu_y}}{\boldsymbol{\mu_x}^3} = \dfrac{2\boldsymbol{R}}{\boldsymbol{\mu_x}^2} \end{cases}$$

$$\begin{aligned} g(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) &\approx \boldsymbol{R} + \frac{1}{\boldsymbol{\mu_x}}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}) - \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}) + \frac{1}{2}\Big[0 - \frac{2}{\boldsymbol{\mu_x}^2}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y})(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}) + \frac{2\boldsymbol{R}}{\boldsymbol{\mu_x}^2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})^2\Big] \\ &= \boldsymbol{R} + \frac{\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}}{\boldsymbol{\mu_x}} - \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}) - \frac{1}{\boldsymbol{\mu_x}^2}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y})(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}) + \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}^2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})^2. \end{aligned}$$

We take the expectation of the expanded series. Since $\mathbb{E}[\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y}] = 0$ and $\mathbb{E}[\bar{\boldsymbol{X}} - \boldsymbol{\mu_x}] = 0$, the first-order linear terms vanish. We focus on the expectations of the second-order terms:

$$\mathbb{E}[\boldsymbol{r}] \approx \boldsymbol{R} + \frac{1}{2}\left[0 \cdot \mathrm{Var}(\bar{\boldsymbol{Y}}) + 2\left(-\frac{1}{\boldsymbol{\mu_x}^2}\right)\mathbb{E}\Big[(\bar{\boldsymbol{Y}} - \boldsymbol{\mu_y})(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})\Big] + \left(\frac{2\boldsymbol{R}}{\boldsymbol{\mu_x}^2}\right)\mathbb{E}\Big[(\bar{\boldsymbol{X}} - \boldsymbol{\mu_x})^2\Big]\right].$$

Substituting the definitions of covariance and variance:

$$\mathbb{E}[\boldsymbol{r}] \approx \boldsymbol{R} + \frac{1}{2}\left[-\frac{2}{\boldsymbol{\mu_x}^2}\mathrm{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) + \frac{2\boldsymbol{R}}{\boldsymbol{\mu_x}^2}\mathrm{Var}(\bar{\boldsymbol{X}})\right] = \boldsymbol{R} - \frac{1}{\boldsymbol{\mu_x}^2}\mathrm{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) + \frac{\boldsymbol{R}}{\boldsymbol{\mu_x}^2}\mathrm{Var}(\bar{\boldsymbol{X}}).$$

Factoring out the common term $1/\boldsymbol{\mu_x}^2$:

$$\mathbb{E}[\boldsymbol{r}] \approx \boldsymbol{R} + \frac{1}{\boldsymbol{\mu_x}^2}\Big[\boldsymbol{R}\,\mathrm{Var}(\bar{\boldsymbol{X}}) - \mathrm{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})\Big].$$

To find the bias of our estimator we just need to subtract the true population value of $\boldsymbol{R}$ from this expectation:

$$\text{Bias}(\boldsymbol{r}) = \mathbb{E}[\boldsymbol{r}] - \boldsymbol{R} \approx \frac{1}{\boldsymbol{\mu_x}^2}\Big[\boldsymbol{R}\,\text{Var}(\bar{\boldsymbol{X}}) - \text{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})\Big].$$

To understand the asymptotic behavior, recall the formulas for variance and covariance under SRSWOR. With population variance ${\boldsymbol{S_x}}^2$ and covariance $S_{\boldsymbol{xy}}$:

$$\text{Var}(\bar{\boldsymbol{X}}) = \left(1 - \frac{n}{N}\right)\frac{{\boldsymbol{S_x}}^2}{n} \quad \text{and} \quad \text{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) = \left(1 - \frac{n}{N}\right)\frac{S_{\boldsymbol{xy}}}{n}.$$

Substituting these into the bias approximation yields:

$$\text{Bias}(\boldsymbol{r}) \approx \frac{1}{\boldsymbol{\mu_x}^2}\left[\boldsymbol{R}\left(1 - \frac{n}{N}\right)\frac{{\boldsymbol{S_x}}^2}{n} - \left(1 - \frac{n}{N}\right)\frac{S_{\boldsymbol{xy}}}{n}\right] = \frac{1}{n}\underbrace{\left[\frac{1-f}{\boldsymbol{\mu_x}^2}(\boldsymbol{R}{\boldsymbol{S_x}}^2 - S_{\boldsymbol{xy}})\right]}_{\text{Population Constants}}.$$

