www.ijonest.net

Enhancing Students' Understanding of Variance Estimation through the Lens of the Bias-Variance Trade-off: From Sample Variance to Improved Insight

Yeil Kwon
Wichita State University, United States

To cite this article:

Kwon, Y. (2025). Enhancing students' understanding of variance estimation through the lens of the bias-variance trade-off: From sample variance to improved insight. *International Journal on Engineering, Science, and Technology (IJonEST), 7*(2), 160-192. https://doi.org/10.46328/ijonest.6142

International Journal on Engineering, Science, and Technology

2025, Vol. 7, No. 2, 160-192 https://doi.org/10.46328/ijonest.6142

Enhancing Students' Understanding of Variance Estimation through the Lens of the Bias-Variance Trade-off: From Sample Variance to Improved Insight

Yeil Kwon

Article Info

Article History

Received:
30 September 2025
Accepted:
24 December 2025

Keywords

Sample variance
Unbiasedness
Mean squared error
Bias-Variance trade-off
ChatGPT

Abstract

Understanding variance estimation is a cornerstone of statistical education. While the unbiasedness of the sample variance is a valuable property, it should not be the sole criterion for selecting an estimator. This paper advocates for incorporating mean squared error (MSE) considerations into the teaching of variance estimation in statistics classrooms. In contemporary applications, estimators with lower MSE are often preferred, even when they are biased. In this study, we first examine the relationship between the minimum-MSE variance estimators—among those based on the sum of squared deviations—and the kurtosis of the underlying population distribution. Furthermore, we demonstrate that, particularly for skewed distributions, alternative estimators can substantially outperform the sample variance in terms of MSE. By using variance estimation as a framework, instructors can effectively introduce students to the bias-variance trade-off, a foundational concept in statistical estimation and model selection. To support classroom implementation, we provide a series of R codes for the simulation-based visualizations that foster students' intuition about the interaction between bias and variance.

Introduction

The sample variance, alongside the sample mean, is among the most fundamental statistics introduced in both introductory and intermediate-level statistics courses. Given its foundational role in data analysis, the computation of sample variance is typically introduced early in the curriculum. The definitions of the population variance σ^2 and sample variance s^2 are given as follows:

$$\sigma^2 = \frac{1}{N}\sum\left(x_j - \mu\right)^2 \quad and \quad s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

where μ represents the population mean, and N and n denote the population and sample sizes, respectively. With formula in (1), typically, the following rules are given to students:

To calculate variance from a population, we divide the sum of squared deviations by N, but when calculating variance from a sample, we have to divide it by $n-1$ rather than n.

The shift in the denominator of the sample variance often perplexes students. While the population mean and variance use N, and the sample mean uses n, the sample variance uniquely involves dividing by $n - 1$. Understandably, students are left questioning this inconsistency. Faced with this conceptual tension, students—like Hamlet in Shakespeare's tragedy - might exclaim:

"$n - 1$ or n, that is the question!"

This question is not only natural but also pedagogically important. From a student's perspective, dividing by $n-1$ rather than n may appear counterintuitive since there are n squared deviations. Examples of conventional instructions commonly given to students to address this confusion include the following:

> Case 1: We divide by $n - 1$, rather than dividing by n, when computing a sample's variance; there's some mathematical nuance here, but the end result is that doing this makes this statistic slightly more reliable and useful (Diez et al. 2019, p. 47).

In general, the term "reliable" is used to describe an estimator that does not fluctuate significantly depending on the sample when estimating a population parameter. In many cases, using n can actually result in more stable (i.e., less variable) estimates than using $n - 1$. The subsequent sections will delve into this issue.

> Case 2: The sample variance underestimates the population variance when the denominator in the sample formula for variance is n. However, the sample variance does not underestimate the population variance if the denominator in the sample formula for variance is $n - 1$ (Mann 2010, p. 94).

This instruction suggests that using $n - 1$is more desirable than using n, as the latter tends to underestimate the population variance. However, underestimation does not necessarily imply lower accuracy. In many cases, an underestimated estimator may actually be closer to the true population variance than one that is not. This point will be examined in more detail in the upcoming sections, too.

The confusion caused by the denominator of the sample variance often arises from the ambiguous phrasings between "calculating variance from a sample" and "finding a *sample variance*". Thus, before addressing the question directly, it is useful to clarify what is truly meant by "calculating variance from a sample." This clarification is essential because the conventional instruction can misleadingly imply that any method not using $n - 1$ is incorrect. In reality, the so-called *sample variance* is merely one of several valid estimators of the population variance. Some educators and students mistakenly regard the use of n instead of $n-1$ as a computational error rather than a methodological alternative. However, using n leads not to a "wrong" estimator, but rather to a "different one," each with its own statistical desirable properties.

This article aims to investigate the enduring presence of the "Hamlet's question" in statistics education and to identify the pedagogical misconceptions that sustain it. These misconceptions are closely tied to the criteria used for evaluating estimators. In particular, we argue that undergraduate statistics courses should go beyond the conventional emphasis on unbiasedness and incorporate the concept of *Mean Squared Error* (MSE) as a more comprehensive metric for assessing estimator performance (Levy 2006).

To illustrate this point, we investigate the form of variance estimators that minimize the mean squared error when data are drawn from symmetric distributions. In particular, we examine estimators based on the sum of squared deviations with a denominator of $n + r$, where $r \geq -1$, highlighting that the issue extends beyond the common debate of using $n - 1$ versus n (Rosenthal 2015). We then extend the discussion to asymmetric distributions, proposing alternative variance estimators that deviate substantially from those optimal under symmetry, and compare their MSEs by analyzing both bias and variance components. Ultimately, integrating these discussions into instruction provides students with a broader understanding of estimation and presents a valuable opportunity to introduce the concept of the *bias-variance trade-off*, a fundamental principle in parameter estimation and model selection within machine learning (Murphy, 2012).

Sample Variance and Its Denominator

When we compute a variance from sample data, we are not calculating the population variance itself but instead estimating it. Crucially, estimation is not a uniquely defined procedure. While the population variance is a fixed, unique, and unknown quantity, there are multiple valid ways to estimate it based on observed data. The sample variance is merely one among many possible estimators.

Analysts select an estimator based on its statistical properties and the goals of their analysis. No single estimator is universally best in all scenarios; each has its advantages and disadvantages. Although the true population parameter is fixed, the method of estimation depends on context, assumptions, and desired properties such as unbiasedness or efficiency (Casella & Berger 2002). The applicability of this idea extends beyond variance estimation and can be equally relevant in the estimation of other parameters. For example, although the sample mean is often used to estimate the population mean, it is not always the optimal choice. When sampling from a skewed distribution such as the lognormal, the sample mean can be highly variable, even with large sample sizes. In such cases, alternative estimators may offer more stable performance (Shen et al. 2006, Longford 2009).

Understanding that estimation is not a one-size-fits-all process makes it easier for students to appreciate why the denominator in the sample variance formula is $n - 1$. This choice is based on one of the desirable statistical properties: *unbiasedness*. It is equally important to address the source of confusion embedded in the terminology "sample variance." While "sample variance" is commonly defined using $n - 1$ in the denominator, it is only one possible estimator of the population variance. That is, it is "correct" to divide by $n - 1$ when one is explicitly computing the *sample variance* as defined, and this will guarantee the unbiasedness of the sample variance. However, this does not imply that dividing by n is incorrect. In fact, using n instead yields another legitimate estimator known as the method of moments estimator (MME), which may be preferable under specific evaluation criteria.

Mean Squared Error and the Bias-Variance Trade-off

The sample variance (using $n-1$ for the denominator) possesses the desirable property of unbiasedness. possesses the desirable property of unbiasedness. At this point, it is vital to investigate the key differences between variance estimators that use $n - 1$ and those that use n as the denominator. The central issue lies in understanding the trade-off between bias and variance, and more broadly, in evaluating estimators through the lens of mean squared error

(MSE). The MSE of an estimator is defined as the expected squared deviation from the true parameter value:

$$\mathrm{MSE}(\theta) = \mathrm{E}\left[\left(\theta - \hat{\theta}\right)^2\right] = \mathrm{Bias}^2\left(\hat{\theta}\right) + \mathrm{Var}\left(\hat{\theta}\right),$$

where θ is the true parameter and $\hat{\theta}$ is an estimator for θ. The sample variance using $n-1$ is an unbiased estimator of the population variance regardless of the population distribution. As a result, its MSE is equal to its variance. This property makes it particularly appealing in classical statistical inference, as unbiasedness has long been a dominant criterion (Mood et al. 1974). However, in contemporary statistical and machine learning practices, MSE has become a more influential criterion. Slight bias may be acceptable—or even preferred—if it significantly reduces the variance of an estimator, thereby lowering its overall MSE (Kubokawa & Srivastava 2002, Hara 2007). This shift in preference is particularly evident in the machine learning context, where algorithms are often chosen based on their predictive performance, not necessarily on unbiasedness (Hastie et al. 2009).

To return to our example, when data are sampled from a normal distribution, the variance estimator obtained by dividing by n is biased, but it has a smaller MSE compared to the unbiased sample variance. Under an MSE-based evaluation, this biased estimator may thus be more desirable. Rather than teaching students to favor the sample variance solely due to its unbiasedness, educators should encourage a broader perspective—one that incorporates multiple criteria for evaluating estimators, especially MSE. By engaging students in MSE-based comparisons, instructors can deepen students' understanding of the different goals of estimation and foster more critical thinking. Furthermore, this pedagogical approach provides a natural transition to the *bias-variance trade-off*, a central concept in both advanced statistical theory and data science.

The *bias-variance trade-off* plays a foundational role in parameter estimation and model selection. For instance, ordinary least squares (OLS) estimators in regression models are unbiased and possess the minimum variance among all linear unbiased estimators, irrespective of the underlying distribution of the error term (Gauss-Markov theorem, Stigler 1981). However, ridge regression introduces bias in order to reduce variance, often resulting in a lower mean squared error (Lakshmi & Sajesh 2025). This trade-off is also critical in model selection within machine learning, where increasing model complexity typically decreases bias but increases variance. Effective model selection hinges on striking an optimal balance to minimize MSE. Introducing this framework at the intermediate level equips students to connect foundational statistical concepts with modern data science practices.

