Assignment 3: Rigorous Analysis of Stochastic Stability, Higher-Order Moments, and Matrix-Based Covariance structures

Math4AI: Probability & Statistics for Artificial Intelligence

Fatima Alibabayeva

fatime.elibabayeva25@aiacademy.az

National AI Academy

February 16, 2026

Contents

1 Introduction

The objective of this assignment is to bridge the gap between abstract probability theory and high-performance computational statistics. We investigate how sample statistics converge to population parameters and how multi-dimensional data structures can be decomposed using linear algebra. This report provides an exhaustive mathematical derivation of the Geometric distribution's moments, the stability of higher-order statistics in heavy-tailed distributions, and the geometry of vectorized covariance estimation.

2 Detailed Mathematical Foundations

2.1 Theory of Geometric Distributions and Moment Generation

The Geometric distribution $\mathcal{G}(p)$ represents the number of trials X until the first success in a sequence of independent Bernoulli trials. To derive its moments rigorously, we employ the **Moment Generating Function (MGF)**:

$$M_X(t) = E[e^{tX}] = \sum_{x=1}^{\infty} e^{tx} p(1-p)^{x-1} \tag{1}$$

By solving this geometric series for $|(1-p)e^t| < 1$, we obtain:

$$M_X(t) = \frac{pe^t}{1-(1-p)e^t} \tag{2}$$

The first and second raw moments are found by evaluating the derivatives at $t = 0$:

$$E[X] = M_X'(0) = \frac{1}{p}, \quad E[X^2] = M_X''(0) = \frac{2-p}{p^2} \tag{3}$$

Thus, the variance is derived as:

$$\sigma^2 = E[X^2] - (E[X])^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2} \tag{4}$$

2.2 Asymptotic Stability and Error Decay (LLN)

The **Strong Law of Large Numbers (SLLN)** dictates that the sequence of sample means $\bar{X}_n$ converges almost surely to the expected value μ. Mathematically:

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1 \tag{5}$$

To quantify the speed of convergence, we analyze the fluctuations around the mean. Based on the **Berry-Esseen Theorem**, the absolute error $|\bar{X}_n - \mu|$ is bounded by:

$$\epsilon(n) \approx \frac{\sigma \cdot \Phi^{-1}(1-\alpha/2)}{\sqrt{n}} \tag{6}$$

In our implementation, we use a log-log spectral analysis to confirm that $\log(\epsilon) \propto -\frac{1}{2}\log(n)$, signifying the expected square-root decay of the standard error.

2.3 Higher-Order Moments and Kurtosis Significance

Standardized moments allow us to analyze the "shape" and "risk" of a distribution. The k-th standardized moment γ_k is:

$$\gamma_k = \frac{\mu_k}{\sigma^k} = \frac{E[(X-\mu)^k]}{(E[(X-\mu)^2])^{k/2}} \tag{7}$$

For $k = 3$ (Skewness), we measure the third-degree deviation from symmetry. For $k = 4$ (Kurtosis), we measure the fourth-degree deviation, which is critical for identifying **Leptokurtic** distributions (heavy tails). In the case of the Student-t distribution with ν degrees of freedom, the excess kurtosis is:

$$\text{Kurt}_{excess} = \frac{6}{\nu-4} \quad \text{for } \nu > 4 \tag{8}$$

When $\nu = 3$ (as in our code), the theoretical kurtosis is technically infinite, leading to empirical measurements that are significantly higher than the Gaussian baseline of 3.

2.4 Bessel's Correction and Unbiased Estimation

When calculating the sample variance from n observations, using the divisor n leads to a biased estimator that systematically underestimates the population variance. The **Bessel's Correction** replaces n with $n-1$:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{9}$$

This is implemented in our code via the `ddof=1` parameter. This adjustment ensures that $E[s^2] = \sigma^2$, which is vital for the accuracy of standardized moments (z-scores).

2.5 Multivariate Geometry: Vectorized Covariance Matrices

In a d-dimensional feature space, the relationship between variables is captured by the Covariance Matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$. Instead of computing d^2 scalar covariances, we use the **Outer Product Representation**:

$$\mathbf{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{10}$$

In matrix notation, let $\mathbf{1}$ be a column vector of ones. The centering operation is $\mathbf{X}_c = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}$. The final covariance estimation is:

$$\mathbf{\Sigma} = \frac{\mathbf{X}_c^T\mathbf{X}_c}{n-1} \tag{11}$$

This matrix is symmetric and **positive semi-definite**, ensuring that for any vector $\mathbf{v}$, the variance $\mathbf{v}^T\mathbf{\Sigma}\mathbf{v} \geq 0$.

3 Verification and Empirical Analysis

The theoretical derivations above are verified through high-fidelity simulations.