Since the term in brackets is independent of the sample size $n$ (treating the finite population correction $f = n/N$ as negligible or bounded), we conclude:

$$\text{Bias}(\boldsymbol{r}) = O\left(\frac{1}{n}\right).$$

This demonstrates that the bias decreases linearly as the sample size increases, confirming that the ratio estimator is asymptotically unbiased (Cochran, 1977).

It is useful to express the bias relative to the true ratio $\boldsymbol{R} = \frac{\boldsymbol{\mu_y}}{\boldsymbol{\mu_x}}$. Dividing the bias approximation by $\boldsymbol{R}$:

$$\frac{\text{Bias}(\boldsymbol{r})}{\boldsymbol{R}} \approx \frac{1}{\boldsymbol{R}\boldsymbol{\mu_x}^2}\Big[\boldsymbol{R}\,\text{Var}(\bar{\boldsymbol{X}}) - \text{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})\Big] = \frac{\text{Var}(\bar{\boldsymbol{X}})}{\boldsymbol{\mu_x}^2} - \frac{\text{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})}{\boldsymbol{\mu_y}\boldsymbol{\mu_x}}.$$

For Simple Random Sampling without Replacement (SRSWOR), we substitute the known variances:

$$\text{Var}(\bar{\boldsymbol{X}}) = \frac{1-f}{n}{\boldsymbol{S_x}}^2 \quad \& \quad \text{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) = \frac{1-f}{n}\rho \boldsymbol{S_x}\boldsymbol{S_y}$$

$$\Rightarrow \frac{\text{Bias}(\boldsymbol{r})}{\boldsymbol{R}} \approx \left(\frac{1-f}{n}\right)\left[\frac{{\boldsymbol{S_x}}^2}{\boldsymbol{\mu_x}^2} - \frac{\rho \boldsymbol{S_x}\boldsymbol{S_y}}{\boldsymbol{\mu_y}\boldsymbol{\mu_x}}\right]$$

$$= \left(\frac{1-f}{n}\right)\left[\left(\frac{\boldsymbol{S_x}}{\boldsymbol{\mu_x}}\right)^2 - \rho\left(\frac{\boldsymbol{S_x}}{\boldsymbol{\mu_x}}\right)\left(\frac{\boldsymbol{S_y}}{\boldsymbol{\mu_y}}\right)\right].$$

Using the coefficients of variation $C_{\boldsymbol{x}} = \frac{\boldsymbol{S_x}}{\boldsymbol{\mu_x}}$ and $C_y = \frac{\boldsymbol{S_y}}{\boldsymbol{\mu_y}}$:

$$\frac{\text{Bias}(\boldsymbol{r})}{\boldsymbol{R}} \approx \left(1 - \frac{n}{N}\right)\frac{1}{n}\Big({C_{\boldsymbol{x}}}^2 - \rho C_{\boldsymbol{x}} C_{\boldsymbol{y}}\Big) = \left(1 - \frac{n}{N}\right)\frac{C_{\boldsymbol{x}}(C_{\boldsymbol{x}} - \rho C_{\boldsymbol{y}})}{n}.$$

The bias is typically positive (overestimation) when $\rho C_{\boldsymbol{y}} < C_{\boldsymbol{x}}$, and is proportional to ${C_{\boldsymbol{x}}}^2$. High positive correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ reduces the bias (Cochran, 1977).

Linear approximation via Taylor Expansion method provides a convenient, closed-form approximation for the variance of a ratio estimator, but it rests on asymptotic theory and first-order linearization. In finite samples, the accuracy of this approximation depends on several factors: the degree of curvature in the ratio functional $g(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})$, the proximity of $\bar{\boldsymbol{X}}$ to zero (which affects the stability of the ratio), and the concentration of the sampling distribution of $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})$ around the population means $(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})$.

