Part 1: Simulation Exercise

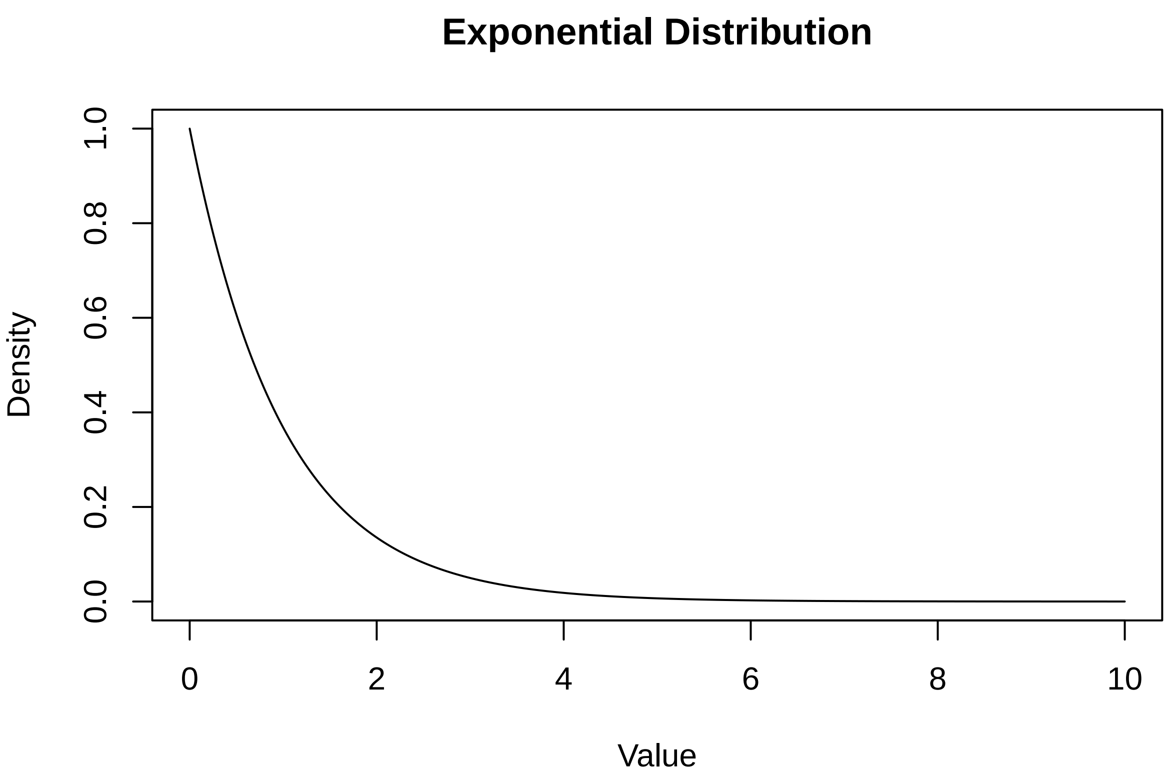
Akhil Kota

2022-07-22

Overview

The exponential distribution is a frequently used distribution with a rate parameter λ, which determines its mean and standard deviation (both of which are $\frac{1}{\lambda}$). In this report, we will be investigating the distribution of means of samples taken from the exponential distribution. The samples will each have 40 random exponential variables, and a total of 1000 samples will be taken to get a good feel for the distribution and its' properties.

For reference, a plot of the exponential distribution is provided below.



Simulations

The simulations were carried out by obtaining 40,000 total random data from the exponential distribution with a set $\lambda = 0.2$. These data were reshaped into a matrix with 1000 rows and 40 columns, representing our 1000 simulations of 40 random exponentials each. The means were taken across each row, so the vector "means" contains the data for our relevant exponential sample mean distribution. The mean and standard deviation of this distribution were then calculated and printed below. Note that the seed is specifically set for reproducibility.