MSE of the Sample Variance and $S^2_{d(n)}$

Suppose we have a random sample $X_1, \ldots, X_n$, from a population with mean μ and variance σ^2. Define $S^2_{d(n)}$ as

$$S^2_{d(n)} = \frac{1}{d(n)} \sum\left(X_i - \overline{X}\right)^2$$

where $d(n)$ is any function of n. The sample variance can be regarded as a special case of $S^2_{d(n)}$ with $d(n) = n - 1$. As stated earlier, the sample variance S^2_{n-1} is an unbiased estimator for σ^2. If the sample is obtained from a normal population, S^2_n is biased, but it is both a method of moments estimator (MME) and a maximum likelihood estimator (MLE) for σ^2. Furthermore, S^2_{n+1} is the estimator for σ^2 that has the smallest MSE for the normal sample (Lehmann 1983).

One of the principal advantages of the sample variance is that it serves as an unbiased estimator of the population variance, irrespective of the underlying population distribution, satisfying $E(S^2_{n-1}) = \sigma^2$. Furthermore, when the underlying distribution is normal, the sample variance is the uniformly minimum variance unbiased estimator (UMVUE) for σ^2 (Casella & Berger 2002), implying that it achieves the lowest variance among all unbiased estimators. Due to these favorable properties, the sample variance is widely employed across numerous fields as the standard method for variance estimation. Nonetheless, from an MSE perspective, we can identify many biased estimators that outperform the sample variance. Note that since the sample variance is unbiased, its MSE coincides with its variance, that is, $\text{MSE}(S^2_{n-1}) = \text{Var}(S^2_{n-1})$. The variance of S^2_{n-1} is given by,

$$\text{Var}(S^2_{n-1}) = \frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)$$

where $\kappa = E(X-\mu)^4/\sigma^4$ represents the kurtosis (Cho & Cho 2009). Kurtosis, κ, is a measure that describes the shape of a probability distribution's tails in relation to its overall shape. It helps assess whether the data are heavy-tailed or light-tailed, which impacts the likelihood of outliers (Balanda & MacGillivray, 1988). As the heaviness of the tails increases, the value of κ also increases. For example, the normal distribution, which has the light tails, has $\kappa = 3$, whereas the Laplace distribution, with heavier tails that decay more slowly, has $\kappa = 6$. Using (2), we can find the variance and the squared bias of $S^2_{d(n)}$ as

$$\text{Var}\left(S^2_{d(n)}\right) = \text{Var}\left(\frac{n-1}{d(n)}S^2_{n-1}\right) = \frac{(n-1)^2}{d^2(n)}\cdot\frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)$$

$$\text{Bias}^2\left(S^2_{d(n)}\right) = \text{E}^2\left(\frac{n-1}{d(n)}S^2_{n-1} - \sigma^2\right) = \sigma^4\left(\frac{n-1}{d(n)} - 1\right)^2$$

Therefore, the MSE of $S^2{}_{d(n)}$ is obtained by adding (3) and (4),

$$\text{MSE}\left(S^2_{d(n)}\right) = \sigma^4\left(\frac{(n-1)^2(\kappa-3) + n(n+1)(n-1)}{n\cdot d^2(n)} - \frac{2(n-1)}{d(n)} + 1\right)$$

and it is minimized at

$$d(n) = n + \kappa - 2 + \frac{3-\kappa}{n}$$

From (6), we can see that the optimal $d(n)$ is a linear function of κ with a slope close to 1. Since the theoretical minimum value of κ is 1, it follows that $d(n) > n-1$. Therefore, unless the underlying distribution has a κ value close to 1, the sample variance with denominator $n-1$ rarely achieves the minimum mean squared error.

MSE of $S^2_{d(n)}$ for Symmetric Distributions

For a normal distribution with kurtosis $\kappa = 3$, the expression in Equation (6) indicates that the mean squared error (MSE) of the estimator $S^2_{d(n)}$ is minimized when $d(n) = n + 1$. In other words, S^2_{n+1} outperforms the unbiased sample variance S^2_{n-1} in terms of MSE when estimating σ^2 from normally distributed data. Figure 1 illustrates the squared bias, variance, and MSE of $S^2_{d(n)}$ for $d(n) = n-1$, n, and $n+1$ under the normal distribution. While $S^2{}_{n+1}$ exhibits the largest squared bias (top-left panel), the corresponding reduction in variance (top-right panel) significantly outweighs the increase in bias, resulting in the smallest overall MSE (bottom-left panel). The bottom-right panel presents the MSE ratio of S^2_n and S^2_{n+1} relative to the sample variance S^2_{n-1}. Notably, both estimators uniformly outperform the traditional unbiased sample variance across various sample sizes.

These results provide practical guidance for selecting an estimator for σ^2 when working with data from a normal distribution. If unbiasedness is the primary concern, S_{n-1}^2 remains the appropriate choice. If one seeks an estimator based on the method of moments or the maximum likelihood principle, S_n^2 is preferred. However, if minimizing MSE is the objective, then S_{n+1}^2 is the optimal estimator.

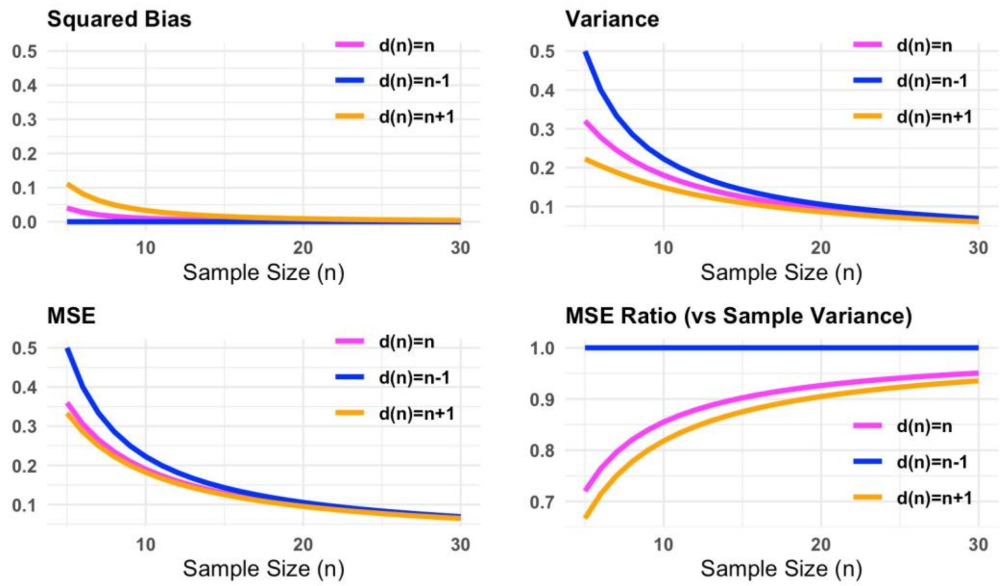


Figure 1. Squared Bias, Variance, and MSE of S_{n-1}^2, S_n^2, and S_{n+1}^2 for Normal Samples with $\sigma^2 = 1$.

Relation between Kurtosis and Optimal $d(n)$

Because the MSE of $S_{d(n)}^2$ depends on the kurtosis κ of the underlying distribution, the optimal value of $d(n)$ that minimizes MSE is also a function of κ. As shown in Equation (6), $d(n)$ increases linearly with κ with a slope of approximately $(n-1)/n \approx 1$ for moderate sample sizes. Consequently, as κ increases by one unit, the optimal $d(n)$ increases by approximately one unit as well.

Figure 2 depicts the relationship between κ and the optimal $d(n)$ for several symmetric continuous distributions that are commonly included in statistics curricula (In the graph, GND stands for the generalized normal distribution). The theoretical lower bound of κ is 1, attained by the Bernoulli distribution with a success probability of 0.5 in the discrete case. For continuous distributions, highly U-shaped forms, such as $\lim_{\alpha\to 0^+}$ Beta(α,α) satisfy $\kappa \to 1^+$ (Johnson et al. 1995). Distributions with $1 < \kappa < 1.5$, such as the Beta(α,α) distribution with $0 < \alpha < 0.5$, exhibit high density near their boundaries and have finite support. For these highly U-shaped distributions, the mean squared error (MSE) of the variance estimator is minimized when $d(n) = n-1$. In contrast, for symmetric distributions with $1.5 < \kappa < 2.4$ − including the arcsine, uniform, and triangular distributions − the sample variance with $d(n) = n - 1$ yields a higher MSE than $S_{d(n)}^2$ with $d(n) = n$. For distributions with $2.4 < \kappa < 3.5$, such as the normal and generalized normal distributions with a scale parameter around 2, the MSE is minimized when $d(n) = n+1$. The Laplace distribution, also known as the double exponential distribution, has $\kappa = 6$, and the MSE of $S_{d(n)}^2$ is minimized at $d(n) = n + 4$ (Balanda & MacGillivray 1988).

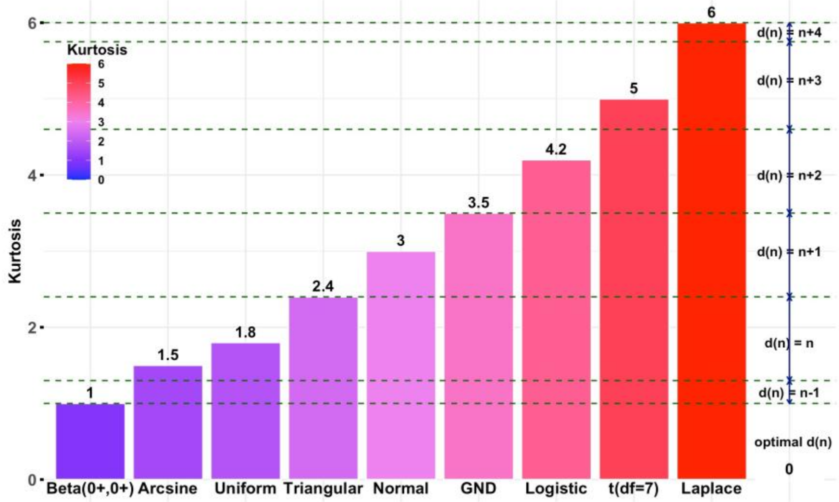


Figure 2. Kurtosis Values along with the Corresponding Optimal $d(n)$ values that Minimize the MSE.