To build confidence in the derived formula and to understand its behavior in realistic settings, we empirically assess its accuracy by comparing it against two computational approaches that make fewer analytical approximations:

1. **Method A (Simulation):** We treat the predicted variance produced by the Taylor Expansion as a hypothesis to be tested against a Monte Carlo "ground truth." By generating a synthetic finite population from a known joint model for $(\boldsymbol{Y}, \boldsymbol{X})$ (e.g., bivariate normal with specified means, variances, and correlation), we can repeatedly draw probability samples (SRSWOR), compute the ratio estimator in each replication, and estimate its empirical variance across replications. This method directly approximates the true sampling variance of $\boldsymbol{r}$ under the specified design, free from linearization error. The discrepancy between the simulation variance and the approximated variance derived above reveals how well the first-order approximation captures the actual variability in $\boldsymbol{r}$.

2. **Method B (Bootstrap):** In practice, we rarely have access to the full finite population; we observe only a single sample. The bootstrap offers a design-based resampling strategy that mimics the sampling process: from one observed SRSWOR sample, we generate many bootstrap resamples (with replacement) to approximate the sampling distribution of $\boldsymbol{r}$. Comparing the bootstrap variance to the Taylor Series prediction tells us whether a practitioner relying on resampling would obtain a similar uncertainty estimate to one using analytic linearization. This is particularly relevant when the ratio estimator is embedded in a complex survey design where closed-form variance formulas are unavailable.

This exercise follows well-established practice in survey sampling and resampling methods; see, for example, Cochran (1977, pp. 162–164) for empirical validations of the approximate variance, and Efron and Tibshirani (1993, pp. Ch.20) for rigorous comparison of the Taylor series linearization (delta-method) and bootstrapping.

We generate a synthetic finite population of size $N = 100{,}000$ with a bivariate normal structure:

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \boldsymbol{\mu_y} \\ \boldsymbol{\mu_x} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\sigma_y}^2 & \rho\,\boldsymbol{\sigma_y}\boldsymbol{\sigma_x} \\ \rho\,\boldsymbol{\sigma_y}\boldsymbol{\sigma_x} & \boldsymbol{\sigma_x}^2 \end{bmatrix} \right).$$

We fix the target means $\boldsymbol{\mu_y} = 15$ and $\boldsymbol{\mu_x} = 10$, standard deviations $\boldsymbol{\sigma_y} = 6$, $\boldsymbol{\sigma_x} = 4$, and correlation $\rho = 0.8$. The population ratio is $\boldsymbol{R} = \dfrac{\boldsymbol{\mu_y}}{\boldsymbol{\mu_x}} = 1.5$.

The choice of $N = 100{,}000$ ensures that the finite population is large enough to approximate the superpopulation model while keeping sampling fractions small enough that finite-population corrections are meaningful. The correlation $\rho = 0.8$ creates a strong linear relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, which is typical in ratio estimation contexts (e.g., where $\boldsymbol{X}$ is an auxiliary variable that explains much of the variation in $\boldsymbol{Y}$). The means and standard deviations are chosen so that $\boldsymbol{\mu_x}$ is comfortably away from zero, avoiding the instability that arises when the denominator can approach zero.

For sampling, we draw SRSWOR samples of size $n = 200$, with $K = 2000$ Monte Carlo replications in Method A. This sample size is modest but realistic for many survey applications, and it is large enough for the central limit theorem to begin taking effect for $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})$. For Method B, we draw a single SRSWOR sample of size $n$ and apply $B = 3000$ bootstrap resamples. The number of bootstrap replicates is chosen to yield a stable variance estimate; $B \geq 2000$ is generally recommended for variance estimation.

In each replication (or bootstrap resample), we compute the ratio estimate

$$\boldsymbol{r} = \frac{\bar{\boldsymbol{Y}}}{\bar{\boldsymbol{X}}}.$$

For the Taylor series prediction, we use the linearization result:

$$\mathrm{Var}(\boldsymbol{r}) \approx \frac{1}{\boldsymbol{\mu_x}^2}\Big[\mathrm{Var}(\bar{\boldsymbol{Y}}) \;-\; 2\,\boldsymbol{R}\,\mathrm{Cov}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}) \;+\; \boldsymbol{R}^2\,\mathrm{Var}(\bar{\boldsymbol{X}})\Big],$$

