Statistical Inference: Exponential Distribution Simulation

Daniel R. Klemfuss

2021-01-03

Table of Contents

1. Overview:

The objective of the simulation exercise is to simulate an exponential distribution in R, and compare the corresponding statistics with the Central Limit Theorem.

2. Simulation Exercise: Exponential Distribution

2.1 Theoretical Exponential Distribution

To better understand the simulation results, start by exploring the theoretical exponential distribution. This distribution is often used to represent the time between events for Poisson processes with a constant average rate, such as queuing. The theoretical probability density function (PDF) can be expressed as $P(x) = \lambda e^{-(\lambda x)}$, while the theoretical cumulative density function (CDF) can be expressed as $P(X <= x) = 1 - e^{-(\lambda x)}$.

The figure below shows how the exponential PDF and CDF changes as λ varies from 0.1 to 2 (note that both the PDF and CDF are defined as zero for any x value less than zero):

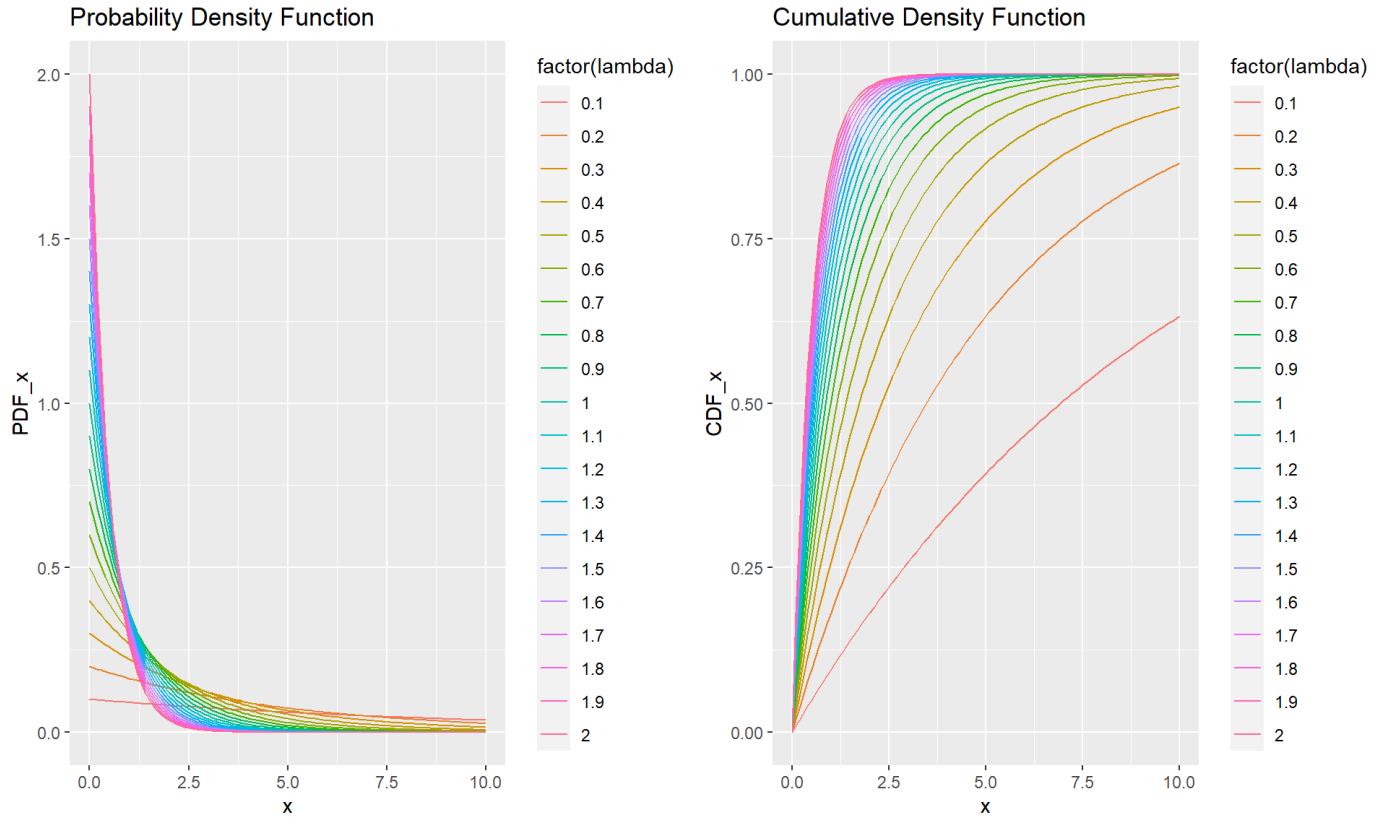

Figure 1: Exponential Distribution for Varying Lambda Values

Based on these plots, the curve becomes 'tighter' as lambda is increased. To determine the theoretical mean and variance values for λ=0.2 for our simulation:

Theoretical Mean (μ) = 1/λ = 1/0.2 = **5**

Theoretical Variance (σ^2) = $1/\lambda^2$ = *1/0.04* = **25**

2.2 Simulations

