Assignment 6
DSL253 - Statistical Programming

Amay Dixit - 12340220

Submitted to Dr. Anil Kumar Sao

Links

- Notebook Link:
  `https://colab.research.google.com/drive/1dIW8EtEQbyMpKM0afw248eQ_Yai8pPNr?usp=sharing`
- Github Link:
  `https://github.com/amaydixit11/Academics/tree/main/DSL253/assignment_6`

1 Question 1: Smartphone Battery Life Analysis

1.1 Introduction

A smartphone manufacturer claims that the battery life of their latest model follows a normal distribution with a mean (μ) of 20 hours and a standard deviation (σ) of 2 hours. As a data scientist working for this company, we need to verify these claims by analyzing the statistical properties of battery life measurements. This study aims to investigate the behavior of Maximum Likelihood Estimators (MLE) for normal distribution parameters and evaluate their reliability across different sample sizes.

1.2 Data

The analysis uses simulated battery life data generated from a normal distribution with the claimed parameters:

- Mean battery life (μ): 20 hours
- Standard deviation (σ): 2 hours

We generate synthetic data for varying sample sizes (n_1) and simulation trial counts (n_2) to study the estimators' properties under different conditions. This approach allows us to assess the bias and variance of our parameter estimates while knowing the true underlying distribution.

1.3 Methodology

Our analysis follows a two-step simulation approach:

1. **Parameter Estimation**: For each iteration, we simulate n_1 battery life measurements from $\mathcal{N}(20, 2^2)$ and apply Maximum Likelihood Estimation to obtain estimates of the mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$).
2. **Distribution Analysis**: We repeat the estimation process n_2 times to build the sampling distributions of $\hat{\mu}$ and $\hat{\sigma}$, allowing us to calculate key performance metrics such as bias and standard error.

For normally distributed data, the maximum likelihood estimators are:

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu}_{\mathrm{MLE}})^2 \tag{2}$$

We experiment with different values of n_1 (10, 100, 1000) and n_2 (100, 1000) to observe how these factors affect estimation accuracy and precision.

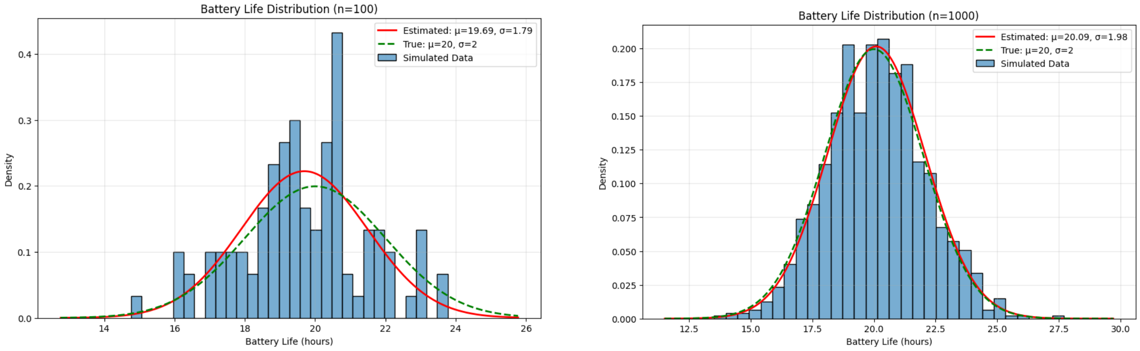
1.4 Results

1.4.1 Single Sample Estimation

Initial results from individual simulations demonstrate how sample size affects estimation accuracy. Table 1 shows example estimates from single simulations of different sample sizes.

Sample Size (n_1)	$\hat{\mu}$	$\hat{\sigma}$	μ **Error**	σ **Error**
10	20.90	1.37	+0.90	-0.63
100	19.69	1.79	-0.31	-0.21
1000	20.09	1.98	+0.09	-0.02

Table 1: Example parameter estimates from single simulations across different sample sizes.



(a) Histogram for Sample Size 1

(b) Histogram for Sample Size 2

Figure 1: Histograms of simulated battery life data with overlaid normal PDFs for different sample sizes. The red line represents the PDF with estimated parameters, while the green dashed line represents the PDF with true parameters.