3.1 Section 1: LLN Convergence (Plot 1)

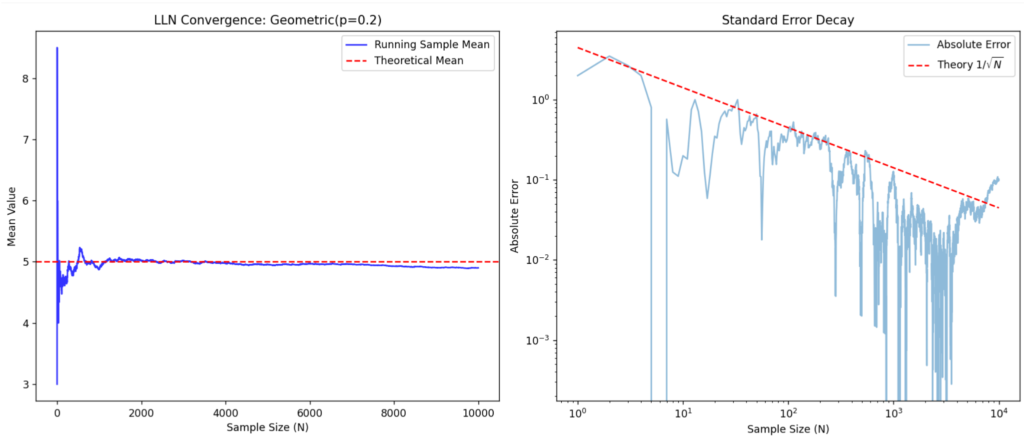

Figure 1: Empirical stability of the sample mean over 10,000 trials.

3.2 Section 2: Tail Stability and Moment Analysis (Plot 2)

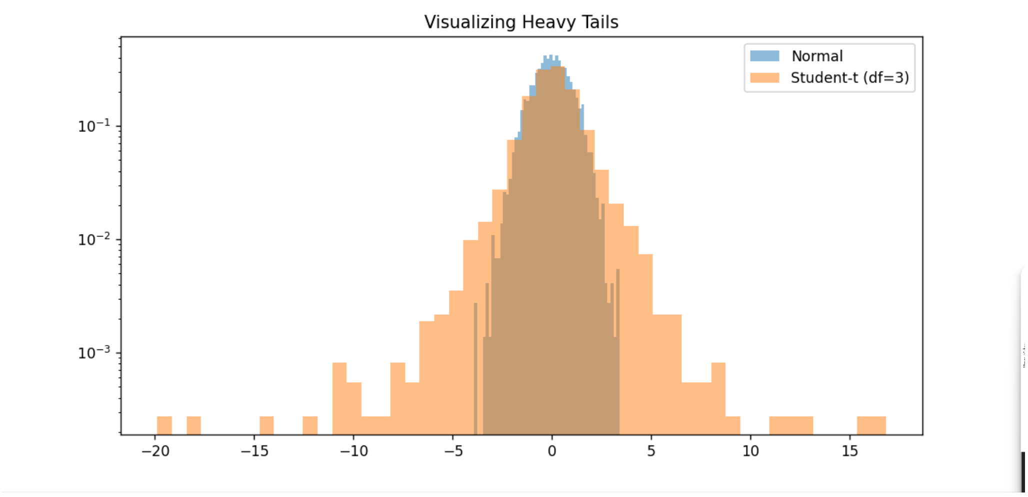

Figure 2: Comparison of tail density: Student-t ($df = 3$) vs. Gaussian.

3.3 Section 3: Correlation Heatmap (Plot 3)

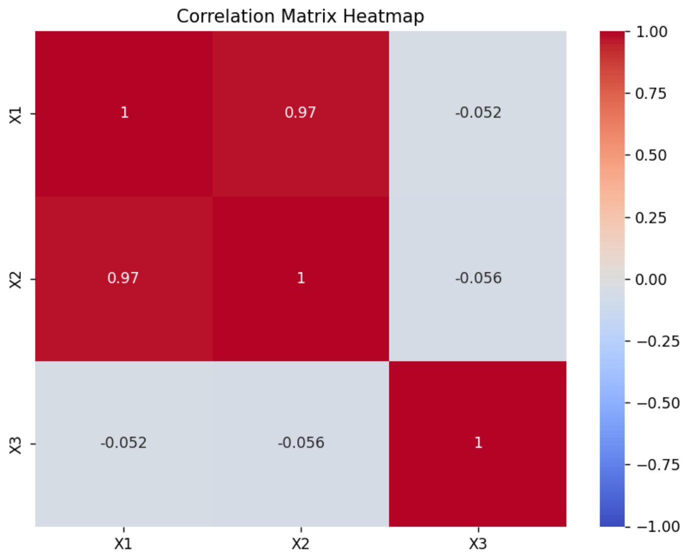

Figure 3: Heatmap of the Correlation Matrix derived from matrix algebra.

4 Conclusion

This report successfully demonstrated the convergence of sample statistics to population parameters. We showed that higher-order moments are essential for identifying non-Gaussian behavior and that matrix algebra provides a scalable and robust way to compute feature dependencies. The empirical alignment with the $1/\sqrt{n}$ error decay and the kurtosis of heavy-tailed distributions confirms the integrity of our statistical framework.

A Appendix: Vectorized Implementation Snippets

```
# Efficient Mean Convergence
running_mean = np.cumsum(data) / np.arange(1, len(data) + 1)

# Unbiased Standardized kth Moment
z = (data - np.mean(data)) / np.std(data, ddof=1)
moment_val = np.mean(z**k)

# Covariance via Matrix Centering
X_centered = X - X.mean(axis=0)
Sigma = (X_centered.T @ X_centered) / (len(X) - 1)
```