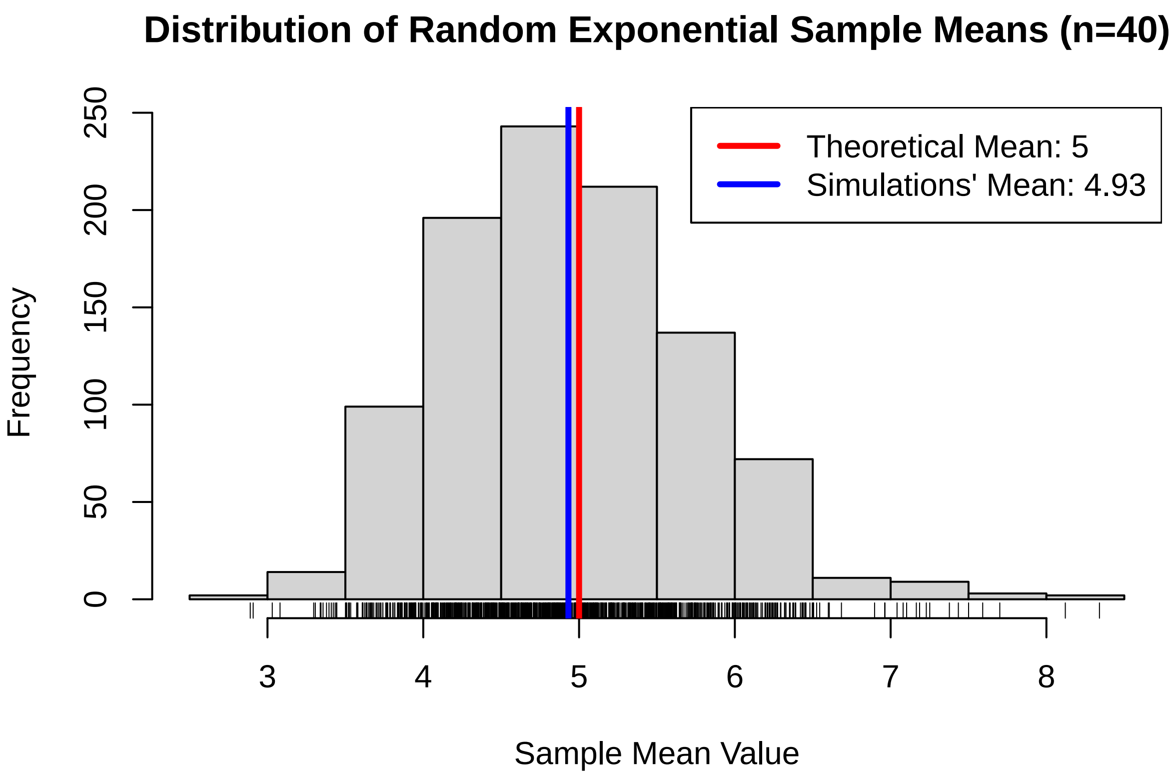
```
lambda <- 0.2 ## Rate parameter in exp distribution
n <- 40 ## Number of exp vars per sample
s <- 1000 ## Number of sims
set.seed(seed = 1067)
dat <- matrix(data = rexp(n*s, rate = lambda), nrow = 1000, ncol = 40)
means <- apply(dat, 1, mean)
distb_mean <- mean(means)
distb_var <- var(means)
distb_sd <- sd(means)
data.frame(
    Mean = distb_mean,
    Variance = distb_var,
    Standard.Deviation = distb_sd,
    row.names = "Summary Statistics")
```

```
##                        Mean  Variance Standard.Deviation
## Summary Statistics 4.931685 0.6270553          0.7918682
```

Sample Mean vs Theoretical Mean

The theoretical mean μ of the exponential distribution is $1/\lambda$, which, with $\lambda = 0.2$, is $\mu = 5$. This means that the theoretical mean of the distribution of sample means of any number of random exponentials should also be $\mu_{\bar{x}} = 5$. The sample mean distribution mean we got from our simulation was $\mu_{sim} = 4.9316849$.

In the plot below, we show the 1000 sample means from the simulation, the mean of the distribution of sample means, and the theoretical mean of the exponential distribution (which is also the theoretical mean of the sample means distribution).