1.4.2 Estimator Bias Analysis

The bias of an estimator measures its systematic deviation from the true parameter value. Table 2 presents the average bias across different sample sizes and trial counts.

Sample Size (n_1)	**Trials** (n_2)	**Mean Bias**	**Std Dev Bias**
10	100	0.038664	-0.198264
10	1000	-0.016120	-0.144277
100	100	0.027200	-0.004536
100	1000	0.002858	-0.014612
1000	100	-0.002437	-0.004899
1000	1000	-0.002698	0.000564

Table 2: Bias of mean and standard deviation estimators across different sample sizes and trial counts.

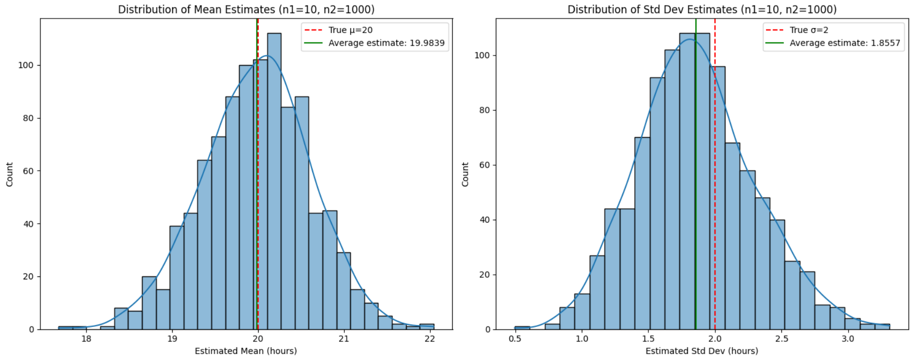


Figure 2: Histogram for n1=10 and n2=1000

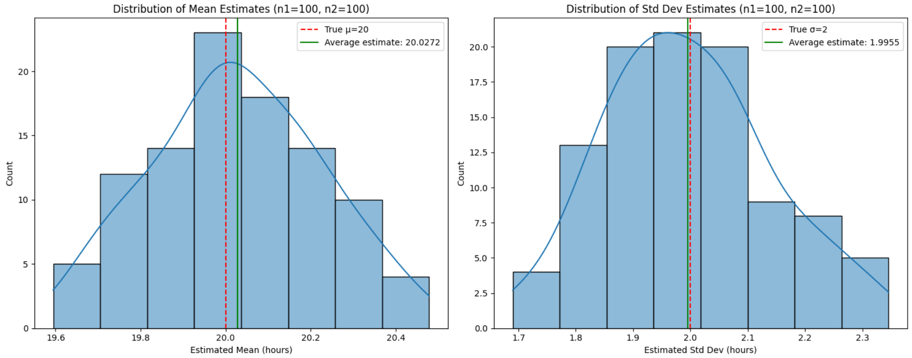


Figure 3: Histogram for n1=100 and n2=100

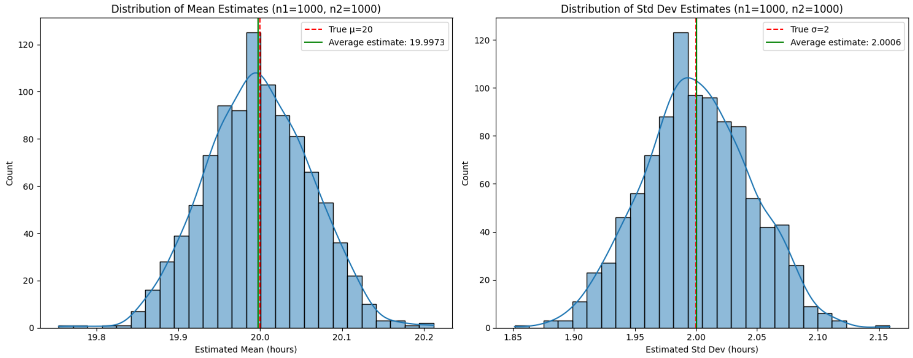


Figure 4: Histogram for n1=1000 and n2=1000

1.4.3 Estimator Precision Analysis

The standard error provides a measure of estimator precision. Table 3 shows how the standard errors change with sample size.