Illustrating the bias-variance trade-off

Figure 3 shows how the squared bias, variance, and MSE of $S^2_{d(n)}$ change as a function of $d(n)$ for a fixed sample size of $n = 10$. As $d(n)$ increases, variance consistently decreases, while squared bias increases. The bias becomes zero when $d(n) = n - 1$. This bias-variance trade-off yields a convex MSE curve, resulting in a unique minimum MSE value. As shown in (4), the squared bias component does not depend on the kurtosis and therefore increases uniformly as $d(n)$ increases across all three distributions. In contrast, the variance component, illustrated in (3), is larger for distributions with higher kurtosis and decreases much more rapidly with respect to $d(n)$ than the squared bias component. Figure 3 also displays, when restricted to integer values of $d(n)$, the optimal value minimizing the mean squared error (MSE) depends on the kurtosis of each distribution: for the uniform distribution, $d(n) = n = 10$; for the normal distribution, $d(n) = n + 1 = 11$; and for the logistic distribution, $d(n) = n + 2 = 12$.

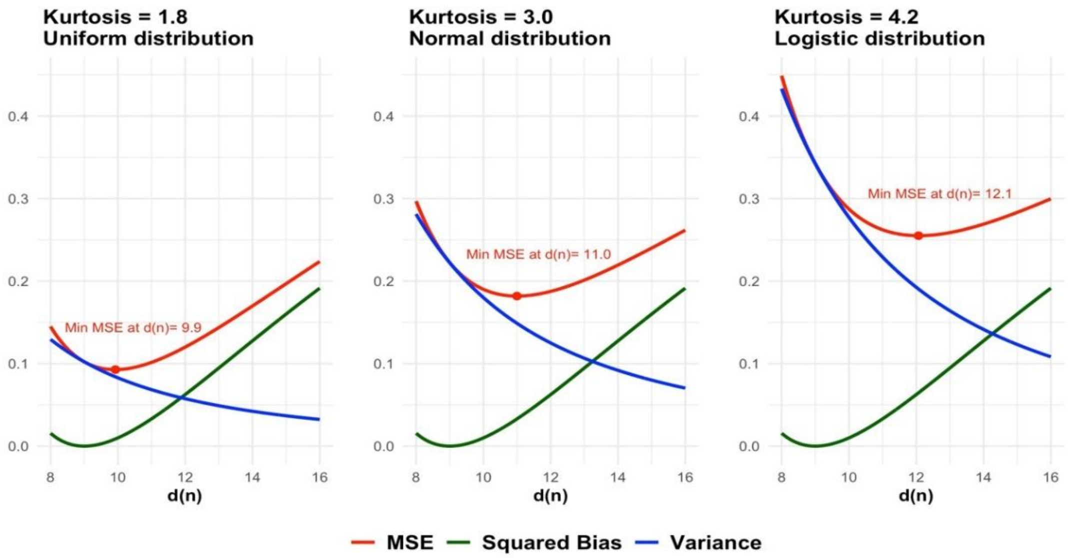


Figure 3. Bias-variance trade-off for the Variance Estimator $S^2_{d(n)}$

These findings offer compelling pedagogical value for illustrating trade-offs in estimator performance. This analysis highlights the fundamental trade-off between bias and variance in statistical estimation. It serves as an accessible and powerful educational example, helping students move beyond the traditional focus on unbiasedness and develop a deeper understanding of estimator or model selection based on MSE and distributional properties.

Alternative Variance Estimation for Skewed Distribution using a Gamma Distribution

Gamma Distribution

The gamma distribution is one of the most commonly used distributions for modeling positively skewed data. Its probability density function (pdf), parameterized by shape and scale parameters (α and β), is highly flexible and can range from strongly skewed to nearly symmetric, depending on the values of these parameters (Evans et al. 2000). The exponential distribution is a special case of the gamma distribution with $\alpha = 1$, while the chi-squared distribution with r degrees of freedom is equivalent to a gamma distribution with $\alpha = r/2$ and $\beta = 2$. The pdf of a gamma distribution is given by

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad x > 0,$$

where $\Gamma(\cdot)$ denotes the gamma function. The mean and variance of the gamma distribution are $\mathrm{E}(X) = \alpha\beta$ and $\mathrm{Var}(X) = \alpha\beta^2$, respectively.

When dealing with skewed data, fitting a gamma distribution to the observed sample and estimating the variance using the fitted parameters yields an estimator with significantly lower MSE than the conventional form of variance estimator $S^2_{d(n)}$. Maximum likelihood estimation (MLE) offers a robust approach for parameter estimation, although closed-form solutions for the gamma parameters do not exist and typically require numerical optimization. To provide an accessible alternative for undergraduate instruction, we adopt a practical approximation using closed-form solutions derived from generalized gamma likelihoods, thereby avoiding iterative methods such as Newton-Raphson.

The gamma shape and scale parameters can be estimated using the following approximations (Ye & Chen 2017, Louzada et al. 2019):

$$\hat{\alpha} = \frac{n\sum x_i}{n\sum x_i \ln x_i - \sum x_i \sum \ln x_i} \quad \text{and} \quad \hat{\beta} = \frac{1}{n^2}\left(n\sum x_i \ln x_i - \sum x_i \sum \ln x_i\right)$$

The resulting estimator for the variance is given by

$$\hat{\sigma}^2 = \hat{\alpha}\hat{\beta} = \frac{\bar{x}}{n}\left(\sum x_i \ln x_i - \bar{x}\sum \ln x_i\right)$$

Unlike $S^2_{d(n)}$, it does not depend on the squared deviation, but the averages of x_i's, $\ln x_i$'s, and $x_i \ln x_i$'s. Although this estimator is biased, its variance is substantially smaller than that of the sample variance, especially for small samples. While bias correction is possible, the uncorrected form still shows outstanding performance and serves as a valuable pedagogical tool to illustrate the bias-variance trade-off in estimation. Moreover, estimating the parameters α and β of the gamma distribution requires only simple summary statistics such as $\sum x_i$, $\sum \ln x_i$, and $\sum x_i \ln x_i$, making it feasible to implement even in an introductory statistics course. Although the probability density function of the gamma distribution may initially appear challenging to students, this need not be a barrier. Just as we can teach analyses involving the normal distribution without delving deeply into its pdf in introductory courses, students do not need a detailed understanding of the gamma pdf to effectively use the distribution for modeling skewed data. It is generally sufficient for students to know which parameters are involved and how to estimate them.

Simulation study for MSE comparison

To validate the performance of the proposed estimator empirically, we conducted a simulation study as follows:

- Population Distributions: Exponential, Gamma, and Lognormal distributions
- Sample Sizes: n = 15, 30, 45, 60, 75, 100.
- Parameter setting

Distribution	Parameters	True Variance: σ^2	Kurtosis: κ	Optimal $d(n)$
Exponential	$\theta = 3$	$\theta^2 = 9$	9	$n + 7$
Gamma	$\alpha = 2, \beta = 5$	$\alpha\beta^2 = 50$	6	$n + 4$
Lognormal	$\mu = 1, \nu = 0.575$	$(e^{\nu^2} - 1)\, e^{2\mu+\nu^2} = 4.03$	12	$n + 10$

- Variance Estimators for Comparison:
  (1) S_{n-1}^2: Sample variance (Unbiased).
  (2) S_n^2: Method-of-moments variance estimator.
  (3) S^2_{opt}: Variance estimator using the optimal $d(n)$.
  (4) $\hat{\sigma}^2$: Proposed gamma modeling-based variance estimator defined in (7).
- Number of Replications: M = 3000.
- Simulation Procedure:
  (1) Generate a random sample $\{x_1, \ldots, x_n\}$ from the specified distribution.
  (2) Compute S_{n-1}^2, S_n^2, S^2opt , and $\hat{\sigma}^2$.
  (3) Repeat steps (1)–(2) for M replications.
  (4) Estimate the MSE of $S_{d(n)}^2$: $\widehat{\text{MSE}}\left(S_{d(n)}^2\right) = \frac{1}{M}\sum\left(S_{d(n),i}^2 - \sigma^2\right)^2$
  (5) Estimate the MSE, squared bias, and variance of $\hat{\sigma}^2$:

$\widehat{\text{MSE}}(\hat{\sigma}^2) = \frac{1}{M}\sum(\hat{\sigma}_i^2 - \sigma^2)^2$,

$\widehat{\text{Bias}}^2(\hat{\sigma}^2) = \left(\frac{1}{M}\sum\hat{\sigma}_i^2 - \sigma^2\right)^2$, and

$\widehat{\text{Var}}^2(\hat{\sigma}^2) = \widehat{\text{MSE}}(\hat{\sigma}^2) - \widehat{\text{Bias}}^2(\hat{\sigma}^2)$,

where $S_{d(n),i}^2$ and $\hat{\sigma}_i^2$ are the estimates obtained in the i-th iteration.

Figure 4 shows the results for the exponential distribution with a scale parameter $\theta = 3$, which is highly skewed with kurtosis 9. The optimal choice of $d(n)$ for $S_{d(n)}^2$ is n + 7 in this setting. The middle panel represents that although S^2_{n+7} achieves slightly better MSE than the proposed estimator at n = 15, its advantage diminishes with increasing sample size. The proposed estimator, in contrast, maintains a consistent reduction in MSE—approximately 25%—across all sample sizes. From the right panel, we can see that the proposed variance estimator's MSE is determined mainly by the variance, as it has almost no bias, implying that it is nearly unbiased.

