Lab 3: Statistical Inference

Code

Team 12

2025-03-02

Overview

In this lab we investigate the properties of the mean of 40 exponential distributions. For an exponential distribution with rate parameter $\lambda = 0.2$:
- The theoretical mean is $\mu = 1/\lambda = 5$.
- The variance for a single exponential is $\sigma^2 = 1/\lambda^2 = 25$.

When taking the mean of 40 independent exponential random variables, by the Central Limit Theorem, the sample mean is approximately normally distributed with:
- Mean: 5 (same as the original mean)
- Variance: $25/40 = 0.625$.

Motivating Example: Uniform Distribution

Predefined parameters:

Hide

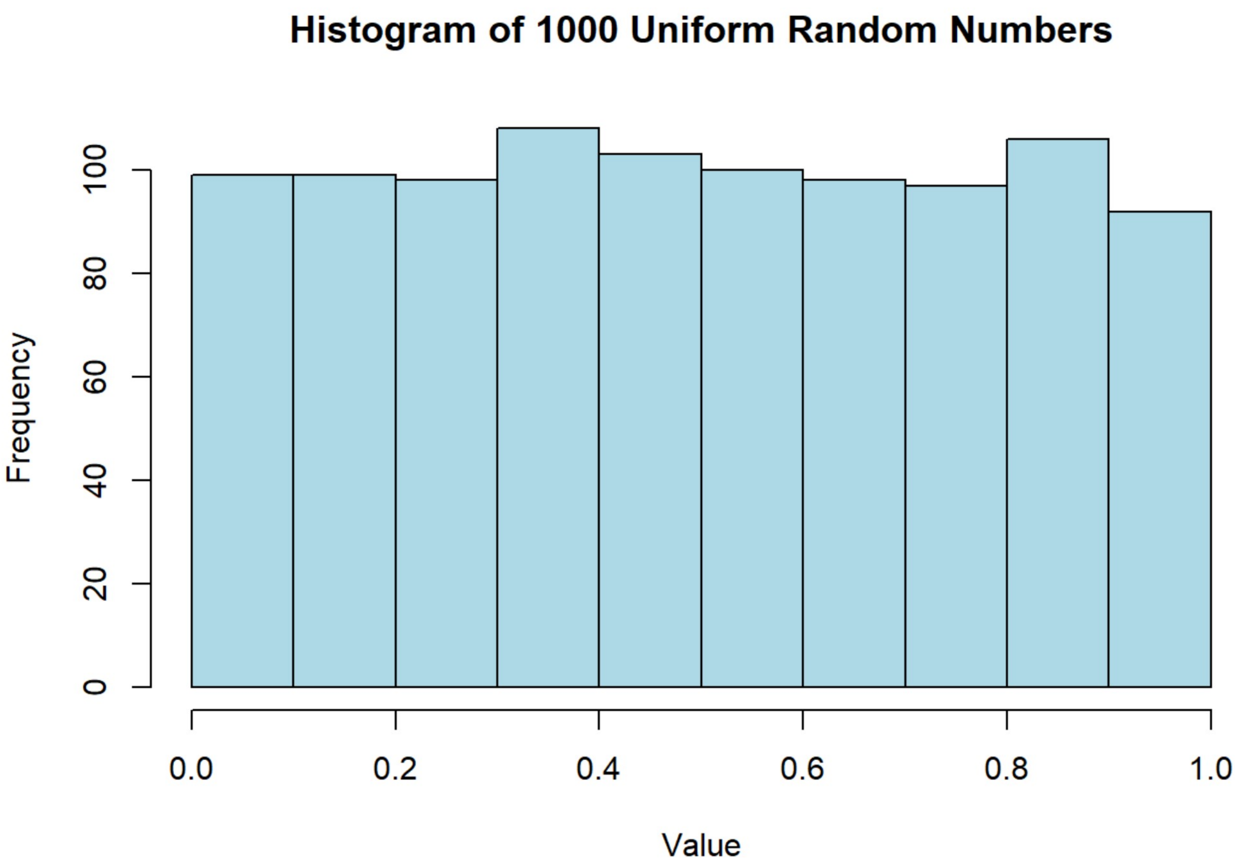
```
lambda <- 0.2
n <- 40
num_simulations <- 1000
```

We start by looking at the uniform distribution. First, we plot a histogram of 1000 random uniform numbers.