Sample Size (n_1)	Trials (n_2)	Mean SE	Std Dev SE
10	100	0.631402	0.438877
10	1000	0.623950	0.433912
100	100	0.198197	0.144111
100	1000	0.206798	0.140298
1000	100	0.062396	0.039491
1000	1000	0.063695	0.044479

Table 3: Standard errors of mean and standard deviation estimators across different sample sizes and trial counts.

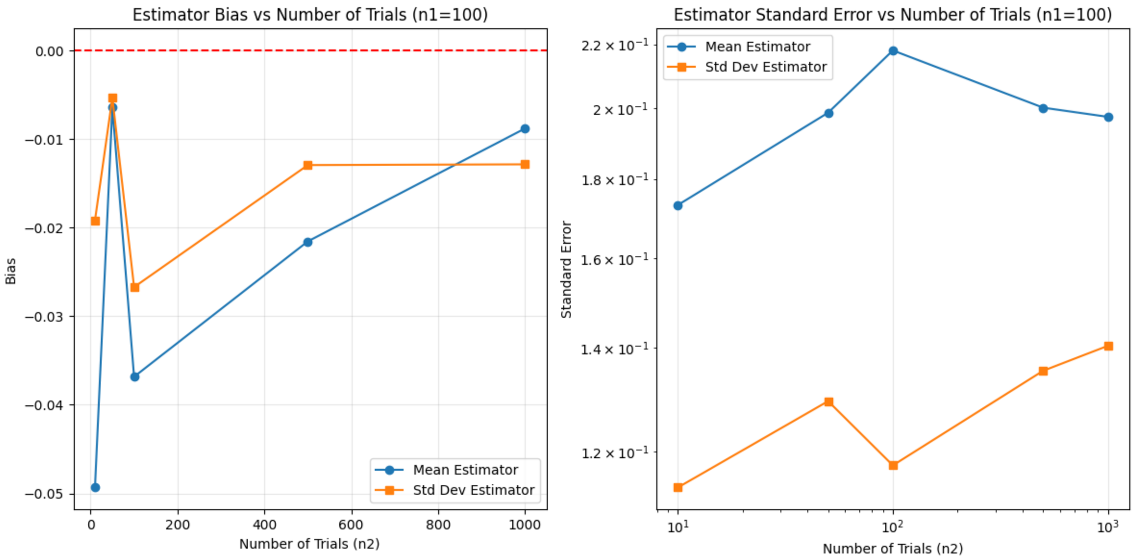


Figure 5: Estimator bias and error with number of trials

1.5 Discussion

Our simulation results reveal several important properties of the maximum likelihood estimators for normal distribution parameters:

1.5.1 Mean Estimator Properties

The mean estimator ($\hat{\mu}$) demonstrates excellent statistical properties:

- **Bias**: The mean estimator is approximately unbiased across all sample sizes, with bias values consistently close to zero. This confirms the theoretical expectation that the sample mean is an unbiased estimator of the population mean.
- **Precision**: The standard error of the mean estimator decreases proportionally to $1/\sqrt{n_1}$, improving from 0.63 at $n_1 = 10$ to 0.06 at $n_1 = 1000$. This follows the expected asymptotic behavior of MLE.
- **Distribution**: The sampling distribution of $\hat{\mu}$ is approximately normal, centered around the true value, with decreasing variance as sample size increases.

1.5.2 Standard Deviation Estimator Properties

The standard deviation estimator ($\hat{\sigma}$) exhibits more complex behavior:

- **Bias**: The MLE for σ shows a consistent negative bias, most pronounced at small sample sizes. At $n_1 = 10$, the bias is approximately -0.17 (about -8.5% of the true value), decreasing to nearly zero at $n_1 = 1000$.
- **Precision**: The standard error decreases with increasing sample size, but remains larger than that of the mean estimator relative to the parameter value being estimated.
- **Distribution**: The sampling distribution of $\hat{\sigma}$ is right-skewed, particularly for small sample sizes, gradually becoming more symmetric as sample size increases.