or, under SRSWOR from a finite population,

$$\mathrm{Var}(\boldsymbol{r}) \approx \Big(1 - \frac{n}{N}\Big)\frac{\boldsymbol{S_r}^2}{n\,\boldsymbol{\mu_x}^2}, \quad S_r^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N}\big(\boldsymbol{y_i} - \boldsymbol{R}\boldsymbol{x_i}\big)^2,$$

see Cochran (1977); Oehlert (1992). In practice, we plug in the finite-population quantities computed from the generated population for $\boldsymbol{\mu_x}$, $\boldsymbol{R}$, and $\boldsymbol{S_r}^2$.

The first expression reveals the underlying logic of the linear approximation: the variance of $\boldsymbol{r}$ is driven by the variability of the numerator $\bar{\boldsymbol{Y}}$ and denominator $\bar{\boldsymbol{X}}$, but it also depends on their covariance. The term $\boldsymbol{R}^2\,\mathrm{Var}(\bar{\boldsymbol{X}})$ captures the contribution of denominator uncertainty, while $-2\boldsymbol{R}\,\mathrm{Cov}(\bar{\boldsymbol{Y}},\bar{\boldsymbol{X}})$ reflects the fact that when $\boldsymbol{Y}$ and $\boldsymbol{X}$ are positively correlated, fluctuations in the denominator partially offset fluctuations in the numerator, reducing overall variability. The second expression is more practical for computation: ${S_r}^2$ is the population variance of the residuals $\boldsymbol{r}_i = \boldsymbol{y_i} - \boldsymbol{R}\boldsymbol{x_i}$, which represent the deviation of each $\boldsymbol{y_i}$ from the proportional relationship implied by the ratio $\boldsymbol{R}$. The finite-population correction $(1-n/N)$ accounts for the fact that sampling without replacement from a finite population reduces variance relative to sampling with replacement.

**Method A (Simulation):** Across $R = 2000$ replications, the empirical variance of $\boldsymbol{r}$ is

$$\widehat{\mathrm{Var}}_{\mathrm{Sim}}(\boldsymbol{r}) \;=\; 0.0007193404.$$

The corresponding Taylor Series prediction, evaluated using the population quantities, is

$$\widehat{\mathrm{Var}}_{\Delta}(\boldsymbol{r}) \;=\; 0.0007160586.$$

The relative error is

$$\frac{\widehat{\mathrm{Var}}_{\mathrm{Sim}}(\boldsymbol{r}) - \widehat{\mathrm{Var}}_{\Delta}(\boldsymbol{r})}{\widehat{\mathrm{Var}}_{\Delta}(\boldsymbol{r})} \;=\; 0.004583.$$

The simulation variance is slightly higher than the Taylor Series prediction (by about 0.46%). This modest upward bias is consistent with the fact that the approximation first-order Taylor polynomial: it ignores higher-order terms in the Taylor expansion of $g(\bar{\boldsymbol{Y}},\bar{\boldsymbol{X}})$ that can contribute additional variability, especially in finite samples. The small magnitude of the discrepancy suggests that, for this configuration, the linearization provides an excellent approximation to the true sampling variance.

**Method B (Bootstrap):** From a single SRSWOR sample and $B = 3000$ bootstrap resamples, the bootstrap variance estimate is

$$\widehat{\mathrm{Var}}_{\mathrm{Boot}}(\boldsymbol{r}) \;=\; 0.0006894320.$$

Its relative difference from the Taylor series prediction is

$$\frac{\widehat{\mathrm{Var}}_{\mathrm{Boot}}(\boldsymbol{r}) - \widehat{\mathrm{Var}}_{\Delta}(\boldsymbol{r})}{\widehat{\mathrm{Var}}_{\Delta}(\boldsymbol{r})} \;=\; -0.037185.$$

The bootstrap variance is about 3.7% lower than the Taylor series prediction. This downward bias is typical of the nonparametric bootstrap for ratio estimators: bootstrap resampling introduces additional discreteness and can underestimate the true sampling variance when the estimator is nonlinear. The bootstrap distribution is centered at the sample ratio $\boldsymbol{r}$ rather than the population ratio $\boldsymbol{R}$, and the resampling process does not fully capture the variability of $\bar{\boldsymbol{X}}$ around $\mu_x$. Nevertheless, the bootstrap estimate remains in the same ballpark as the analytic approximation, which is reassuring for practical applications.

