Stat 303 Term Project

Rubar Akyıldız-Melih Akay

2026-01-09

4.1 Descriptive Statistics

In this section, we generate a simulated dataset from a Gamma distribution and analyze its descriptive statistics to assess the suitability of the model.

Data Generation and Visualization

We simulate a dataset of size $n = 100$ using a Gamma distribution with shape parameter $k = 3$ and scale parameter $\theta = 2$.

```
# Setting seed for reproducibility
set.seed(303)

# True parameters
true_k <- 3
true_theta <- 2
n <- 100

# Generating the data
data_sample <- rgamma(n, shape = true_k, scale = true_theta)

# calculating sample mean and variance
sample_mean <- mean(data_sample)
sample_var <- var(data_sample)

# Reporting the values
cat("Sample Mean (X_bar):", round(sample_mean, 4), "\n")
```

```
## Sample Mean (X_bar): 6.3157
```

```
cat("Sample Variance (S^2):", round(sample_var, 4), "\n")
```

```
## Sample Variance (S^2): 18.1508
```

```
# Plotting the Histogram
hist(data_sample, breaks = 15, probability = TRUE,
     main = "Histogram",
     xlab = "Data Values", col = "lightblue", border = "white")
```

```
# Adding a density line to look at the shape
lines(density(data_sample), col = "darkblue", lwd = 2)
```

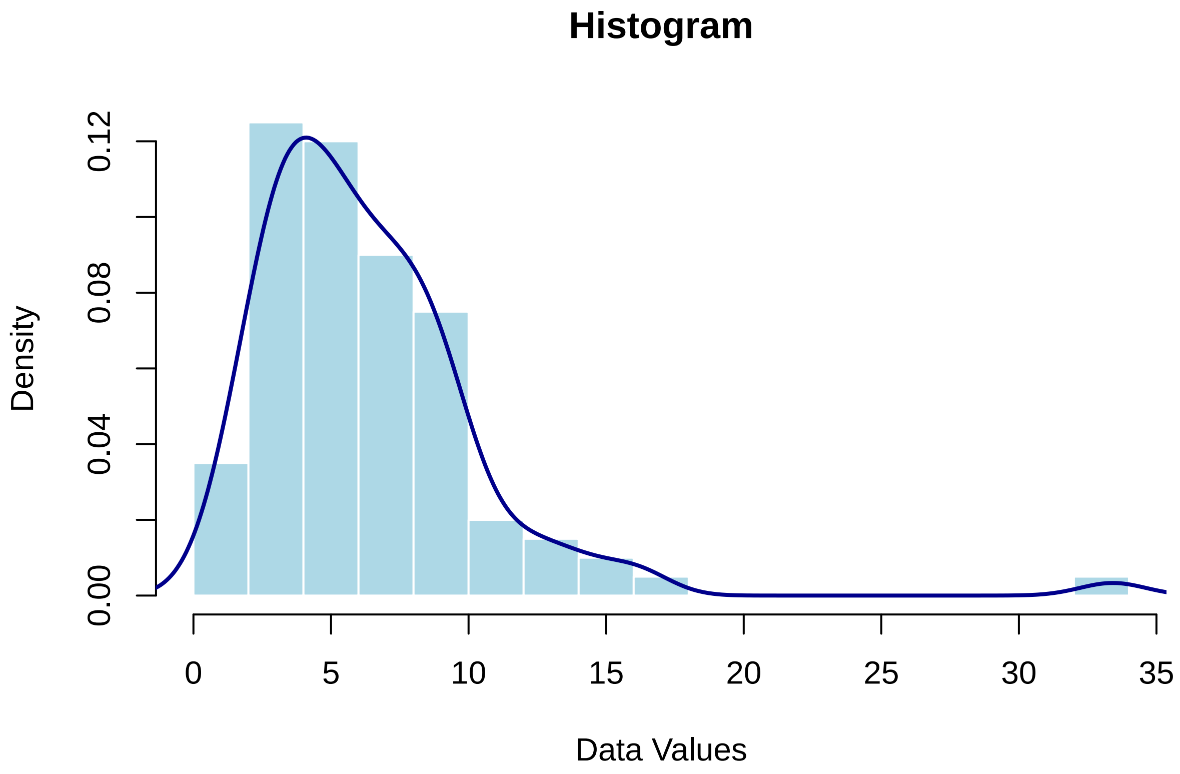


Figure 1: Histogram of Simulated Data

Remarks on the Distribution of Data

The data is **positively skewed** (skewed to the right) rather than symmetric, as can be seen in the histogram. The characteristics of the Gamma family of distributions are in good agreement with this.