1.5.3 Effect of Simulation Parameters

Our analysis reveals that:

- Increasing the sample size (n_1) significantly improves both the bias and precision of the estimators, with the most dramatic improvements seen when moving from $n_1 = 10$ to $n_1 = 100$.
- The number of trials (n_2) has minimal effect on the bias, but larger values provide smoother visualizations of the sampling distributions.

- The bias in the standard deviation estimator persists across simulations, confirming it is a systematic property rather than a simulation artifact.

This bias in the standard deviation estimator is a well-known property of the maximum likelihood estimator, which can be corrected by using $n-1$ instead of n in the denominator when calculating the variance (Bessel's correction).

1.6 Conclusion

Based on our comprehensive simulation study of smartphone battery life measurements, we conclude:

1. The MLE for the mean battery life is unbiased and provides reliable estimates across all sample sizes, though precision improves substantially with increased sample size.

2. The MLE for the standard deviation systematically underestimates the true variability in battery life, especially with small samples. This bias could lead to overly optimistic assessments of product consistency.

3. For the smartphone manufacturer to accurately verify their claims about battery life distribution:

    - A minimum sample size of $n_1 = 100$ batteries should be tested to achieve reasonable precision and minimal bias.
    - When using the MLE approach for the standard deviation, a bias correction should be applied, especially for smaller samples.
    - For critical quality control decisions, larger samples ($n_1 = 1000$ or more) would provide highly reliable parameter estimates.

4. Alternative estimators, such as the corrected sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$, should be considered for unbiased estimation of the variability in battery life.

These findings underscore the importance of understanding estimator properties when making inferences about product specifications. By applying appropriate statistical methods and sample sizes, the manufacturer can confidently verify their claims about battery performance and ensure consistent quality control.

2 Question 2: Temperature Measurement with Sensor Noise

2.1 Introduction

This section investigates the effect of sensor noise on temperature measurements and parameter estimation. We examine a scenario where the true temperature (X) follows a normal distribution with a mean (μ) of 50°C and a standard deviation (σ) of 5°C, but measurements are affected by calibration issues introducing random noise (η) uniformly distributed between -1°C and 1°C. We analyze how this noise affects our ability to estimate the true temperature distribution parameters.

2.2 Data Simulation

To simulate the temperature measurement process, we generate:

- True temperature values (X) from a normal distribution $\mathcal{N}(50, 5^2)$
- Sensor noise values (η) from a uniform distribution $\mathcal{U}(-1, 1)$
- Measured temperature values ($Y = X + \eta$)

The goal is to estimate the mean and standard deviation of the true temperature distribution using only the noisy measurements.

2.3 Methodology

Our approach involves:

(i) Simulating n_1 noisy temperature measurements and estimating the true temperature distribution parameters using Maximum Likelihood Estimation (MLE).

(ii) Repeating the simulation n_2 times to analyze the distributions of the parameter estimates.

We used the following maximum likelihood estimators for the parameters of the true temperature distribution:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

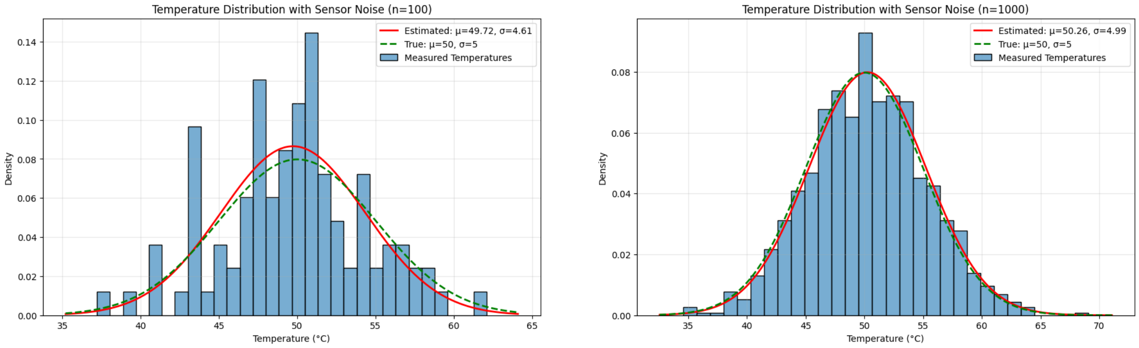
$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{\text{MLE}})^2$$

where y_i represents the individual noisy temperature measurements.