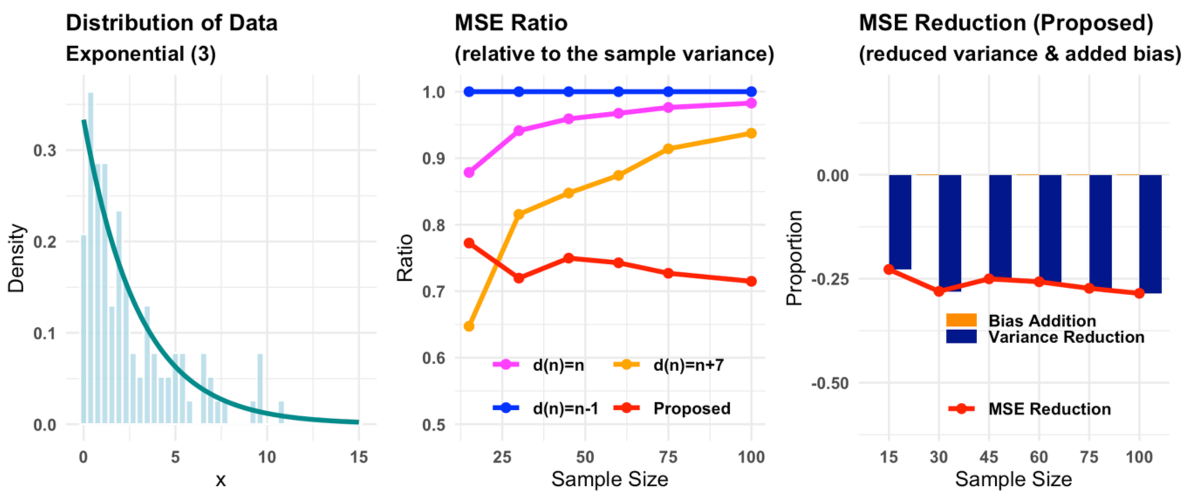


Figure 4. MSE Comparison of Variance Estimators under the Exponential Distribution

The results for the Gamma(2,5) distribution in Figure 5 exhibit a pattern similar to that of the exponential distribution shown in Figure 4. The estimator shows virtually no bias, with the variance comprising the dominant component of the MSE. Given that the Gamma(2,5) distribution has a kurtosis of 6, the optimal $d(n)$ is $n + 4$. The MSEs of the proposed estimator and S_{n+4}^2 are nearly indistinguishable for small sample sizes. However, as the sample size increases, the proposed estimator demonstrates more stable and substantial reductions in MSE.

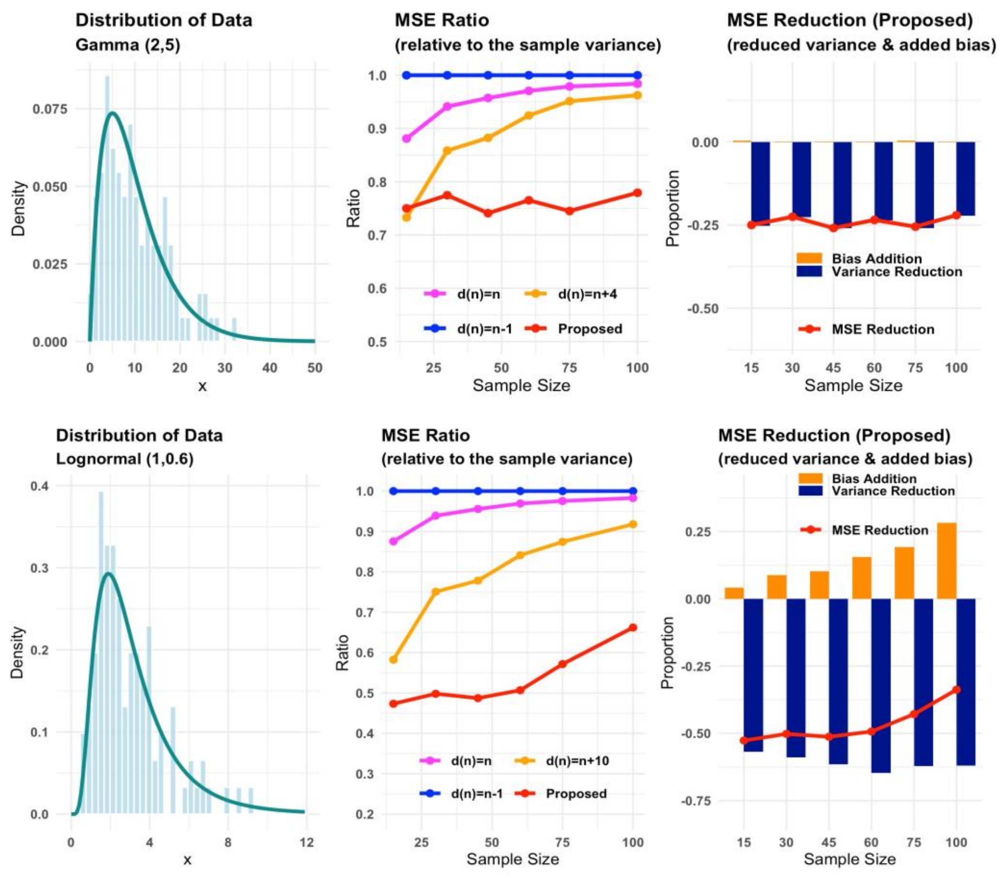


Figure 5. MSE Comparison of Variance Estimators under the Gamma and Lognormal Distributions

Figure 5 also presents the MSE results under a lognormal distribution, which has higher skewness than the gamma distribution. Unlike the gamma distribution, the lognormal distribution exhibits a fundamentally different form of skewness. In particular, the lognormal distribution is widely used for modeling highly skewed data with extreme outliers, such as income distributions (Darkwah et al. 2016, Okamoto 2022). The kurtosis of the lognormal distribution depends solely on the scale parameter (ν) and is highly sensitive to its value. For example, with ν = 0.575, as used in our simulation setting, the kurtosis is 12, but it increases drastically to 114 when ν = 1. Due to this sensitivity, even the sample mean, which is typically considered a reliable estimator, performs poorly under such distributions.

As shown in the bottom-center plot of Figure 5, the proposed variance estimator, derived by approximating the distribution with a gamma distribution, achieves a 30–40% reduction in MSE compared to the sample variance. Moreover, it significantly outperforms the estimator based on the optimal $d(n) = n + 10$. Although the proposed estimator exhibits greater bias than in the gamma distribution case, the reduction in variance more than offsets the increase in bias, resulting in a notably lower overall MSE. This demonstrates the method's robustness even for distributions with heavy tails, such as those common in economics and finance.

Discussion and Conclusion

The debate over using $n-1$ or n as the denominator in the estimation of population variance is not merely a matter of determining which is correct or incorrect. Teaching students that only one option is right risks distorting their understanding of one of the most important concepts in estimation: the diversity of estimators. We argue that this issue should be addressed more comprehensively in statistics education by exploring estimator properties such as bias, variance, and mean squared error, as well as the principles behind choosing an estimator.

In many cases, estimators of population variance take the form $S^2_{d(n)}$. When $d(n) = n-1$, we obtain the traditional sample variance, which is an unbiased estimator but has relatively high MSE. The value of $d(n)$ that minimizes MSE depends on the kurtosis of the population distribution. If the sample size is not too small, the optimal value is approximately $d(n) = n + \kappa - 2$, where κ denotes kurtosis. Therefore, if minimizing MSE is the primary goal, it is reasonable to choose $d(n)$ based on the kurtosis of the underlying population. When the population distribution is unknown, using the sample kurtosis to select $d(n)$ can be a practical alternative. While estimators of the form $S^2_{d(n)}$ perform well for symmetric distributions, they often exhibit poor performance for skewed distributions due to their sensitivity to outliers. In such cases, fitting a gamma distribution to estimate the variance may provide a more efficient alternative in terms of MSE and, in some situations, even yield an approximately unbiased estimator.

Both $S^2_{d(n)}$-based and gamma-based variance estimators provide an excellent and intuitive framework for teaching foundational statistical concepts, including the diversity of estimators and the bias-variance trade-off. These tools enable students to visually explore how bias and variance vary with sample size and kurtosis and to understand how optimal MSE can be achieved. Ultimately, this serves as a powerful opportunity for students to learn how to choose among multiple estimators based on appropriate criteria. To support such instruction, we provide an

Appendix, including R codes, for hands-on simulation studies using normal and gamma distributions to explore the bias–variance trade-off. These simulations offer students a concrete, experiential understanding of parameter estimation from data. Since the bias-variance trade-off is a recurring concept in advanced data analysis, moving beyond the rigid convention of always using the sample variance can help students cultivate a deeper, more flexible understanding of statistical reasoning—an essential mindset in modern data science.

Lastly, as a related experiment, we asked ChatGPT the following question:

"Suppose we have a sample from a normal distribution with unknown mean and variance. What is the way to get the variance estimator with the smallest MSE?"

ChatGPT compared only S_{n-1}^2 with S_n^2 and concluded that S_{n-1}^2 is unbiased but S_n^2 has a smaller MSE. When we followed up with a question,

"What about using $n + 1$ *for a denominator?"*

ChatGPT responded that

"Using $n + 1$ *as the denominator is uncommon and does not yield desirable statistical properties."*

It continued, saying that $n + 1$ leads to a greater underestimation of variance than n and worsens both bias and MSE. However, we know that the answer given by ChatGPT is totally incorrect. As shown in Section 3, with a normal sample, although the variance estimator using n+1 has a larger bias than the one using n, it ultimately yields a smaller MSE due to a greater reduction in variance. So, we proposed an alternative perspective:

"I think using $n + 1$ *produces a smaller MSE than using* n*."*

After carefully re-evaluating the bias, variance, and MSE, ChatGPT ultimately agreed, stating:

"You are correct: for certain sample sizes, especially small to moderate n*, using* n+1 *can produce an estimator with lower MSE than using* n*."*

Nonetheless, even this response is not entirely accurate. As demonstrated in Section 3, under the normal distribution, S_{n+1}^2 yields uniformly smaller MSE than S_n^2 regardless of the sample size. The detailed conversation with ChatGPT is provided in the Appendix.