The **Gamma model is a suitable option** for this dataset since the variable of interest is continuous and strictly positive ($x > 0$), as well as because of the observed skewness. The simulation's validity is further supported by the descriptive statistics ($\bar{X}$ and S^2).

4.2 Point Estimation

(A) Method of Moments (MoM)

Let $X_1, ..., X_n$ be independent and identically distributed random variables from a Gamma distribution with parameters k (shape) and θ (scale). The probability density function is given by:

$$f(x|k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}, \quad x > 0$$

The first two population moments of the Gamma distribution are:

$$E[X] = k\theta$$
$$Var(X) = k\theta^2$$

The Method of Moments estimators are obtained by equating the population moments to the sample moments ($\overline{X}$ and S^2).

1. From the expectation equation:
$$\mu = k\theta \implies \theta = \frac{\mu}{k}$$
2. Substituting θ into the variance equation:
$$\sigma^2 = k\left(\frac{\mu}{k}\right)^2 = \frac{\mu^2}{k}$$
3. Solving for k:
$$k = \frac{\mu^2}{\sigma^2}$$
4. Solving for θ:
$$\theta = \frac{\sigma^2}{\mu}$$

The MoM estimators are obtained by replacing sample moments ($\overline{X}, S^2$) for the population moments (μ, σ^2):

$$\hat{k}_{MM} = \frac{\overline{X}^2}{S^2}$$

$$\hat{\theta}_{MM} = \frac{S^2}{\overline{X}}$$

(B) Maximum Likelihood Estimation

The likelihood function for the random sample $X_1, ..., X_n$ is:

$$L(k,\theta) = \prod_{i=1}^{n} \frac{1}{\Gamma(k)\theta^k} x_i^{k-1} e^{-x_i/\theta}$$

$$L(k,\theta) = \left[\Gamma(k)\theta^k\right]^{-n} \left(\prod_{i=1}^{n} x_i\right)^{k-1} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right)$$

The log-likelihood func. , $l(k,\theta) = \ln L(k,\theta)$, is:

$$l(k,\theta) = -n\ln\Gamma(k) - nk\ln\theta + (k-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

To find the MLEs, we calculate the partial derivatives with respect to θ and k and set them to zero (score functions).

1. Derivative with respect to θ:

$$\frac{\partial l}{\partial \theta} = -\frac{nk}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0$$

Multiplying by θ^2:

$$-nk\theta + \sum_{i=1}^{n} x_i = 0 \implies \hat{\theta} = \frac{\sum x_i}{nk} = \frac{\overline{X}}{\hat{k}}$$

Thus, we found that:

$$\hat{\theta}_{MLE} = \frac{\overline{X}}{\hat{k}_{MLE}}$$

2. Derivative with respect to k:

$$\frac{\partial l}{\partial k} = -n\frac{\Gamma'(k)}{\Gamma(k)} - n\ln\theta + \sum_{i=1}^{n} \ln x_i = 0$$

Here, $\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$ is the digamma function. Substituting $\hat{\theta} = \frac{\overline{X}}{\hat{k}}$ into this equation:

$$-n\psi(\hat{k}) - n\ln\left(\frac{\overline{X}}{\hat{k}}\right) + \sum_{i=1}^{n} \ln x_i = 0$$

$$-n\psi(\hat{k}) - n(\ln\overline{X} - \ln\hat{k}) + \sum_{i=1}^{n} \ln x_i = 0$$

Dividing by n and rearranging the terms:

$$\ln(\hat{k}) - \psi(\hat{k}) = \ln(\overline{X}) - \frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

Comparison of Solution Methods

Unlike the MoM estimators, the MLE for k ($\hat{k}_{MLE}$) does not have a **closed-form solution**. This is because it is hard to isolate $\hat{k}$ algebraically on one side of the equation due to the non-linear appearance of the parameter k inside both the logarithmic function and the digamma function ($\psi(k)$)$.

Therefore, $\hat{k}_{MLE}$ must be computed by using **numerical optimization methods** to find the root of the derived equation. Once $\hat{k}_{MLE}$ is found numerically, $\hat{\theta}_{MLE}$ can be calculated directly.