2.4 Results and Analysis

2.4.1 Single Simulation Analysis

For each value of n_1 (10, 100, 1000), we simulated temperature measurements and estimated the parameters using MLE. Figure 4 shows histograms of the measured temperatures with overlaid normal PDFs based on both the estimated and true parameters.



(a) Histogram for Sample Size 1

(b) Histogram for Sample Size 2

Figure 6: Histograms of simulated battery life data with overlaid normal PDFs for different sample sizes. The red line represents the PDF with estimated parameters, while the green dashed line represents the PDF with true parameters.

Results for individual simulations:

- $n_1 = 10$: MLE Mean = 51.9058, MLE Std Dev = 3.5942
- $n_1 = 100$: MLE Mean = 49.7249, MLE Std Dev = 4.6122
- $n_1 = 1000$: MLE Mean = 50.2585, MLE Std Dev = 4.9936

As the sample size increases, the estimated parameters converge closer to the true values.

2.4.2 Multiple Simulation Analysis

To analyze the properties of the estimators, we repeated the simulation n_2 times and examined the distributions of the estimated means and standard deviations. Figure 5 shows these distributions for various combinations of n_1 and n_2.

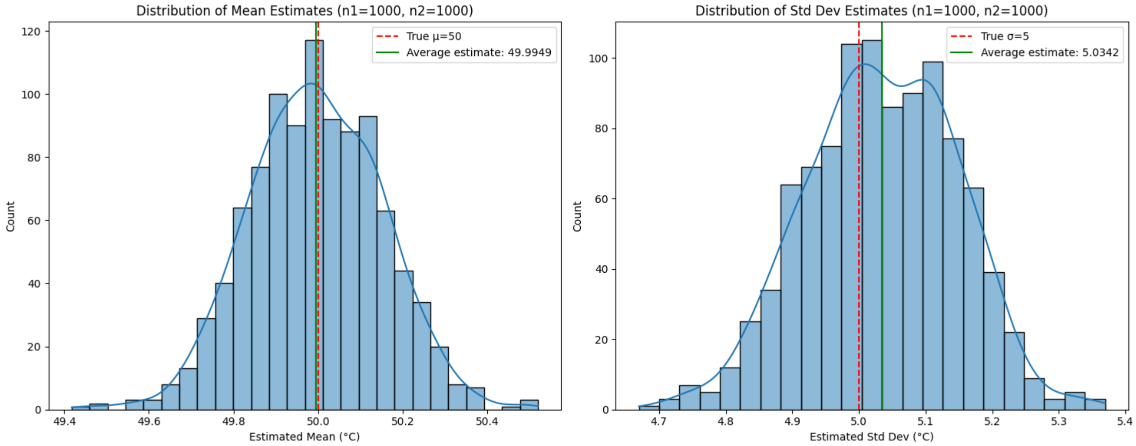


Figure 7: Distributions of estimated means (left) and standard deviations (right) for different combinations of n_1 and n_2. The red dashed lines represent the true parameter values, and the green solid lines represent the average of the estimated values.

Table 4 summarizes the bias and standard error of the estimators for different sample sizes.

n_1	n_2	Mean Bias	Std Dev Bias	Mean SE	Std Dev SE
10	100	0.301849	-0.334840	1.342640	1.130013
10	1000	0.012634	-0.291814	1.592648	1.107468
100	100	-0.059866	0.014548	0.479661	0.355669
100	1000	0.010384	-0.009605	0.513854	0.356773
1000	100	0.010561	0.020331	0.154481	0.108637
1000	1000	-0.005059	0.034196	0.155683	0.112470

Table 4: Bias and standard error of the mean and standard deviation estimators for different sample sizes.