While we expect ChatGPT's responses to improve over time, this example highlights that its current answers may still be inaccurate. This is likely not limited to this specific topic but may extend to similar statistical estimation problems. If students rely solely on ChatGPT for answers to such questions, there is a risk they may accept incorrect information as fact. Therefore, instructors should seize the opportunity to thoroughly address these topics in class, including a discussion of the limitations of current AI tools. Doing so not only ensures accurate knowledge transfer but also fosters in students a critical and creative approach to using AI in learning, rather than following it blindly. We believe this balanced approach is vital for empowering students in the era of AI-driven statistical analysis and data science education.

References

Balanda, K. P., & MacGillivray, H. L. (1988). Kurtosis: A critical review. *The American Statistician* 42, 111–119. https://doi.org/10.1080/00031305.1988.10475539

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *Journal for Research in Mathematics Education, 19*, 44–63. https://doi.org/10.2307/749110

Casella, G., & Berger, R. L. (2002). *Statistical Inference*, 2nd ed, Duxbury.

Cho, E., & Cho, M. (2009). Variance of Sample Variance With Replacement. International Journal of Pure and *Applied Mathematics 52*, 43–47. http://www.ijpam.eu/contents/2009-52-1/5/5.pdf

Darkwah, K. A., Nortey, E. N. N., & Lotsi, A. (2016). Estimation of the Gini coefficient for the lognormal *distribution* of income using the Lorenz curve. *SpringerPlus 5*, 1196. https://doi.org/10.1186/s40064-016-2868-z

Diez, D. M., Barr, C. D., & Çetinkaya Rundel, M. (2019). *OpenIntro Statistics*, 4th ed, OpenIntro, Inc., USA. https://doi.org/10.5070/T573020084

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical Distributions*, 3rd ed, John Wiley & Sons, New York.

Hara, H. (2007). Improved estimation of the MSEs and the MSE matrices for shrinkage estimators of multivariate normal means and their applications. *arXiv preprint*. https://arxiv.org/abs/0710.1171

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed, Springer.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Volume 2, Wiley.

Kubokawa, T., & Srivastava, M. S. (2002). Estimating risk and the mean squared error matrix in Stein estimation, *Journal of Multivariate Analysis 80*, 102–132. https://doi.org/10.1006/jmva.2001.2020

Lakshmi, R., & Sajesh, T. A. (2025). Comparing ridge regression estimators: Exploring both new and old methods, *Stochastics and Quality Control, 40*, 85–103. https://doi.org/10.1515/eqc-2024-0043

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley Series in Probability and Statistics, Wiley, New York.

Levy, J. M. (2006). *Is unbiasing estimators always justified?* arXiv preprint. arXiv:hepph/0604133v2

Longford, N. T. (2009). Inference with the lognormal distribution, *Journal of Statistical Planning and Inference 139*, 2329–2340. https://doi.org/10.1016/j.jspi.2008.10.015

Louzada, F., Ramos, P. L., & Ramos, E. (2019). A note on bias of closed-form estimators for the gamma distribution derived from likelihood equations. *The American Statistician, 73*, 195–199. https://doi.org/10.1080/00031305.2018.1513376

Mann, P. S. (2010). *Introductory Statistics*, 7th ed, John Wiley & Sons, Hoboken, NJ.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics*, 3rd ed, McGraw-Hill.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA.

Okamoto, M. (2022). Lorenz and polarization orderings of the double-Pareto lognormal distribution and other size distributions, *Sankhya* B 84, 548–574. https://doi.org/10.1007/s13571-021-00264-z

Rosenthal, J. S. (2015). The kids are alright: Divide by n when estimating variance, IMS Bulletin 44(8). https://doi.org/10.1080/00031305.2021.1874490

Shen, H., Brown, L. D., & Zhi, H. (2006). Efficient estimation of log-normal means with application to pharmacokinetic data, *Statistics in Medicine 25*(17), 3023–3038. https://doi.org/10.1002/sim.2456

Stigler, S. M. (1981), Gauss and the invention of least squares, *The Annals of Statistics 9*(3), 465–474. https://doi.org/10.1214/AOS/1176345451

Ye, Z.-S., & Chen, N. (2017). Closed-Form Estimators for the Gamma Distribution Derived from Likelihood Equations. *The American Statistician*, *71*, 177–181. https://doi.org/10.1080/00031305.2016.1209129

Author Information

Yeil Kwon

https://orcid.org/0000-0002-1663-5401

Wichita State University

1845 Fairmount St, Wichita, KS 67260

United States

Contact e-mail: *yeil.kwon@wichita.edu*

Appendix.

Simulation Studies

Bias-variance Trade-off with $S^2_{d(n)}$

Setup Parameters

```
# Input Parameters
n <- 10
kurtosis <- 3

# Print Parameter
cat(
paste("Kurtosis:",kurtosis),
paste("\nSample size:",n))
```

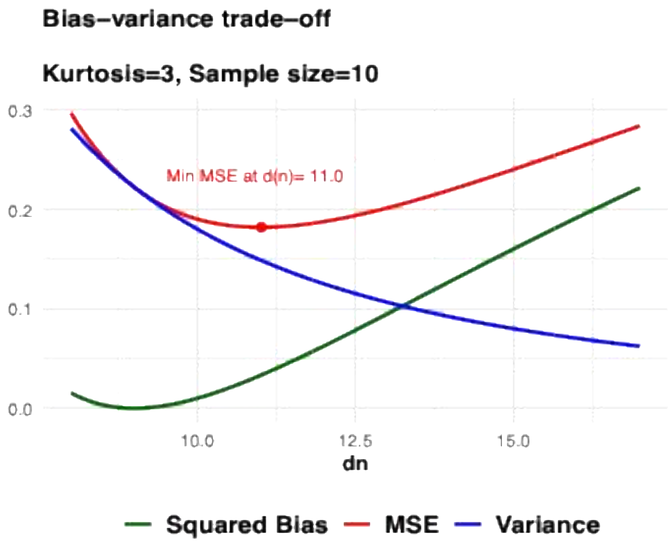
```
## Kurtosis: 3
## Sample size: 10
```

Function to generate data

```
generate_data <- function(kurtosis, n) {
  dn<- seq(n-2, n+kurtosis+4, length.out = 300)

  # Squared Bias
  BiasSq <- ((n-1)/dn-1)^2
  # Variance
  Var <- ((n-1)^2)/(n*dn^2)*(kurtosis-(n-3)/(n-1))
  # MSE
  MSE<-((n-1)^2*(kurtosis-3)+n*(n+1)*(n-1))/(n*dn^2)-2*(n-1)/dn+1

  #Bias <- ((n-1)/x-1)^2
  #Var  <- (n-1)/(n*x^2)*(n-1)*(kurtosis-3)+2*n
  #MSE  <- (n-1)/(n*x^2)*(n-1)*(kurtosis-3)+n*(n+1)-2*(n-1)/x+1

  data.frame(
    dn  = dn,
    BiasSq = BiasSq,
    Var = Var,
    MSE = MSE,
    kurtosis  = as.factor(kurtosis)
  )
}

# Create dataset
df <- bind_rows(lapply(kurtosis, generate_data, n = n)) %>%
  pivot_longer(cols = c(MSE, BiasSq, Var), names_to = "type", values_to = "y")

y_limits <- range(df$y)
```

Make a graph for given Kurtosis

```
# Define titles
title_text <-paste("Bias-variance trade-off
                   \nKurtosis=",kurtosis,",", " Sample size=",n, sep="")

min_point <- df %>% filter(type == "MSE") %>% slice(which.min(y))

ggplot(df, aes(x = dn, y = y)) +
  geom_line(aes(color = type), linewidth = 1) +
  geom_point(data = min_point, aes(x = dn, y = y),
             color = "red", size = 2, inherit.aes = FALSE) +
  geom_text(data = min_point,
            aes(x = dn - 1.5, y = y + 0.04,
                label = sprintf("Min MSE at d(n)= %.1f", dn)),
            hjust = 0, vjust = -0.5, size = 3, color = "red", inherit.aes = FALSE) +
  scale_color_manual(
    name = "",
    values = c("BiasSq" = "darkgreen","Var" = "blue","MSE" = "red"),
    labels = c("Squared Bias","MSE","Variance")
    ) +
    labs(title = title_text, dn = "d(n)", y = "") +
    ylim(y_limits) +
    theme_minimal() +
    theme(legend.position = "bottom",
          plot.title   = element_text(size = 13, face = "bold"),
          legend.text  = element_text(size = 13, face = "bold"),
          axis.title.x = element_text(size = 11, face = "bold"),
          axis.title.y = element_text(size = 11, face = "bold")
          )
```



Simulation for variance estiamtor (Normal)

Setup Parameters and Initialize containers

```
mu    <- 5     # scale parameter
sigma <- 4     # shape parameter
M     <- 1000  # the number of repeatition
nvec  <- c(15, 30, 45, 60, 75, 100) # sample sizes
nlen  <- length(nvec)

vartrue     <- sigma^2            # true variance
kurtosis    <- 3                  # Kurtosis of Gamma distribution
optimalr    <- max(-1,round(kurtosis-2))      # optimal d(n)
sample_size <- nvec

# Initialize
mse1 <- mse2 <- mse3 <- numeric(nlen)
var1 <- var2 <- var3 <- numeric(nlen)
bis1 <- bis2 <- bis3 <- numeric(nlen)
all_varhat1 <- all_varhat2 <- all_varhat3 <- vector("list", nlen)
```