4.3 Comparison on Observed Data

The generated MoM and MLE methods are applied to a simulated dataset in this section. To assess and compare the estimates both numerically and visually, we create a random sample from a Gamma distribution with given parameters.

Data Generation

We simulate a dataset of size $n = 100$ with shape parameter $k = 3$ and scale parameter $\theta = 2$.

```r
# Setting seed for reproducibility
set.seed(303)

# True parameters
true_k <- 3
true_theta <- 2
n <- 100

# Generating data from Gamma dist.
data_sample <- rgamma(n, shape = true_k, scale = true_theta)

# Descriptive statistics
cat("Sample Mean:", mean(data_sample), "\n")
```

```
## Sample Mean: 6.315663
```

```r
cat("Sample Variance:", var(data_sample), "\n")
```

```
## Sample Variance: 18.15082
```

```r
# Method of Moments
# k = mean^2 / var, theta = var / mean
k_mom <- (mean(data_sample)^2) / var(data_sample)
theta_mom <- var(data_sample) / mean(data_sample)

#  Maximum Likelihood Estimation
# Equation to solve: log(k) - psi(k) - (log(x_bar) - mean(log(x))) = 0
target_val <- log(mean(data_sample)) - mean(log(data_sample))

mle_eqn <- function(k) {
  log(k) - digamma(k) - target_val
}

# Mle roots
mle_root <- uniroot(mle_eqn, interval = c(0.1, 20), extendInt = "yes")
k_mle <- mle_root$root
theta_mle <- mean(data_sample) / k_mle

# Displaying the results
results_table <- data.frame(
  Method = c("True Values", "MoM", "MLE"),
  Shape_k = c(true_k, k_mom, k_mle),
  Scale_theta = c(true_theta, theta_mom, theta_mle)
)

knitr::kable(results_table, digits = 4, caption = "Comparison of Estimates")
```

Table 1: Comparison of Estimates

Method	Shape_k	Scale_theta
True Values	3.0000	2.0000
MoM	2.1976	2.8739
MLE	2.8685	2.2017

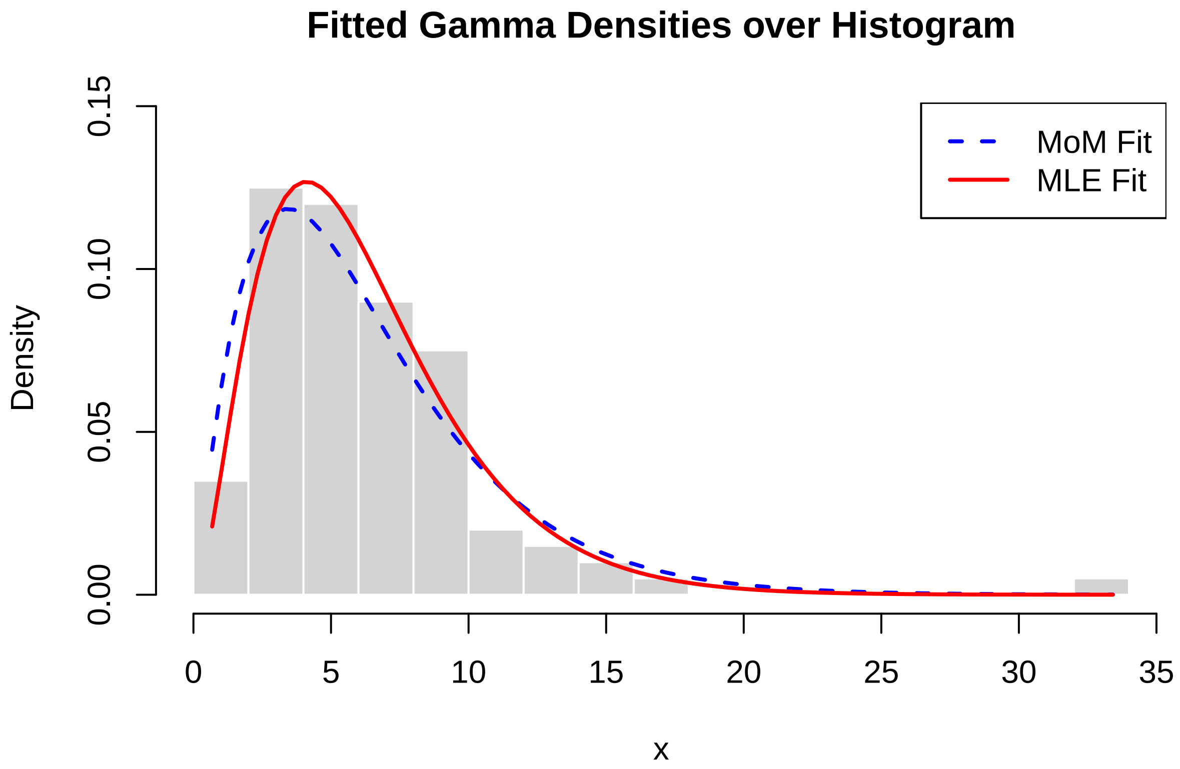
```r
#  Plot Histogram
hist(data_sample, probability = TRUE, breaks = 15,
     main = "Fitted Gamma Densities over Histogram",
     xlab = "x", col = "lightgray", border = "white",
     ylim = c(0, max(density(data_sample)$y) * 1.2))

#  x-axis sequence for curves
x_vals <- seq(min(data_sample), max(data_sample), length.out = 100)

#  Add MoM Curve
lines(x_vals, dgamma(x_vals, shape = k_mom, scale = theta_mom),
      col = "blue", lwd = 2, lty = 2)

#  Add MLE Curve
lines(x_vals, dgamma(x_vals, shape = k_mle, scale = theta_mle),
      col = "red", lwd = 2)

#  Add Legend
legend("topright", legend = c("MoM Fit", "MLE Fit"),
       col = c("blue", "red"), lty = c(2, 1), lwd = 2)
```