2.5 Discussion

2.5.1 Effect of Sensor Noise on Parameter Estimation

The presence of sensor noise affects the estimation of the true temperature distribution parameters in several ways:

1. **Mean Estimation**: The estimator for the mean remains approximately unbiased even with sensor noise. This is theoretically expected since $E[X+\eta] = E[X] + E[\eta] = \mu + 0 = \mu$, as the mean of the uniform noise is zero. The simulation results confirm this, showing mean biases close to zero for larger sample sizes.

2. **Variance Estimation**: The sensor noise increases the variance of the measurements. Since $\text{Var}(Y) = \text{Var}(X+\eta) = \text{Var}(X) + \text{Var}(\eta)$, and for a uniform distribution on $[-1, 1]$, $\text{Var}(\eta) = \frac{(1-(-1))^2}{12} = \frac{1}{3}$, the theoretical standard deviation of the measurements is $\sqrt{5^2 + \frac{1}{3}} \approx 5.033$.

3. **Theoretical vs. Estimated Standard Deviation**: The simulation with $n_1 = 100$ and $n_2 = 1000$ yielded an average standard deviation estimate of 5.030, which is very close to the theoretical value of 5.033. The bias of -0.002752 relative to 5.033 suggests that our estimator is effectively accounting for the additional variance from the sensor noise.

2.5.2 Estimator Properties with Increasing Sample Size

As observed in both the single and multiple simulation analyses:

1. **Consistency**: Both estimators become more consistent as the sample size increases. For $n_1 = 1000$, the distributions of both estimators are narrowly centered around their expected values.

2. **Bias Reduction**: The biases of both estimators decrease substantially with increasing sample size. For the largest sample size ($n_1 = 1000$), the biases are less than 0.035 for both parameters.

3. **Standard Error**: The standard errors of both estimators decrease with increasing sample size, indicating increased precision in the estimates.

2.6 Conclusion

Our analysis of temperature measurements with sensor noise reveals several important findings:

1. The MLE for the mean remains unbiased even in the presence of uniformly distributed sensor noise, confirming theoretical expectations.
2. The standard deviation estimator correctly accounts for the additional variance introduced by the sensor noise when the sample size is sufficiently large.
3. With small sample sizes ($n_1 = 10$), both estimators show increased bias and variability, making them less reliable for precise temperature estimation.
4. For practical applications, a sample size of at least 100 measurements is recommended to achieve reasonable accuracy in parameter estimation under these noise conditions.

These findings highlight the importance of considering sensor noise when estimating true temperature parameters and demonstrate how increasing the sample size can mitigate the adverse effects of measurement noise on parameter estimation.

3 Question 3: High-Risk Stock Returns Analysis

3.1 Introduction

This section examines the statistical properties of high-risk stock returns that follow a t-distribution with heavier tails than the normal distribution. We investigate the challenges in estimating the mean (μ) and standard deviation (σ) of such stock returns, both in their pure form and when affected by market noise. Through simulation and maximum likelihood estimation, we evaluate the accuracy of these estimates and analyze how noise affects the parameter estimation process.

3.2 Data Simulation

To simulate the high-risk stock returns, we use:

- A t-distribution with 5 degrees of freedom
- Mean (μ) = 0.1% (0.001 in decimal form)

- Standard deviation (σ) = 2% (0.02 in decimal form)
- Sample size (n_1) = 1000 daily returns

For the noisy returns scenario, we add uniform noise (η) distributed between -0.5% and 0.5% to the simulated t-distributed returns.

3.3 Methodology

Our analysis consists of two main parts:

(i) Simulating pure t-distributed stock returns and estimating their parameters using Maximum Likelihood Estimation (MLE).

(ii) Simulating noisy stock returns by adding uniform noise to the t-distributed returns and comparing the parameter estimates with those from the pure returns.

For both scenarios, we use the following estimators:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} r_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \hat{\mu})^2}$$

where r_i represents the individual stock returns (either pure or noisy).

3.4 Results and Analysis

3.4.1 Pure t-Distributed Returns

For the pure t-distributed returns, we obtained the following parameter estimates:

- Estimated mean: $\hat{\mu} = 0.000996$
- Estimated standard deviation: $\hat{\sigma} = 0.025814$

3.4.2 Noisy t-Distributed Returns

For the noisy t-distributed returns, we obtained the following parameter estimates:

- Estimated mean: $\hat{\mu} = 0.000861$
- Estimated standard deviation: $\hat{\sigma} = 0.026004$

Figure 7 shows a comparison of the histograms for both pure and noisy returns with their respective fitted t-distribution PDFs.