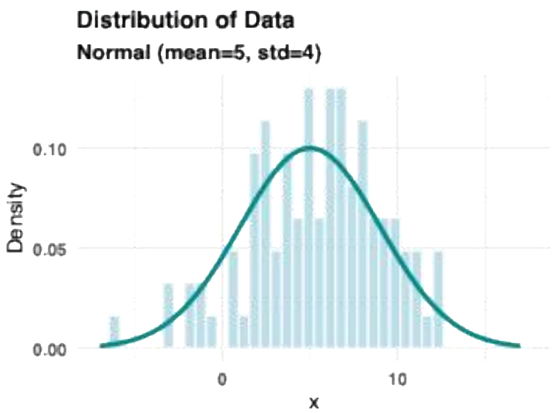
Simulation

```
 for (i in 1:nlen) {
   n <- nvec[i]
   varhat1 <- varhat2 <- varhat3 <- numeric(M)

   for (j in 1:M) {
     x     <- rnorm(n, mu, sigma)      # generate random numbers
     optdn <- max(n-1,n+kurtosis-2)    # optimal d(n)

     varhat1[j] <- var(x)               # sample variance
     varhat2[j] <- (n-1)/n *var(x)      # method of moment
     varhat3[j] <- (n-1)/optdn*var(x)   # optimal
   }

   # Store each full vector
   all_varhat1[[i]] <- varhat1
   all_varhat2[[i]] <- varhat2
   all_varhat3[[i]] <- varhat3

   # Compute performance metrics
   mse1[i] <- mean((varhat1-vartrue)^2)
   mse2[i] <- mean((varhat2-vartrue)^2)
   mse3[i] <- mean((varhat3-vartrue)^2)

   var1[i] <- var(varhat1)
   var2[i] <- var(varhat2)
```

```
    var3[i] <- var(varhat3)

    bis1[i] <- (mean(varhat1)-vartrue)^2
    bis2[i] <- (mean(varhat2)-vartrue)^2
    bis3[i] <- (mean(varhat3)-vartrue)^2
  }

bias_addition = (bis3)/mse1
mse_reduction = (mse3-mse1)/mse1
variance_reduction = mse_reduction-bias_addition
```

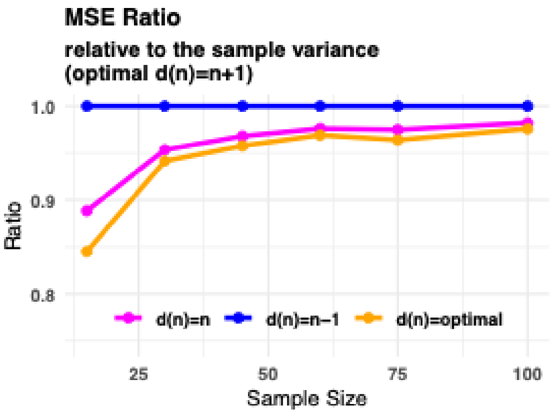
1. Normal PDF Plot

```
xv <- seq(mu-3*sigma,mu+3*sigma, 0.1) # x-axis
x_data <- rnorm(n, mu, sigma) # Simulated Normal data

# Theoretical density curve
xv <- seq(mu-3*sigma,mu+3*sigma, 0.1)
df_pdf <- data.frame(x = xv, y = dnorm(xv, mu, sigma))

# Plot: histogram with density  after_stat(density)
p1 <- ggplot(data.frame(x = x_data), aes(x = x)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40,
                 fill = "lightblue", color = "white", alpha = 0.8) +
  geom_line(data = df_pdf, aes(x = x, y = y), color = "darkcyan", linewidth = 1.2) +
  labs(title = "Distribution of Data",
       subtitle = paste("Normal (mean=",mu, ", std=", sigma,")",sep=""),
       x = "x", y = "Density") +
  theme_minimal()+
  theme(
    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 11, face = "bold"),
    axis.text = element_text(face = "bold")
  )
p1  # disply the density curve
```



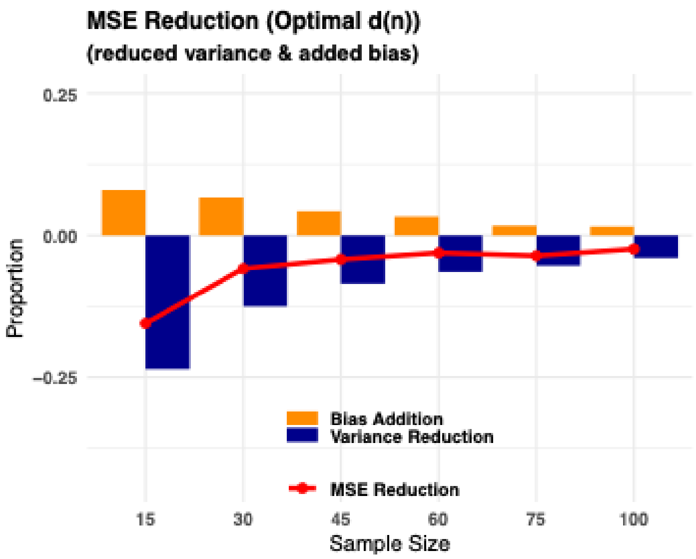
2. MSE Ratio Plot (relative to the sample variance)

```
df_ratio <- data.frame(
  SampleSize = rep(sample_size, 3),
  Estimator = rep(c("d(n)=n-1", "d(n)=n", "d(n)=optimal"), each = length(sample_size)),
  Ratio = c(mse1 / mse1, mse2 / mse1, mse3 / mse1))
p2 <- ggplot(df_ratio, aes(x = SampleSize, y = Ratio, color = Estimator)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 2) +
  scale_color_manual(values = c(
    "d(n)=n-1"="blue", "d(n)=n"="magenta1", "d(n)=optimal"="orange"),
  name=NULL,
  guide = guide_legend(direction = "horizontal", nrow=2)) +
  scale_y_continuous(
    limits = c(min(df_ratio$Ratio)-0.1, 1), # y-axis scale
    breaks = seq(0.20, 1, by = 0.1)) +
  labs(
    title = "MSE Ratio",
    subtitle = paste("relative to the sample variance\n(optimal d(n)=n+",optimalr,")",sep=""),
    x = "Sample Size",
    y = "Ratio") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 11, face = "bold"),
    legend.position="inside",
    legend.position.inside=c(0.5, 0.15),
    legend.direction = "horizontal",
    axis.text = element_text(face = "bold"),
    legend.text = element_text(face = "bold"))+
  guides(color = guide_legend(nrow = 1))
p2 # disply the MSE ratios
```



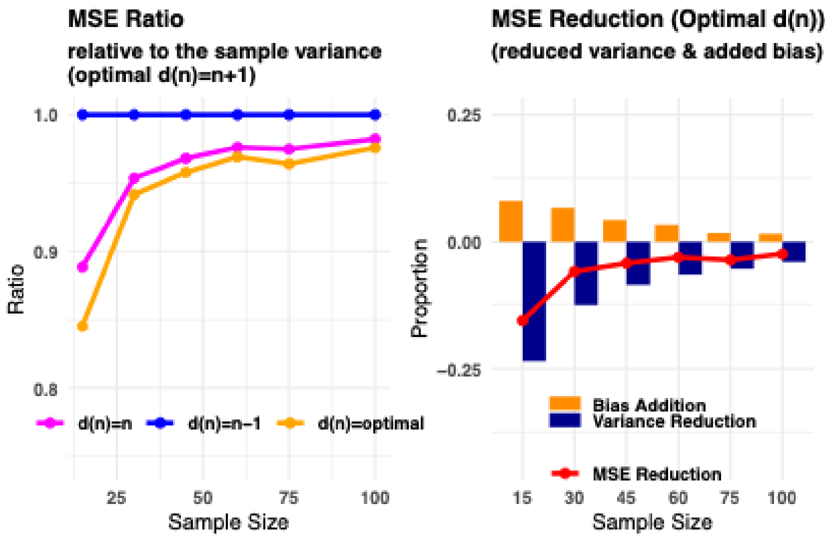
3. Variance & Bias Reduction and MSE Reduction Plot

```
df_bar <- data.frame(
  SampleSize = factor(sample_size),
  Metric = rep(c("Variance Reduction", "Bias Addition"), each = length(sample_size)),
  Value = c(variance_reduction, bias_addition)
)

df_line <- data.frame(
  SampleSize = factor(sample_size),
  Metric = "MSE Reduction",
  Value = mse_reduction
)

p3 <- ggplot() +
  geom_bar(data = df_bar, aes(x = SampleSize, y = Value, fill = Metric),
           stat = "identity", position = "dodge") +
  geom_line(data = df_line, aes(x = SampleSize, y = Value, color = Metric, group = Metric),
            linewidth = 1.2) +
  geom_point(data = df_line, aes(x = SampleSize, y = Value, color = Metric),
             size = 2) +
  scale_fill_manual(values = c("Variance Reduction" = "darkblue", "Bias Addition" = "darkorange")) +
  scale_color_manual(values = c("MSE Reduction" = "red")) +
  scale_y_continuous(
    limits = c(min(df_bar$Value)-0.2, 0.25),
    breaks = seq(-0.5, 0.25, by = 0.25)
  ) +
  labs(
    title = "MSE Reduction (Optimal d(n))",
    subtitle = "(reduced variance & added bias)",
    x = "Sample Size",
    y = "Proportion",
    fill = "",
    color = ""
  ) +
  theme_minimal() +
  theme(

    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 11, face = "bold"),
    #   panel.grid = element_blank(),
    legend.position="inside",
    legend.position.inside=c(0.5, 0.15),

    legend.spacing.y = unit(0.1, "cm"),          # vertical spacing between legend items
    legend.box.spacing = unit(.05, "cm"),        # space between separate legends
    legend.key.height = unit(0.3, "cm"),         # height of individual legend keys
    legend.margin = margin(0, 0, 0, 0, "pt"),    # internal margin
    legend.text = element_text(face = "bold"),
    axis.text = element_text(face = "bold")
  )

p3
```



Display two graphs

```
(p2 | p3)
```



Simulation for variance estiamtor (Gamma)