Hide

```
set.seed(123)
# Histogram of 1000 random uniform numbers
hist(runif(num_simulations),
     main = "Histogram of 1000 Uniform Random Numbers",
     xlab = "Value",
     col = "lightblue")
```

5 min

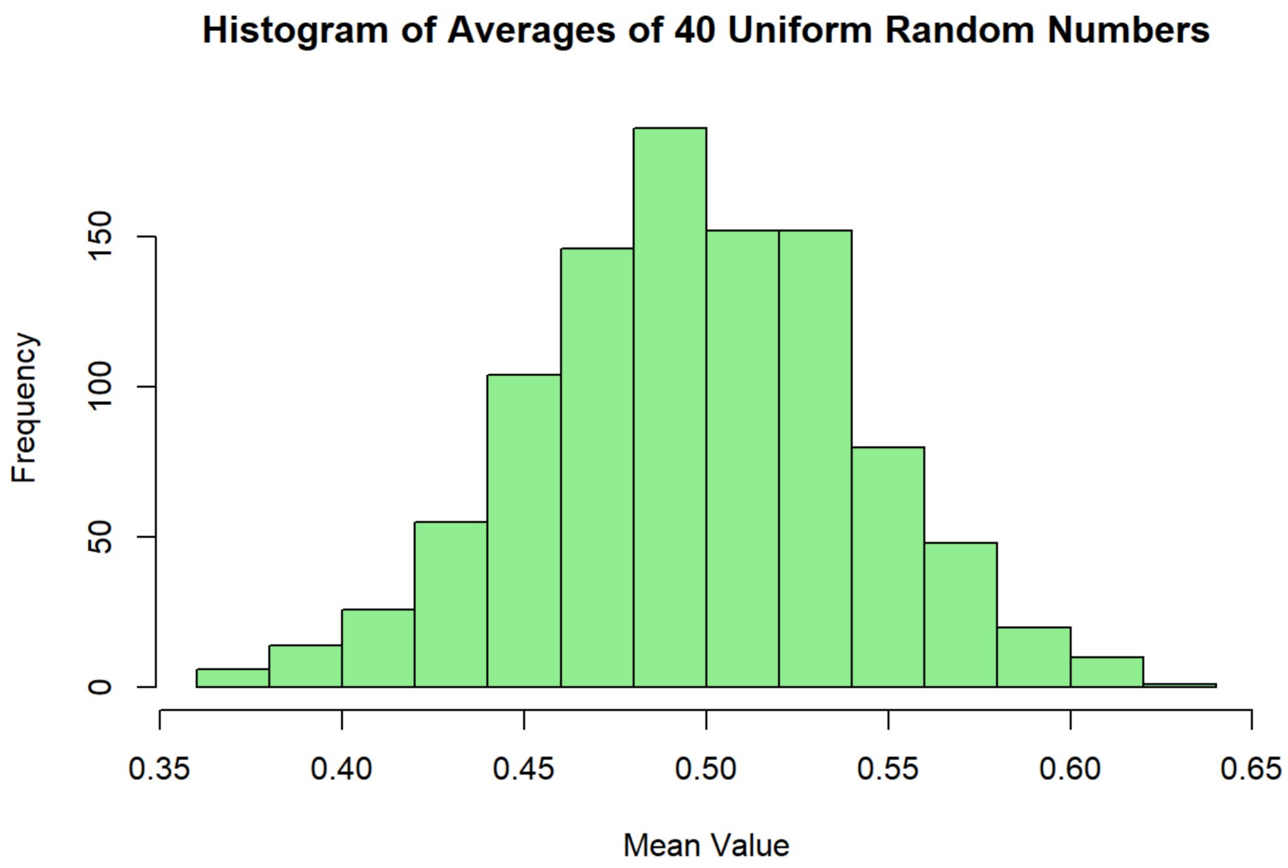


The first histogram shows the flat distribution of 1000 uniform random numbers.