The exponential distribution can be simulated in R with `rexp(n,lambda)`, where n is the number of observations and lambda (λ) is the rate parameter. The mean of exponential distributions is 1/λ and the standard deviation is also 1/λ. For the purposes of the simulation, use a λ value of 0.2 and investigate the averages of 40 randomly generated points of the exponential distribution for 1000 simulations.

Executing 1000 simulations:

```
# Set seed for reproducibility:
set.seed(1)
# Define parameters for simulation:
lambda <- 0.2; n <- 40; num.sims <- 1000;
# Run simulations
sim <- as.data.frame(replicate(num.sims, rexp(n,lambda)))
# Determine the mean/variance for each of the simulations:
sim.mean <- data.frame(value = unlist(lapply(sim, mean), use.names=F))
sim.var <- data.frame(value = unlist(lapply(sim, var), use.names=F))
```

From these simulations, there are 1000 records of simulation means (saved as `sim.mean`) and simulation variances (saved as `sim.var`). These values will be used to determine the overall sample mean and variance in the next sections.

2.3 Sample Mean versus Theoretical Mean

From the exploratory analysis, the Theoretical Mean for an exponential distribution with λ=0.2 is 5.

Using the 1000 simulations, the distribution of simulation means is shown below. This plot contains a histogram showing the frequency of a mean occurring (in bins of 0.2), along with the simulation probability density function (shown as a black curve). The vertical dashed line shows the sample mean for the simulated exponential distribution.

```
gghistogram(sim.mean, x="value", y="..density..", binwdith=0.2, add="mean", add_density=T, fill="lightgray")
+
  ggtitle("Simulation Density Plot (Mean)") + xlab("Simulation Mean Value")
```