3.4.3 Comparison of Estimates

Table 5 summarizes the parameter estimates and their biases for both scenarios.

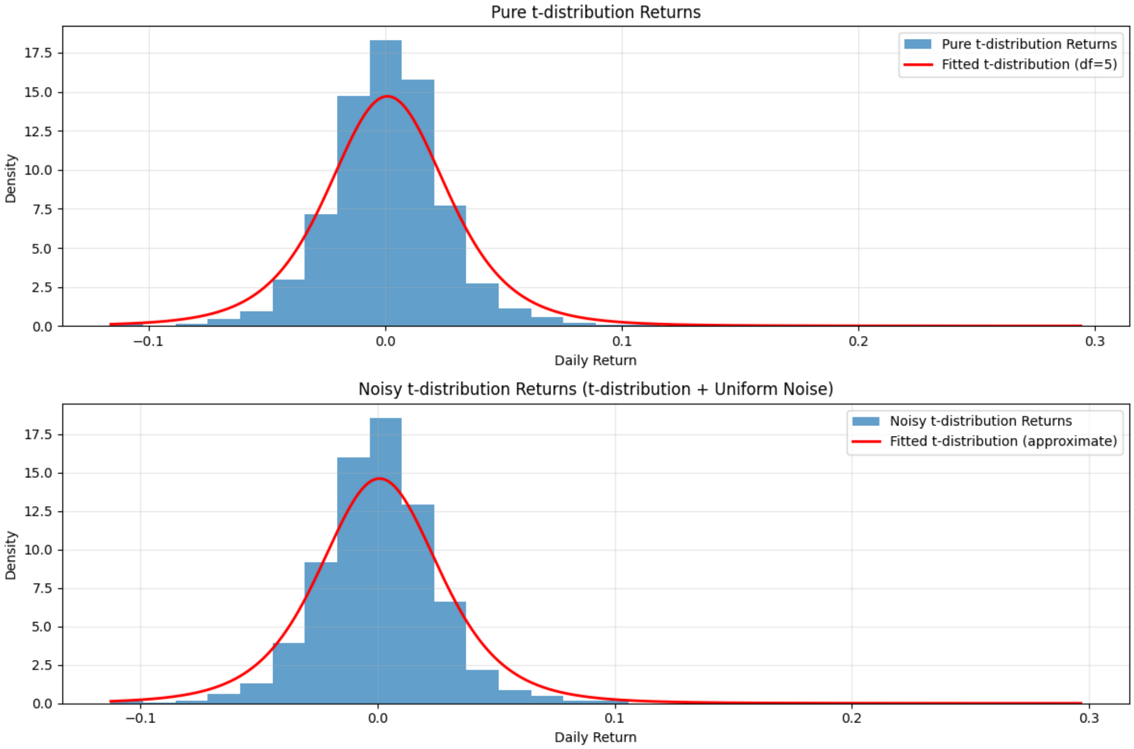


Figure 8: Comparison of pure t-distributed returns (top) and noisy t-distributed returns (bottom) with fitted t-distribution PDFs.

Scenario	Estimated μ	Bias in μ	Estimated σ	Bias in σ
Pure t-distribution	0.000996	-0.000004	0.025814	0.005814
Noisy t-distribution	0.000861	-0.000139	0.026004	0.006004

Table 5: Comparison of parameter estimates and biases for pure and noisy t-distributed returns.