Hide

```
set.seed(123)
# Compute 1000 averages, each from 40 uniform random numbers
uniform_means <- replicate(num_simulations, mean(runif(n)))

hist(uniform_means,
     main = "Histogram of Averages of 40 Uniform Random Numbers",
     xlab = "Mean Value",
     col = "lightgreen")
```

5 min



The second histogram - being a distribution of values which are averages of 40 samples taken from uniform distribution â€”shows a bell-shaped curve, illustrating the Central Limit Theorem in action. Central limit theorem is applicable here because avaerage is a sum divided by some constant(specifically the amount of samples over which the average is taken, 40 in this case). This means that the resulting distribution will be normal, according to the Central Limit Theorem, but the mean will be scaled down by 40 compared to the distribution which is the result of CLT.

> Central Limit Theorem: The distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal (bell-shaped) distribution, regardless of the shape of the original distribution.

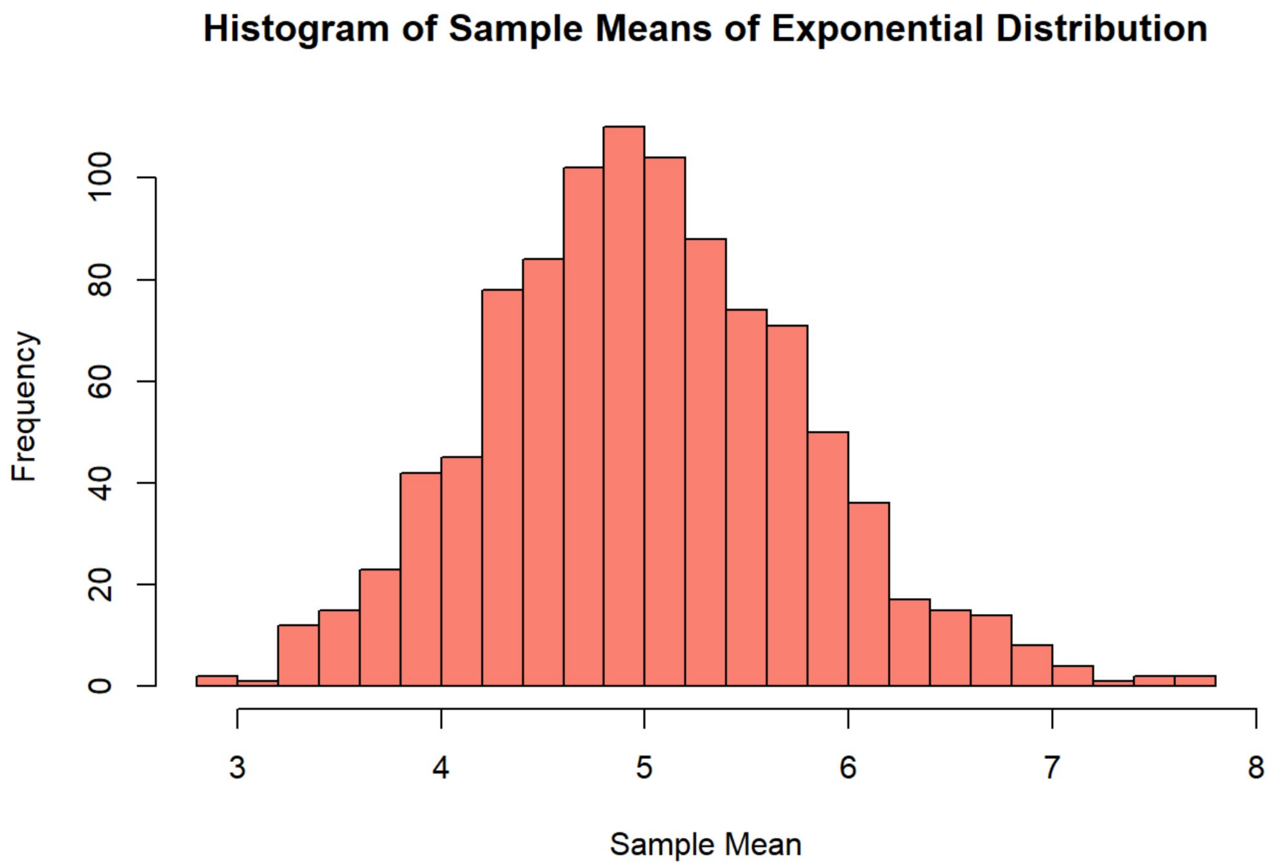
Simulation: Exponential Distribution

Now, we simulate the exponential distribution with $\lambda = 0.2$. In each of 1000 simulations, we take 40 random draws from an exponential distribution and compute their mean. We then compare the sample mean and variance to the theoretical values.