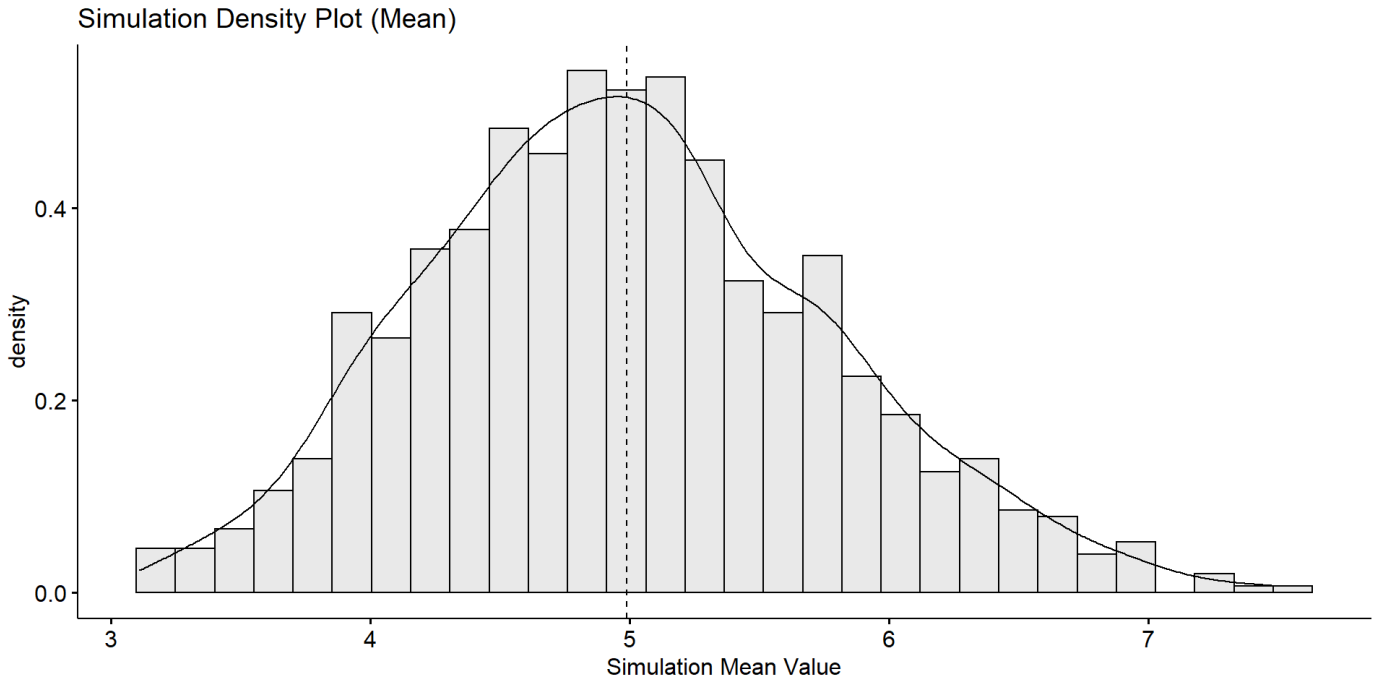

Figure 2: Simulation Sample Mean Distribution

To determine the sample mean, take the average of the simulation data means. For the 1000 simulations, this results in:

```
mean(sim.mean$value)
```

```
## [1] 4.990025
```

Therefore, the average sample mean is **4.99**, and the difference between the sample mean and theoretical mean is **-0.01**.

2.4 Sample Variance versus Theoretical Variance

From the exploratory analysis, the Theoretical Variance for an exponential distribution with λ=0.2 is 25.

From the 1000 simulations, the distribution of variance is shown below. This plot contains a histogram showing the frequency of the variance occurring (in bins of 1.0), along with the simulation probability density function (shown as a black curve). The vertical dashed line shows the sample variance for the simulated exponential distribution.

```
gghistogram(sim.var, x="value", y="..density..", binwidth=1, add="mean", add_density=T, fill="lightgray") +
  ggtitle("Simulation Density Plot (Variance)") + xlab("Simulation Variance Value")
```