Interpreting the Differences

When the fitted densities are compared, the **Maximum Likelihood Estimation (MLE)** obviously fits the observed data better than the Method of Moments (MoM). In terms of values, the shape parameter's MLE estimate ($\hat{k} \approx 2.87$) is significantly closer to the true value ($k = 3$) than the MoM estimate ($\hat{k} \approx 2.20$).

This distinction can be seen visually in the plot: the peak and the right-skewed tail of the histogram are perfectly captured by the **MLE curve (red solid line)**. On the other hand, the MoM curve (blue dashed line) fails to accurately represent the central tendency of the data and underestimates the peak height. This shows how effective MLE is at estimating parameters for Gamma distributions.

5. Simulation Study

In this section, we conduct a Monte Carlo simulation to evaluate the performance of MoM and MLE estimators. We consider two scenarios with different skewness levels:

- **Scenario 1 (High Skewness):** $k = 1, \theta = 2$
- **Scenario 2 (Moderate Skewness):** $k = 5, \theta = 1$

For each scenario, we use sample sizes of $n \in \{20, 50, 100\}$. The number of replications is set to $R = 2000$. We compare the estimators based on Bias, Variance, and Mean Squared Error.

Table 2: Simulation Results: Bias, Variance, and MSE

Scenario	n	Method	Parameter	Bias	Variance	MSE
Scenario 1	20	MoM	k	0.2133	0.2326	0.2781
Scenario 1	20	MLE	k	0.1416	0.1379	0.1579
Scenario 1	20	MoM	theta	-0.0987	0.7558	0.7655
Scenario 1	20	MLE	theta	-0.0877	0.4944	0.5021
Scenario 1	50	MoM	k	0.0902	0.0723	0.0805
Scenario 1	50	MLE	k	0.0476	0.0365	0.0388
Scenario 1	50	MoM	theta	-0.0541	0.3423	0.3452
Scenario 1	50	MLE	theta	-0.0359	0.2032	0.2045
Scenario 1	100	MoM	k	0.0435	0.0406	0.0425
Scenario 1	100	MLE	k	0.0256	0.0180	0.0186
Scenario 1	100	MoM	theta	-0.0092	0.1999	0.2000
Scenario 1	100	MLE	theta	-0.0176	0.0991	0.0994
Scenario 2	20	MoM	k	0.6577	4.7982	5.2307
Scenario 2	20	MLE	k	0.8281	4.5751	5.2609
Scenario 2	20	MoM	theta	0.0010	0.1357	0.1357
Scenario 2	20	MLE	theta	-0.0446	0.1019	0.1039
Scenario 2	50	MoM	k	0.2210	1.3186	1.3674
Scenario 2	50	MLE	k	0.2677	1.1394	1.2111
Scenario 2	50	MoM	theta	0.0018	0.0501	0.0501
Scenario 2	50	MLE	theta	-0.0145	0.0403	0.0406
Scenario 2	100	MoM	k	0.1180	0.6514	0.6653
Scenario 2	100	MLE	k	0.1373	0.5408	0.5596
Scenario 2	100	MoM	theta	0.0014	0.0264	0.0264
Scenario 2	100	MLE	theta	-0.0070	0.0208	0.0208