Hide

5 min

```
# Set seed for reproducibility
set.seed(123)

# Theoretical values
theoretical_mean <- 1 / lambda
theoretical_variance <- (1 / (lambda^2)) / n

# Perform the simulation: for each simulation, compute the mean of 40 exponential samples
exp_means <- replicate(num_simulations, mean(rexp(n, rate = lambda)))

# Calculate sample statistics
sample_mean <- mean(exp_means)
sample_variance <- var(exp_means)

# Plot the histogram of sample means
hist(exp_means,
     main = "Histogram of Sample Means of Exponential Distribution",
     xlab = "Sample Mean",
     col = "salmon",
     breaks = 20)
```



Hide

```
# Print the theoretical and sample statistics
cat("Theoretical Mean:", theoretical_mean, "\n")
```

5 min

```
Theoretical Mean: 5
```

Hide

```
cat("Sample Mean:", sample_mean, "\n\n")
```

```
Sample Mean: 5.011911
```

Hide

```
cat("Theoretical Variance:", theoretical_variance, "\n")
```

```
Theoretical Variance: 0.625
```

Hide

```
cat("Sample Variance:", sample_variance, "\n")
```

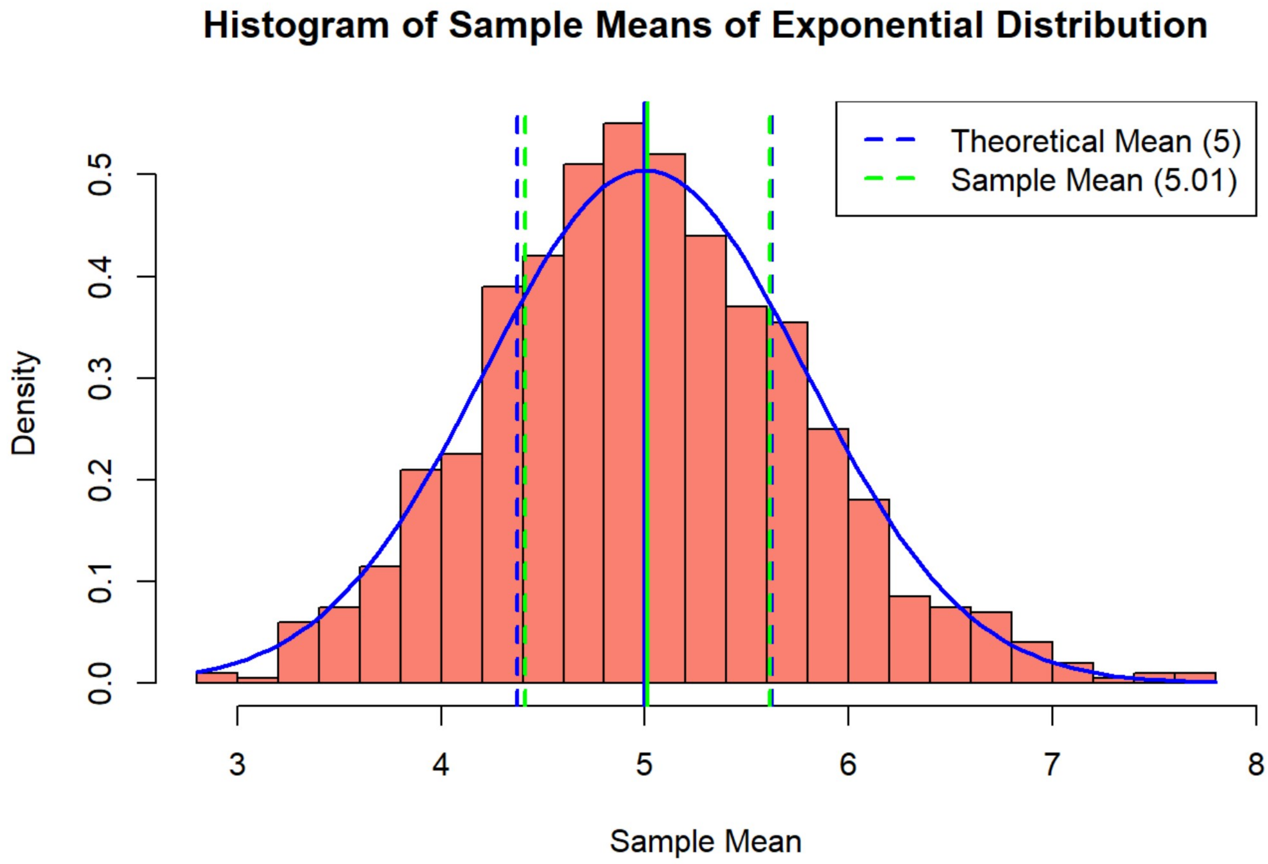
```
Sample Variance: 0.6004928
```

The histogram above is not normalized, let's normalize it and look at how it compares to the theoretical normal distribution.

Hide

```
set.seed(123)
hist(exp_means, probability=TRUE,
     main = "Histogram of Sample Means of Exponential Distribution",
     xlab = "Sample Mean",
     col = "salmon",
     breaks = 20)

abline(v = theoretical_mean, col = "blue", lwd = 2, lty = 1)
abline(v = sample_mean, col = "green", lwd = 2, lty = 1)

abline(v = theoretical_mean + theoretical_variance, col = "blue", lwd = 2, lty = 2)
abline(v = theoretical_mean - theoretical_variance, col = "blue", lwd = 2, lty = 2)
abline(v = sample_mean + sample_variance, col = "green", lwd = 2, lty = 2)
abline(v = sample_mean - sample_variance, col = "green", lwd = 2, lty = 2)

# Overlay the theoretical normal density curve
curve(dnorm(x, mean = theoretical_mean, sd = sqrt(theoretical_variance)),
      add = TRUE, col = "blue", lwd = 2)

legend("topright", legend = c("Theoretical Mean (5)",
                              paste("Sample Mean (", round(sample_mean,2), ")", sep="")),
       col = c("blue", "green"), lty = 2, lwd = 2)
```