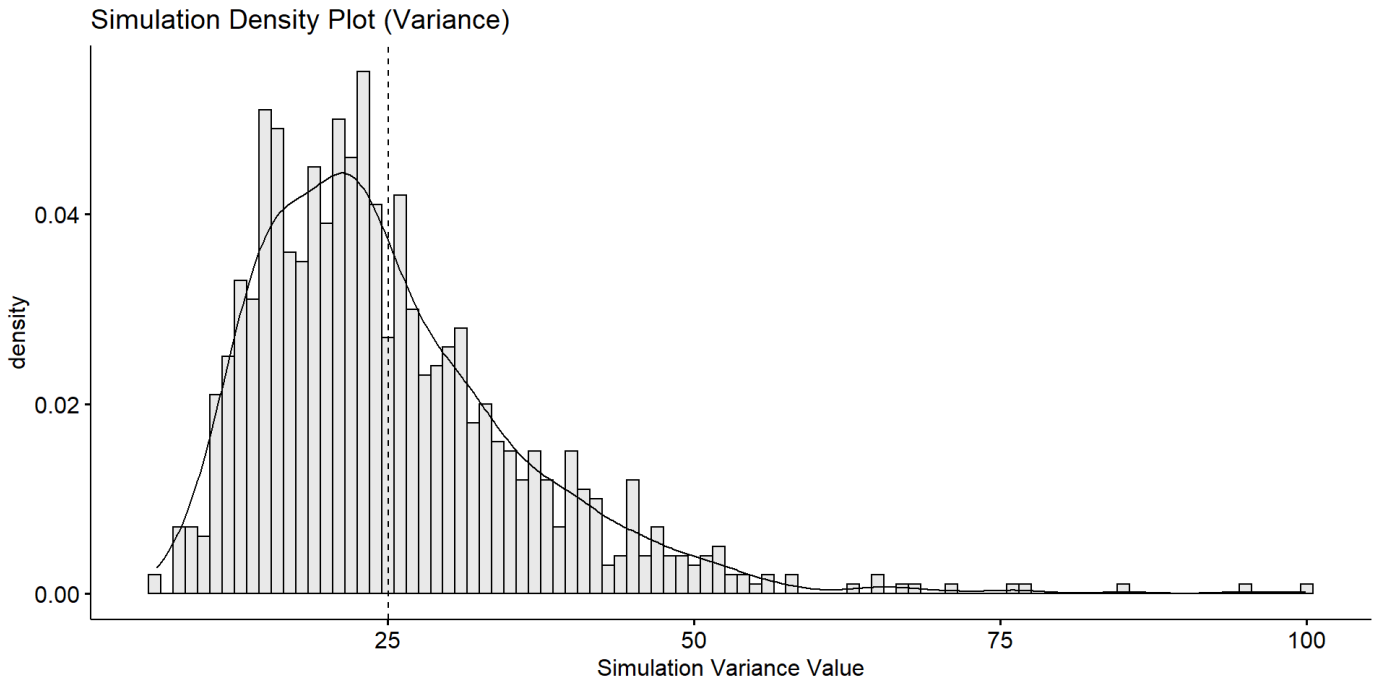

Figure 3: Simulation Sample Variance Distribution

To determine the sample variance, take the average of the simulation data variances. For the 1000 simulations, this results in:

```
mean(sim.var$value)
```

```
## [1] 25.06459
```

Therefore, the average sample mean is **25.06**, and the difference between the sample variance and theoretical variance is **0.06**.

2.5 Distribution

The Central Limit Theory states that the distribution of averages of Independent and Identically Distributed (IID) random variates becomes that of a standard normal as the sample size increases. This should result in a standard normal distribution for the following test statistic (t):

```
  t = (Estimate - Mean of estimate)/(Std. Error of estimate) = (Xbar,n - μ) / (σ/sqrt(n))
```

Where the Estimate ($X_{bar,n}$) is equal to our simulated mean values, the mean of estimate (μ) is equal to the population mean (in this case, use theoretical mean), and the standard error (Denominator) is calculated using standard deviation (σ) and sample sizes (n).

To calculate this statistic (t) for the data:

```
xbar <- sim.mean$value # Sample mean for 1000 simulations
mu <- 5 # Known theoretical mean for exponential dist. with lambda=0.2
sigma <- sqrt(sim.var$value) # Standard Deviation is sqrt of variance
n <- 40 # 40 observations for each sample

t <- data.frame(value=(xbar-mu)/(sigma/sqrt(n))) # t-statistic for each simulation
```

Notice that since the x_{bar} and μ values are subtracted, the center of the simulation t-statistic distribution is zero (allowing comparison against a standard normal distribution, with a mean of zero and a variance of one).