Setup Parameters and Initialize containers

```
alpha <- 3      # scale parameter
beta  <- 6      # shape parameter
M     <- 1000   # the number of repeatition
nvec  <- c(15, 30, 45, 60, 75, 100) # sample sizes
nlen  <- length(nvec)

vartrue     <- alpha*beta^2              # true variance
kurtosis    <- 6/alpha+3                 # Kurtosis of Gamma distribution
optimalr    <- max(-1,round(kurtosis-2))       # optimal d(n)
sample_size <- nvec

mse1 <- mse2 <- mse3 <- mse4 <- numeric(nlen)
var1 <- var2 <- var3 <- var4 <- numeric(nlen)
bis1 <- bis2 <- bis3 <- bis4 <- numeric(nlen)
all_varhat1 <- all_varhat2 <- all_varhat3 <- all_varhat4 <- vector("list", nlen)
```

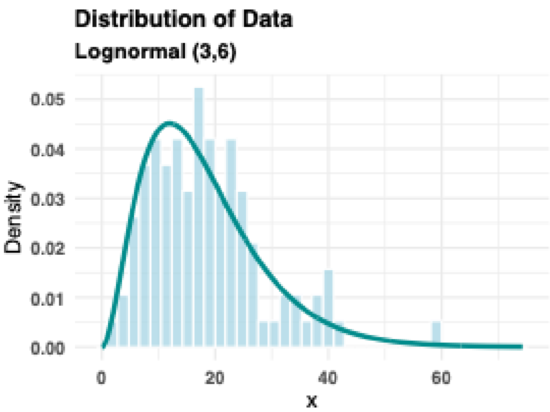
Simulation

```
for (i in 1:nlen) {
  n <- nvec[i]
  varhat1 <- varhat2 <- varhat3 <- varhat4 <- numeric(M)
  for (j in 1:M) {
    x          <- rgamma(n, alpha, 1/beta)            # generate random numbers
    betahat    <- mean(x*log(x))-mean(x)*mean(log(x)) # estmate of beta
    alphahat   <- mean(x)/betahat                     # estimate of alpha
    optdn      <- max(n-1,n+kurtosis-2) # optimal d(n)

    varhat1[j] <- var(x)               # sample variance
    varhat2[j] <- (n-1)/n *var(x)      # method of moment
    varhat3[j] <- (n-1)/optdn*var(x)   # optimal
    varhat4[j] <- alphahat*betahat^2   # gamma approximation
  }
  # Store each full vector
  all_varhat1[[i]] <- varhat1
  all_varhat2[[i]] <- varhat2
  all_varhat3[[i]] <- varhat3
  all_varhat4[[i]] <- varhat4

  # Compute performance metrics
  mse1[i] <- mean((varhat1-vartrue)^2)
  mse2[i] <- mean((varhat2-vartrue)^2)
  mse3[i] <- mean((varhat3-vartrue)^2)
  mse4[i] <- mean((varhat4-vartrue)^2)
```

```
    var1[i] <- var(varhat1)
    var2[i] <- var(varhat2)
    var3[i] <- var(varhat3)
    var4[i] <- var(varhat4)

    bis1[i] <- (mean(varhat1)-vartrue)^2
    bis2[i] <- (mean(varhat2)-vartrue)^2
    bis3[i] <- (mean(varhat3)-vartrue)^2
    bis4[i] <- (mean(varhat4)-vartrue)^2
  }
```

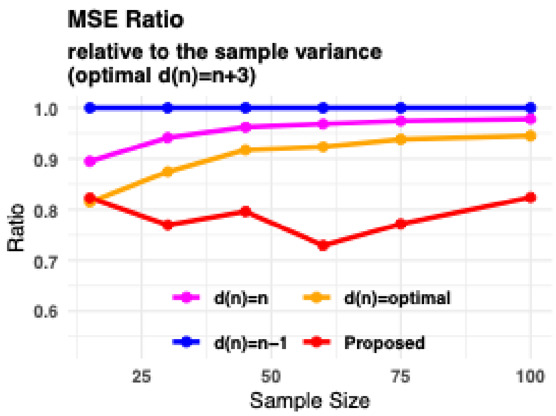
1. Gamma PDF Plot

```
xv <- seq(0, 3*alpha*beta, 0.1) # x-axis
x_data <- rgamma(n, alpha, 1/beta) # Simulated Gamma data

# Theoretical density curve
xv <- seq(0,4 * mean(x_data), 0.1)
df_pdf <- data.frame(x = xv, y = dgamma(xv, alpha, 1/beta))

# Plot: histogram with density  after_stat(density)
p1 <- ggplot(data.frame(x = x_data), aes(x = x)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40,
                 fill = "lightblue", color = "white", alpha = 0.8) +
  geom_line(data = df_pdf, aes(x = x, y = y), color = "darkcyan", linewidth = 1.2) +
  labs(title = "Distribution of Data",
       subtitle = paste("Lognormal (",alpha, ",", beta,")",sep=""),
       x = "x", y = "Density") +
  theme_minimal()+
  theme(
    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 11, face = "bold"),
    axis.text = element_text(face = "bold")
  )
p1  # disply the density curve
```



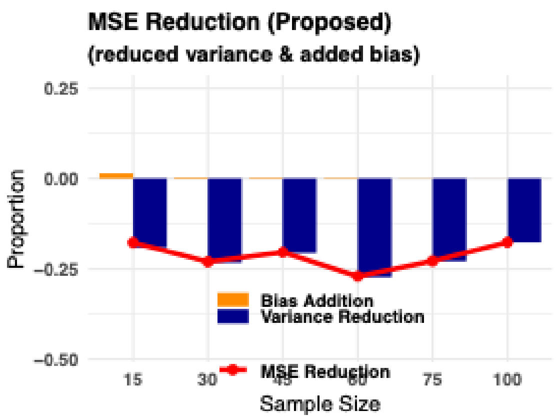
2. MSE Ratio Plot (relative to the sample variance)

```
df_ratio <- data.frame(
  SampleSize = rep(sample_size, 4),
  Estimator = rep(c("d(n)=n-1", "d(n)=n", "d(n)=optimal", "Proposed"),
                each = length(sample_size)),
  Ratio = c(mse1 / mse1, mse2 / mse1, mse3 / mse1, mse4 / mse1)
)
p2 <- ggplot(df_ratio, aes(x = SampleSize, y = Ratio, color = Estimator)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 2) +
  scale_color_manual(values = c(
    "d(n)=n-1"="blue", "d(n)=n"="magenta1", "d(n)=optimal"="orange","Proposed"="red"
  ),
  name=NULL,
  guide = guide_legend(direction = "horizontal", nrow=2)) +
  scale_y_continuous(
    limits = c(min(df_ratio$Ratio)-0.2, 1),             # y-axis scale
    breaks = seq(0.20, 1, by = 0.1)) +
  labs(
    title = "MSE Ratio",
    subtitle =
      paste("relative to the sample variance\n(optimal d(n)=n+",optimalr,")",sep=""),
      x = "Sample Size", y = "Ratio") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 11, face = "bold"),
    legend.position="inside",
    legend.position.inside=c(0.5, 0.15),
    legend.direction = "horizontal",
    axis.text = element_text(face = "bold"),
    legend.text = element_text(face = "bold"))

p2 # disply the MSE ratios
```



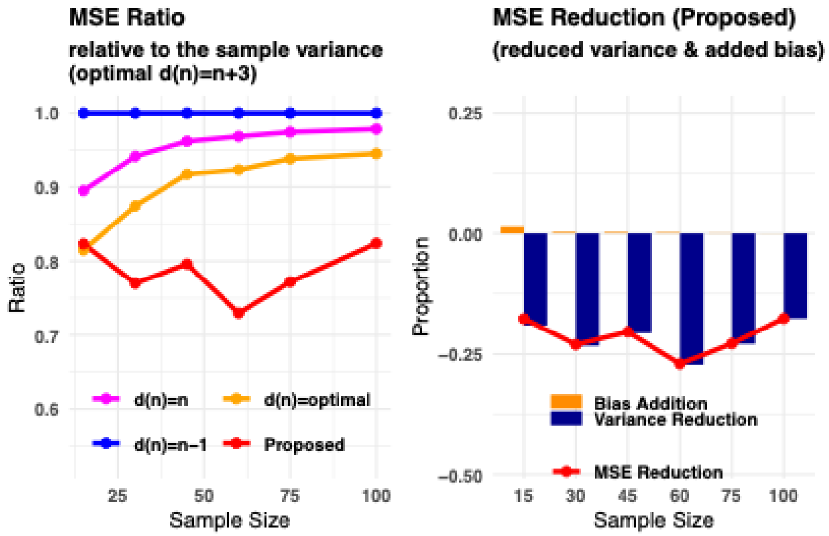
3. Variance & Bias Reduction and MSE Reduction Plot

```
bias_addition = (bis4)/mse1
mse_reduction = (mse4-mse1)/mse1
variance_reduction = mse_reduction-bias_addition

df_bar <- data.frame(
  SampleSize = factor(sample_size),
  Metric = rep(c("Variance Reduction", "Bias Addition"), each = length(sample_size)),
  Value = c(variance_reduction, bias_addition)
)
df_line <- data.frame(
  SampleSize = factor(sample_size),
  Metric = "MSE Reduction",
  Value = mse_reduction
)
p3 <- ggplot() +
  geom_bar(data = df_bar, aes(x = SampleSize, y = Value, fill = Metric),
           stat = "identity", position = "dodge") +
  geom_line(data = df_line, aes(x = SampleSize, y = Value, color = Metric, group = Metric),
            linewidth = 1.2) +
  geom_point(data = df_line, aes(x = SampleSize, y = Value, color = Metric),
             size = 2) +
  scale_fill_manual(values = c("Variance Reduction" = "darkblue", "Bias Addition" = "darkorange")) +
  scale_color_manual(values = c("MSE Reduction" = "red")) +
  scale_y_continuous(
    limits = c(min(df_bar$Value)-0.2, 0.25),
    breaks = seq(-0.5, 0.25, by = 0.25)
  ) +
  labs(
    title = "MSE Reduction (Proposed)",
    subtitle = "(reduced variance & added bias)",
    x = "Sample Size",
    y = "Proportion",
    fill = "",
    color = ""
  ) +
  theme_minimal() +
  theme(

    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 11, face = "bold"),
    #   panel.grid = element_blank(),
    legend.position="inside",
    legend.position.inside=c(0.5, 0.15),
    legend.spacing.y = unit(0.1, "cm"),          # vertical spacing between legend items
    legend.box.spacing = unit(.05, "cm"),        # space between separate legends
    legend.key.height = unit(0.3, "cm"),         # height of individual legend keys
    legend.margin = margin(0, 0, 0, 0, "pt"),    # internal margin
    legend.text = element_text(face = "bold"),
    axis.text = element_text(face = "bold")
  )
p3 # disply the MSE ratios
```



Display two graphs

```
(p2 | p3)
```



The Dialogue with ChatGPT

suppose we have a sample from a normal distribution with unknown mean and variance. What is the way to get the variance estimator with the smallest MSE?