The table summarizes the numerical comparison. The simulation variance aligns almost perfectly with the Taylor Series prediction, confirming that the linearization captures the bulk of the sampling variability. The bootstrap variance, while slightly lower, is still reasonably close, suggesting that resampling-based inference can serve as a viable alternative when analytic formulas are unavailable or when the sampling design is more complex.

| | Taylor Series | Simulation | Bootstrap |
|---|---|---|---|
| Variance estimate | 0.0007160586 | 0.0007193404 | 0.0006894320 |
| Relative error vs Taylor Series | – | 0.004583 | -0.037185 |

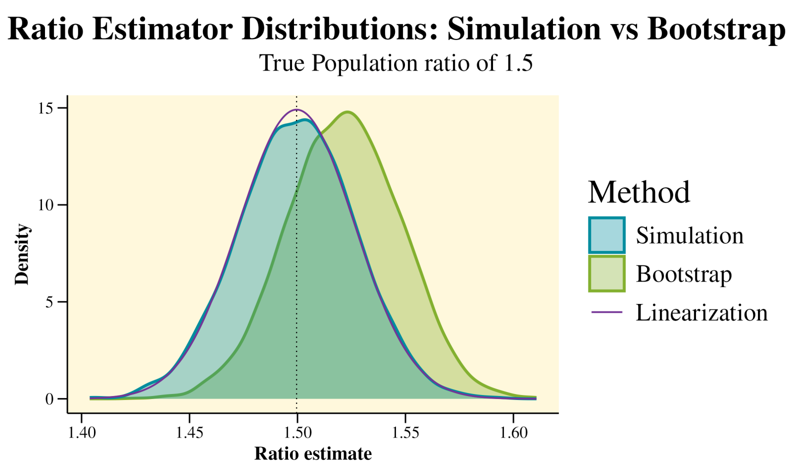



Figure 2: Comparison of variance estimates for the ratio estimator: Taylor series prediction versus simulation and bootstrap results. Dashed line indicates $\boldsymbol{R}$.

The density plot (Figure 2) visualizes the sampling distribution of the ratio estimator $\boldsymbol{r}$ across simulation replications and bootstrap resamples. Both distributions are approximately centered near the true ratio $\boldsymbol{R} = 1.5$ (dashed line) and exhibit similar spread, though the bootstrap distribution is slightly more concentrated, consistent with its lower variance estimate. The near-normality of both distributions validates the asymptotic normal approximation that underlies the Taylor expansion.

In this configuration, the simulation-based variance is *above* the linearized prediction by 0.46 %, while the bootstrap variance is *below* the linearized prediction by 3.72 %. Discrepancies can arise from finite-sample effects, the curvature of the ratio functional, and the extent to which $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}})$

concentrate near $(\boldsymbol{\mu_y}, \boldsymbol{\mu_x})$. As expected from linearization theory (Oehlert, 1992; Cochran, 1977; Scheaffer et al., 2012), agreement tends to improve with larger $n$, stronger concentration of $\bar{\boldsymbol{X}}$ around $\boldsymbol{\mu_x}$, and when the linear model through the origin is an adequate approximation.

The results illustrate a key trade-off: the Taylor expansion is computationally trivial once derived, but it relies on assumptions that may not hold in small samples or with highly nonlinear estimators. The simulation approach provides a gold standard but requires knowledge of the superpopulation model and is computationally intensive. The bootstrap offers a middle ground: it is design-based, requires no model assumptions, and can be applied to complex estimators, but it may exhibit modest bias in finite samples. The takeaway is that the linearized variance formula is highly reliable for ratio estimation under SRSWOR when the sample size is moderate and the denominator variable is well-behaved. When in doubt, or when dealing with more complex designs, bootstrap validation provides a useful robustness check.

# References

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2nd edition.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.

Scheaffer, R. L., Mendenhall, W., Ott, R. L., and Gerow, K. (2012). *Elementary survey sampling*. Cengage Learning, 7th edition.