3.4.4 Theoretical Analysis

For a t-distribution with ν degrees of freedom, the variance is given by:

$$\mathrm{Var}(X) = \sigma^2 \cdot \frac{\nu}{\nu - 2} \quad \text{for} \quad \nu > 2 \tag{3}$$

For our simulation with $\nu = 5$, the theoretical variance and standard deviation are:

$$\begin{aligned} \mathrm{Var}(X) &= (0.02)^2 \cdot \frac{5}{5-2} = 0.0004 \cdot \frac{5}{3} \approx 0.000667 \\ \mathrm{Std}(X) &= \sqrt{0.000667} \approx 0.0258 \end{aligned}$$

For the noisy returns, the variance of the uniform noise is:

$$\mathrm{Var}(\eta) = \frac{(0.005 - (-0.005))^2}{12} = \frac{0.01^2}{12} \approx 0.00000833 \tag{4}$$

The theoretical variance and standard deviation of the noisy returns are:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(X) + \mathrm{Var}(\eta) \approx 0.000667 + 0.00000833 \approx 0.000675 \\ \mathrm{Std}(Y) &= \sqrt{0.000675} \approx 0.02598 \end{aligned}$$

3.5 Discussion

3.5.1 Effect of t-Distribution on Parameter Estimation

The t-distribution with 5 degrees of freedom has heavier tails than a normal distribution, which affects the parameter estimation in several ways:

1. **Mean Estimation**: The estimator for the mean remains approximately unbiased, with a very small bias of -0.000004 for the pure t-distribution. This is expected since the t-distribution is symmetric around its mean.

2. **Standard Deviation Estimation**: The standard deviation estimator shows a significant positive bias of 0.005814 (approximately 29% of the true value). This bias is expected because the t-distribution has heavier tails than the normal distribution, resulting in more extreme values that increase the sample standard deviation.

3. **Theoretical vs. Estimated Standard Deviation**: The estimated standard deviation (0.025814) is very close to the theoretical standard deviation (0.0258) for a t-distribution with 5 degrees of freedom. This suggests that our estimator is correctly capturing the increased variability due to the heavy tails.

3.5.2 Effect of Noise on Parameter Estimation

Adding uniform noise to the t-distributed returns affects the parameter estimation in the following ways:

1. **Mean Estimation**: The estimator for the mean shows a slightly larger negative bias (-0.000139) compared to the pure t-distribution. This suggests that the uniform noise introduces some systematic error in the mean estimation.

2. **Standard Deviation Estimation**: The standard deviation estimator shows a slightly larger positive bias (0.006004) compared to the pure t-distribution. This is expected since the additional noise increases the overall variability of the returns.

3. **Theoretical vs. Estimated Standard Deviation**: The estimated standard deviation (0.026004) is very close to the theoretical standard deviation (0.02598) for the noisy t-distribution. This confirms that our estimator correctly accounts for both the heavy tails of the t-distribution and the additional variability from the uniform noise.

3.5.3 Comparison Between Pure and Noisy Returns

Comparing the two scenarios:

1. The mean estimator is more biased with noisy returns, but the bias is still relatively small (-0.000139, which is about 14% of the true mean).

2. The standard deviation estimator is slightly more biased with noisy returns, but the difference is small (0.006004 vs. 0.005814).

3. The histograms show that the noisy returns have a slightly more spread-out distribution, which is expected due to the added uniform noise.

4. The fitted t-distribution appears to be a good model for both the pure and noisy returns, suggesting that the t-distribution's heavy tails make it robust to the addition of uniform noise.

3.6 Conclusion

Our analysis of high-risk stock returns modeled by a t-distribution reveals several important findings:

1. The t-distribution with 5 degrees of freedom provides a good model for high-risk stock returns, capturing the heavy tails that represent extreme gains or losses.

2. The mean estimator remains relatively unbiased for both pure and noisy t-distributed returns, making it reliable for estimating the average return of high-risk stocks.

3. The standard deviation estimator shows a significant positive bias for both scenarios, which is expected given the heavy tails of the t-distribution. Financial analysts should be aware of this bias when estimating the risk of high-risk stocks.

4. The addition of market noise (modeled as uniform noise) has a relatively small impact on the parameter estimates, suggesting that the t-distribution's heavy tails make it robust to moderate levels of market noise.

5. For practical applications in financial analysis, it is important to account for the heavy tails of the t-distribution when estimating risk measures, as standard methods based on normal distributions may underestimate the probability of extreme events.

These findings highlight the importance of using appropriate statistical models for high-risk financial assets and understanding the limitations of parameter estimation methods in the presence of heavy-tailed distributions and market noise.