Plotting this statistic for the simulation data (red curve), and comparing to the standard normal distribution (green curve):

```
qplot(t$value, geom = "blank") +
  geom_histogram(aes(y = ..density..), binwidth=0.1, col=I('black'),fill=I('grey')) +
  stat_density(geom = "line", aes(colour = "Simulation"), size=1.25) +
  stat_function(fun = dnorm, n=1000, args=list(mean=0, sd=1), aes(colour = "Normal"),size=1.25) +
  geom_vline(xintercept=0, col="green",size=1.25,linetype="dashed") +
  geom_vline(xintercept=mean(t$value), col="red",size=1.25,linetype="dashed") +
  scale_colour_manual(name = "",
                      values = c("red", "green"),
                      breaks = c("Simulation", "Normal"),
                      labels = c("Simulation", "Normal")) +
  ggtitle("Simulation Density Plot (T-statistic)") +
  xlab("Simulation t-statistic Value") +
  ylab("Probability Density") +
  theme(legend.position = "bottom", legend.direction = "horizontal")
```

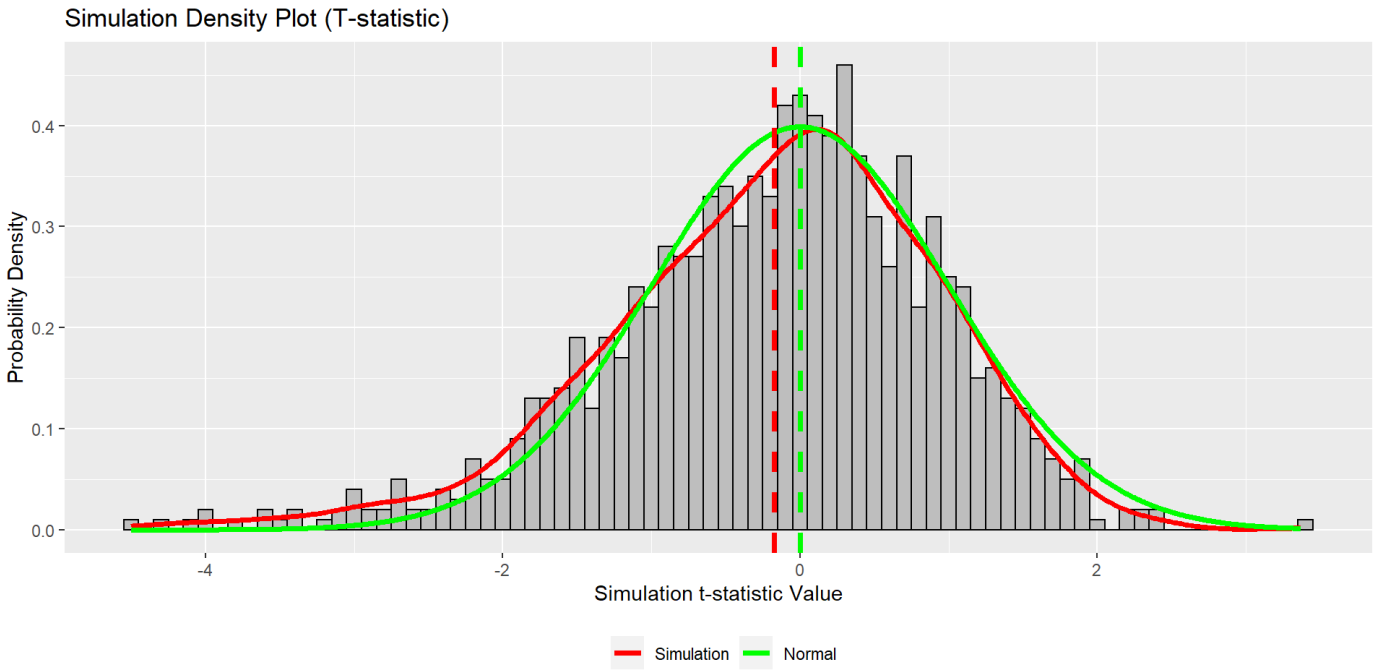

Figure 4: Simulation Sample T-statistic Distribution vs. Normal Distribution

From the distribution, the mean t-statistic for the simulations is **-0.17**.

3. Conclusions

The Central Limit Theorem is confirmed through the results in Section 2.5. The plot of simulated t-statistics revealed that the distribution is approximately normal, with a small difference between the mean t-statistic and zero. Despite the exponential distribution being very different from a Gaussian distribution, this simulation showed that the sample distributions are very similar to normal distributions with large enough sample sizes. This provides confidence in using methods such as t-tests which generally require that data is normal.