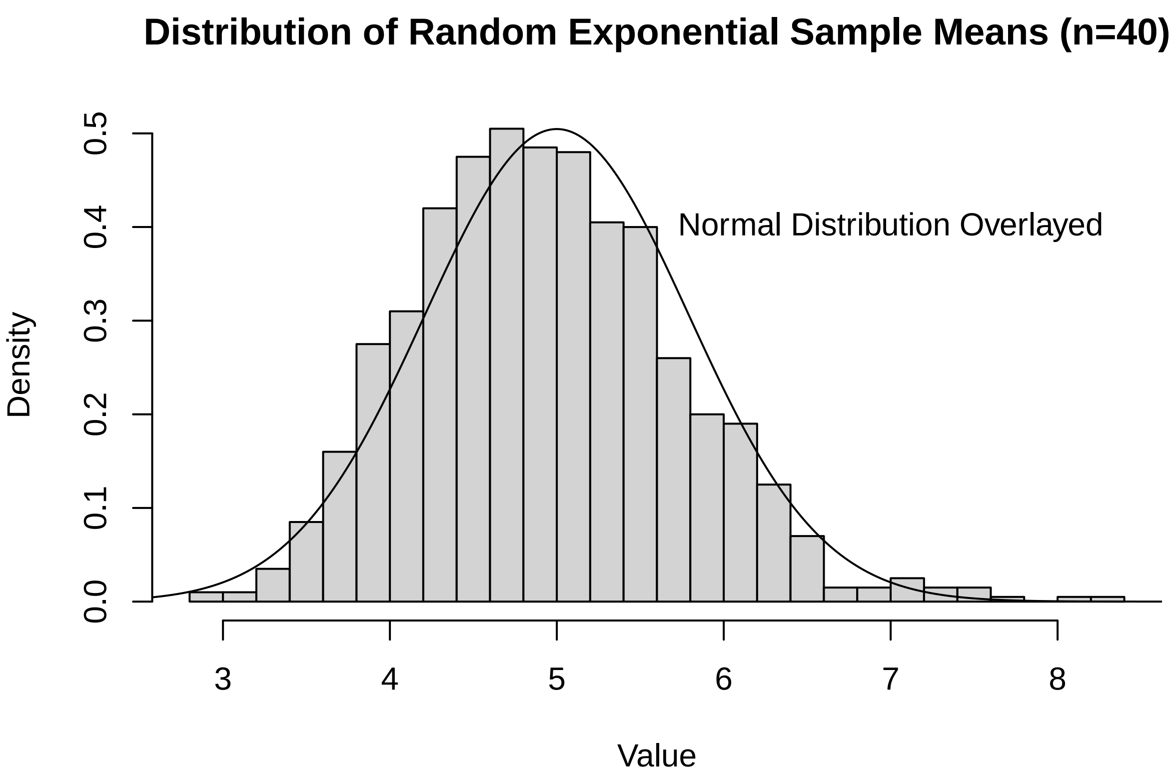
Sample Variance vs Theoretical Variance

The variance of the exponential distribution itself here is $(\frac{1}{\lambda})^2 = 25$. However, the act of taking samples of several random exponentials, averaging across the samples, and looking at the distribution of sample means leads to a significantly reduced variance, depending on the sample size used. The theoretical variance of our sample mean distribution becomes $\sigma_{\bar{x}}^2 = \frac{1}{n\lambda^2} = \frac{25}{40} = 0.625$. This corresponds to a theoretical standard error of the mean of $\sigma_{\bar{x}} = \sqrt{5/8} = 0.79$.

From the simulation, we obtained values of $\sigma_{sim}^2 = 0.6270553$ and so $\sigma_{sim} = 0.7918682$. The code is in the Simulations section above. These are both slightly larger than the theoretical variance and sd, but correspond very well with the theoretical values.

Similarity to Normal Distribution

Now, we'll take a closer look at our sample mean distribution once again. By the Central Limit Theorem, this distribution should be approximately normal now. Let's see if this is true visually by overlaying a normal distribution with theoretical values $\mu_{\bar{x}} = 5$ and $\sigma_{\bar{x}} = \sqrt{5/8}$.



We see that, compared to the normal distribution, our sample mean distribution of exponential samples is relatively close to the normal distribution, but slightly right-skewed, which makes sense given the severe right-skew of the exponential distribution itself. A larger sample size for our samples would move this distribution even closer to normal (by CLT).

Now, let's look at the quantiles of our distribution and compare these to normal distribution quantiles.

```
pct <- seq(0,1,length.out=11)
dfquantiles <- data.frame(Cumulative.Probability.Below = pct,
    Distribution.Quantiles = quantile(means, probs=pct),
    Normal.Quantiles = qnorm(pct, 5, sqrt(5/8)))
dfquantiles
```

```
##       Cumulative.Probability.Below Distribution.Quantiles Normal.Quantiles
```

```
## 0%                                 0.0          2.889311        -Inf
## 10%                                0.1          3.945709    3.986845
## 20%                                0.2          4.247200    4.334640
## 30%                                0.3          4.485637    4.585425
## 40%                                0.4          4.680479    4.799712
## 50%                                0.5          4.868820    5.000000
## 60%                                0.6          5.084070    5.200288
## 70%                                0.7          5.321477    5.414575
## 80%                                0.8          5.579661    5.665360
## 90%                                0.9          5.984343    6.013155
## 100%                               1.0          8.340485         Inf
```

The extreme probability values are to be ignored, as the normal distribution spans infinitely, and our sample mean distribution does not. However, the other quantile values match up quite nicely between our distribution and the normal distribution. Again, we see the right-skew of the sample mean distribution in the fact that every single quantile is slightly lower than that of the normal, but it is only a relatively small difference, especially compared to the difference between the exponential distribution itself and the normal. As the sample size is increased, the sample mean distribution will tend closer and closer towards the normal distribution, by CLT.