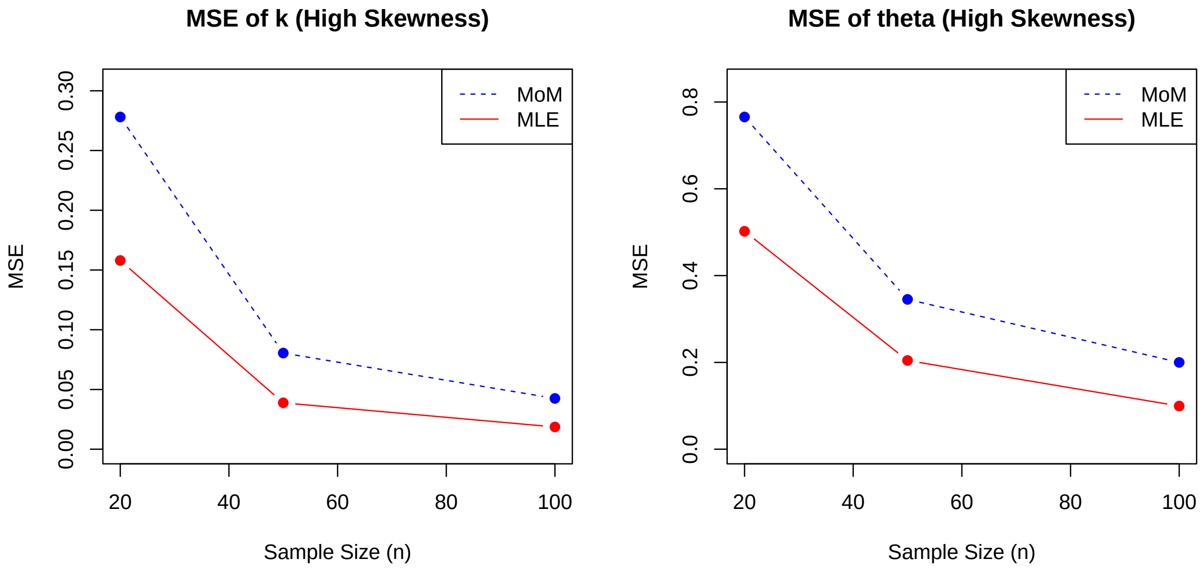


Figure 2: MSE vs Sample Size Comparison

6. Discussion and Conclusions

In this project, we used a Monte Carlo simulation method to evaluate the finite-sample performance of the MoM and MLE for the Gamma distribution parameters. We get the following conclusions based on the MSE plots and the results shown in Table 2:

1. Effects of Sample Size

The consistency property for both estimators is clearly supported by the simulation results. In every case; the bias, variance, and MSE drop as the sample size (n) rises from 20 to 100. This shows that when more data becomes available, both approaches converge to the original parameter values.

2. Comparison (MoM vs. MLE)

In general, the **Maximum Likelihood Estimator** outperformed the Method of Moments.

- The MLE curve (red line) is constantly positioned below the MoM curve (blue dashed line), as can be seen in the MSE plots, indicating a decreased estimation error.
- This advantage is especially obvious in **Scenario 1 (High Skewness)**. For example, for $n = 20$, the shape parameter $\hat{k}_{MLE}$ (0.1579) has a significantly smaller MSE than $\hat{k}_{MoM}$ (0.2781).
- MoM is easier to compute, but because MLE uses the entire likelihood function, it gives more accurate estimates (lower variance).

3. Skewness's Effect

The estimation difficulty was affected by the degree of skewness. MLE performed noticeably better than MoM in Scenario 1 ($k = 1$), when the distribution is substantially skewed. The performance difference between the two approaches decreased in Scenario 2 ($k = 5$, moderate skewness), particularly for small sample sizes ($n = 20$), although MLE continued to be the better estimator asymptotically ($n = 100$).

Practical Recommendation

The **Maximum Likelihood Estimation** approach should be considered for real-world Gamma dist. data, even if the Method of Moments provides simple-to-calculate closed-form answers. MLE offers more precise and reliable estimations, particularly for skewed data and bigger sample sizes, even if it needs numerical optimization because $\hat{k}_{MLE}$ has no closed form.