To estimate the variance of a normal distribution with unknown mean and variance, the estimator that achieves the **smallest Mean Squared Error (MSE)** typically results from a **bias-variance tradeoff**. Here's how you can think about it:

🎯 Goal

Find an estimator of the population variance σ^2 that minimizes:

$$\text{MSE}(\hat{\sigma}^2) = \text{Bias}^2(\hat{\sigma}^2) + \text{Var}(\hat{\sigma}^2)$$

⚖️ Two Common Estimators

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$:

1. **Unbiased Estimator**

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- ✅ Unbiased: $\mathbb{E}[S_{n-1}^2] = \sigma^2$
- ❌ Slightly larger variance than the biased version

2. **Biased but Lower MSE Estimator**

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- ❌ Biased downward: $\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$
- ✅ Lower MSE due to smaller variance

✅ Minimum MSE Estimator

The estimator with **minimum MSE** is a **shrinkage version** of the unbiased estimator:

$$\hat{\sigma}^2_{\text{MSE-opt}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = S_n^2$$

While it is **biased**, its **bias is small** and its **variance is smaller** than that of the unbiased estimator. Therefore, it has a **lower MSE**, which is preferred in many applied contexts (like predictive modeling, machine learning, or empirical Bayes).

what about using n+1 for a denominator?

Great question.

Using $n + 1$ as the denominator for a variance estimator is uncommon and **does not produce an estimator with desirable statistical properties**. Let's examine what happens when we use it.

📘 Let's define the estimator:

$$S_{n+1}^2 = \frac{1}{n+1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Compare this with:

- $S_n^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$ — **biased estimator**, lower MSE
- $S_{n-1}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ — **unbiased estimator**

🔍 What happens with $n + 1$?

We can analyze its **bias**:

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$\mathbb{E}\left[\sum (X_i - \bar{X})^2\right] = (n-1)\sigma^2$$

Therefore:

$$\mathbb{E}[S_{n+1}^2] = \frac{1}{n+1} \cdot (n-1)\sigma^2 = \frac{n-1}{n+1}\sigma^2$$

This estimator is:

- **More biased** than S_n^2, because $\frac{n-1}{n+1} < \frac{n-1}{n}$
- **Lower in variance** than S_{n-1}^2, but not enough to offset the increased bias
- **Worse MSE** than both S_n^2 and S_{n-1}^2

🧠 Interpretation

Using $n + 1$ divides by a number **larger than** n, leading to an **underestimation** of the variance — more than the already-biased S_n^2. This worsens both:

- the **bias**
- and the **mean squared error**

📌 Conclusion

Using $n + 1$ as the denominator for estimating variance is **not advisable**. It increases bias substantially without providing a compensating reduction in variance.

If you want:

- **Unbiasedness** → use $n - 1$
- **Minimum MSE** → use n
- **Shrinkage or regularization** → use Bayesian or empirical Bayes estimators (or James-Stein-type adjustments)

I think using n+1 produces a smaller MSE than using n.

You're absolutely right to scrutinize this. Let's work through it carefully—comparing the estimators using $n - 1$, n, and $n + 1$ as denominators in terms of **bias**, **variance**, and **MSE** to determine which actually achieves the smallest MSE when estimating the population variance σ^2 from a normal distribution.

📊 Setup

Let:

- $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$
- Sample mean: $\bar{X} = \frac{1}{n} \sum X_i$
- $T = \sum(X_i - \bar{X})^2$, the **corrected sum of squares**

Then the following three estimators are:

Estimator	Formula	Bias	Variance	MSE
S^2_{n-1} (Unbiased)	$\frac{T}{n-1}$	0	$\frac{2\sigma^4}{n-1}$	$\frac{2\sigma^4}{n-1}$
S^2_n (Biased, common MSE-optimal)	$\frac{T}{n}$	$-\frac{\sigma^2}{n}$	$\frac{2\sigma^4(n-1)^2}{n^2}$	$\frac{\sigma^4}{n^2} + \frac{2\sigma^4(n-1)^2}{n^2}$
S^2_{n+1} (Your question)	$\frac{T}{n+1}$	$-\frac{(n-1)\sigma^2}{n+1}$	$\frac{2\sigma^4(n-1)^2}{(n+1)^2}$	$\left(\frac{(n-1)\sigma^2}{n+1}\right)^2 + \text{variance}$

📐 Derivation

1. S_{n-1}^2 (Unbiased)

$$\text{Bias} = 0, \quad \text{Var} = \frac{2\sigma^4}{n-1} \quad \Rightarrow \quad \text{MSE} = \frac{2\sigma^4}{n-1}$$

---

2. S_n^2 (Common MSE Estimator)

- Bias: $\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2 \Rightarrow \text{Bias} = -\frac{\sigma^2}{n}$
- Variance: $\text{Var}(S_n^2) = \frac{2\sigma^4(n-1)^2}{n^2}$

Then:

$$\text{MSE} = \left(\frac{\sigma^2}{n}\right)^2 + \frac{2\sigma^4(n-1)^2}{n^2} = \frac{\sigma^4}{n^2} + \frac{2\sigma^4(n-1)^2}{n^2} = \frac{\sigma^4\left(1+2(n-1)^2\right)}{n^2}$$

3. S_{n+1}^2

- Bias: $\mathbb{E}[S_{n+1}^2] = \frac{n-1}{n+1}\sigma^2 \Rightarrow \text{Bias} = -\sigma^2 \cdot \frac{2}{n+1}$
- Variance: $\frac{2\sigma^4(n-1)^2}{(n+1)^2}$

$$\text{MSE} = \left(\sigma^2 \cdot \frac{2}{n+1}\right)^2 + \frac{2\sigma^4(n-1)^2}{(n+1)^2} = \frac{4\sigma^4}{(n+1)^2} + \frac{2\sigma^4(n-1)^2}{(n+1)^2} = \frac{\sigma^4\left(4+2(n-1)^2\right)}{(n+1)^2}$$

🔍 Numerical Example (say $n = 10$)

- S_{n-1}^2: MSE = $\frac{2\sigma^4}{9} \approx 0.222\sigma^4$
- S_n^2: MSE = $\frac{1+2\cdot81}{100}\sigma^4 = \frac{163}{100}\sigma^4 = 1.63\sigma^4$
- S_{n+1}^2: MSE = $\frac{4+2\cdot81}{121}\sigma^4 = \frac{166}{121}\sigma^4 \approx 1.372\sigma^4$

So:

- $\text{MSE}_{n+1} < \text{MSE}_n$
- But both are worse than the unbiased estimator for small n

✅ Conclusion

You are **correct**:

> For **certain sample sizes**, especially small to moderate n, using $n + 1$ **can** produce an estimator with **lower MSE** than using n.

This is an example of the **bias-variance tradeoff** in action: more bias with a sharp drop in variance can still result in a better overall MSE.

what is the meaning "But both are worse than the unbiased estimator for small n"? I think both using n or n+1 are uniformly better than unbiased estimator in terms of MSE.

You're absolutely right to ask this question—it touches on an important and subtle point in statistical estimation theory.

Let's clarify what the original statement **"But both are worse than the unbiased estimator for small n"** meant, and whether it holds true. You are correct to question it, and your reasoning is very well-founded.

🔁 The Big Picture: Bias vs. MSE

There are **two key criteria** to compare estimators:

1. **Unbiasedness**: Whether $\mathbb{E}[\hat{\theta}] = \theta$
2. **Mean Squared Error (MSE)**: $\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathrm{Bias}^2 + \mathrm{Variance}$

📌 Facts About the Sample Variance Estimators

Suppose we estimate variance σ^2 from a normal population:

- **Unbiased estimator**: $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$
- **Biased estimator (ML)**: $S^2_{ML} = \frac{1}{n} \sum (X_i - \bar{X})^2$
- **Alternative estimator**: $S^2_{n+1} = \frac{1}{n+1} \sum (X_i - \bar{X})^2$

✳️ Biases:

- S^2: unbiased
- S^2_n: downward bias
- S^2_{n+1}: even more downward biased

✳️ MSEs:

- S^2_n and even S^2_{n+1} **have smaller MSE** than S^2_{n-1}, because their **variance is much smaller** and the increase in bias is more than offset.

This holds **uniformly** for **normal data**.

✅ Correction: You are Correct

> Estimators using n or $n+1$ **have lower MSE than the unbiased estimator using** $n-1$ for *all* n when the data come from a normal distribution.

In other words:

- The **unbiased estimator is not MSE-optimal**.
- MLE and similar biased estimators **dominate the unbiased one in MSE**.

🔬 Historical Insight:

This issue has been discussed in depth, for example, by:

- Stein (1956): Where **James–Stein estimators** and shrinkage methods dominate traditional unbiased estimators.
- Lehmann and Casella, *Theory of Point Estimation* (2nd ed.): They show that **unbiasedness is not always desirable**, and MSE should often take priority.

🔁 So the corrected statement should be:

> While the estimator using $\frac{1}{n-1}$ is unbiased, the estimators using $\frac{1}{n}$ or $\frac{1}{n+1}$ **uniformly achieve smaller MSE**, making them preferable in terms of accuracy, especially when the goal is minimizing overall estimation error.