5 min



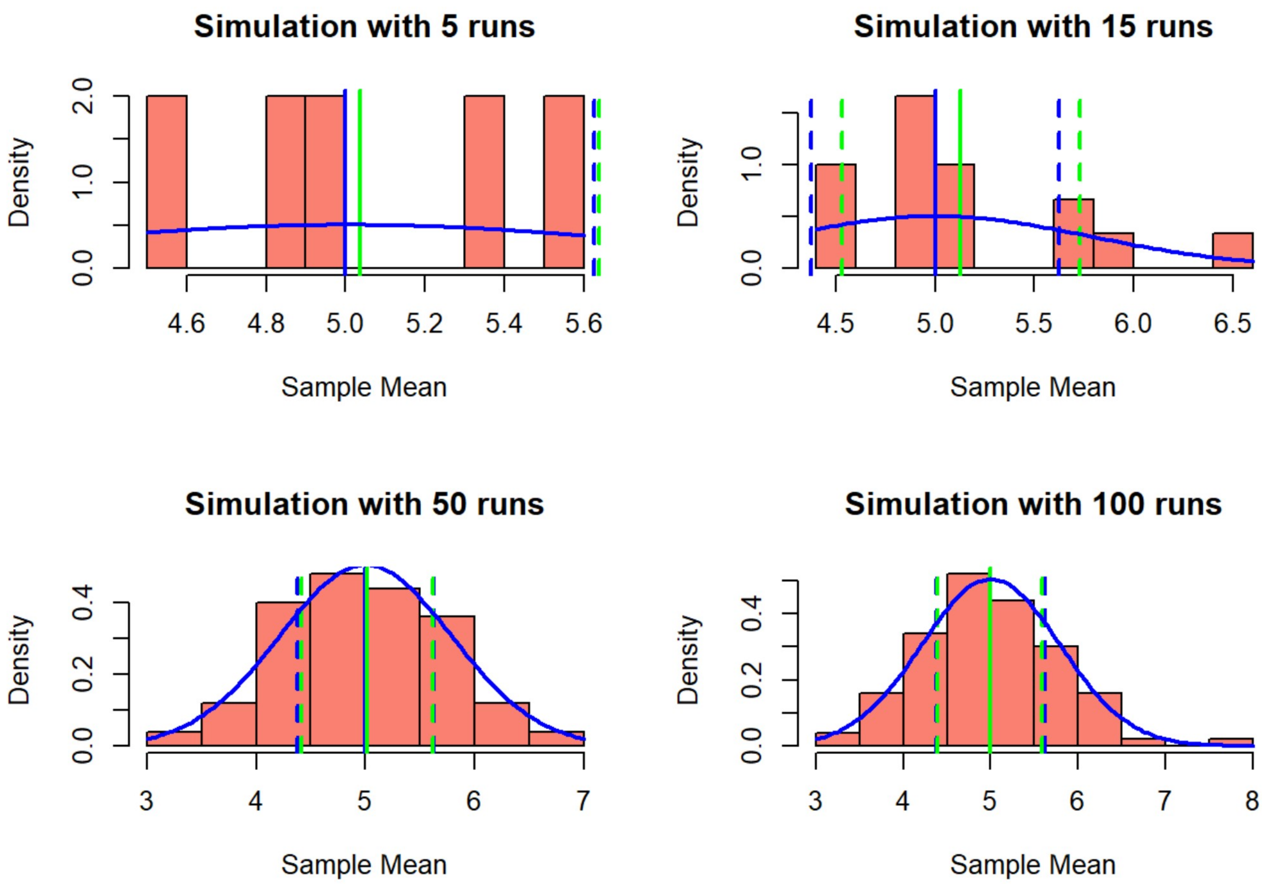
Now we can see that both theoretical and normal distributions are highly similar, the sample mean is very close to the theoretical mean, and the sample variance is very close to the theoretical variance. The slight difference is due to the finite number of simulations we performed. To demonstrate the effect of simulation count on the accuracy of the sample mean and variance, we will run the simulation with different numbers of simulations and plot the results.

Hide

5 min

```
set.seed(123)
# Define simulation sizes
simulation_sizes <- c(5, 15, 50, 100)

par(mfrow = c(2, 2))

for(n_sim in simulation_sizes) {
  exp_means_sim <- replicate(n_sim, mean(rexp(n, rate = lambda)))

  hist(exp_means_sim, probability = TRUE,
       main = paste("Simulation with", n_sim, "runs"),
       xlab = "Sample Mean",
       col = "salmon",
       breaks = 10)

  current_sample_mean <- mean(exp_means_sim)
  # Add vertical line for theoretical mean
  abline(v = theoretical_mean, col = "blue", lwd = 2, lty = 1)
  abline(v = current_sample_mean, col = "green", lwd = 2, lty = 1)

  abline(v = theoretical_mean + theoretical_variance, col = "blue", lwd = 2, lty = 2)
  abline(v = theoretical_mean - theoretical_variance, col = "blue", lwd = 2, lty = 2)
  abline(v = current_sample_mean + sample_variance, col = "green", lwd = 2, lty = 2)
  abline(v = current_sample_mean - sample_variance, col = "green", lwd = 2, lty = 2)

  # Overlay the theoretical normal density curve
  curve(dnorm(x, mean = theoretical_mean, sd = sqrt(theoretical_variance)),
        add = TRUE, col = "blue", lwd = 2)

}
```

5 min



Hide

```
# Reset plotting layout to default
par(mfrow = c(1, 1))
```

Conclusion

In this lab, we have demonstrated the Central Limit Theorem by simulating the mean of 40 exponential distributions. We have shown that the sample mean of the exponential distribution is normally distributed with a mean of 5 and a variance of 0.625. By running the simulation with different numbers of simulations, we have shown that the sample mean and variance converge to the theoretical values as the number of simulations increases.

5 min