Statistical Inference, Part 1: Simulation

ABC

January 17, 2019

Part 1: Simulation

Overview

In this project I will be examining the exponential probability distribution by generating 40,000 values and examining the distribution characteristics along with the characteristics of the sampling distribution of the mean with sample size of 40. The distribution being examined has a single variable, `lambda` (λ). The mean and standard deviation of the distribution are $1/\lambda$.

Simulations

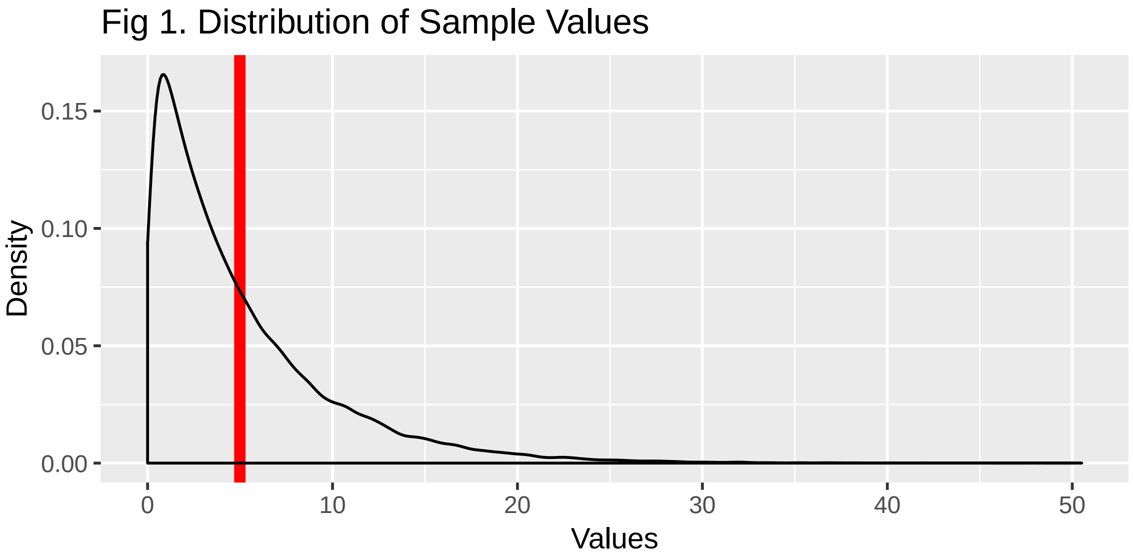
First, I created a variable called `sample` and loaded it with `rexp(40*1000, 0.2)`, 40,000 random values of the exponential distribution with `lambda = 0.2`. This vector was dimensioned into a 1000 x 40 matrix.

The `mean()` function was applied across the rows of the matrix, yielding the sampling distribution of the mean, `sample_bar`.

```
set.seed(54321)
sample <- matrix(rexp(40*1000, 0.2), 1000, 40)
sample_bar <- apply(sample, 1, mean)
```

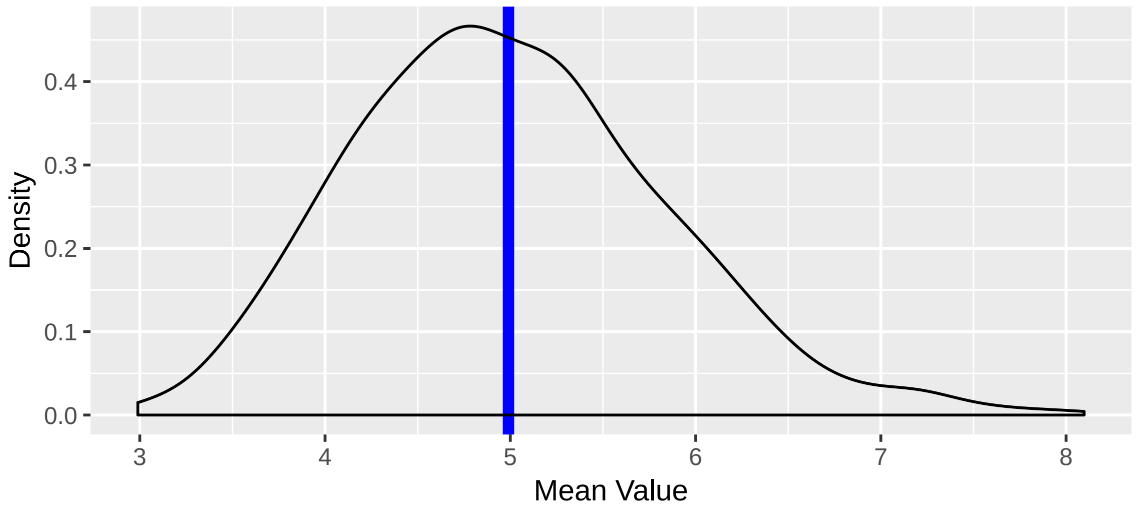
Sample Mean vs Theoretical Mean

```
sample_Summ <- summary(c(sample))
sample_bar_Summ <- summary(c(sample_bar))
sample_bar_mean <- sample_bar_Summ[[4]]
```



Taking the `summary()` of `sample` we see that the mean is 4.99. This is very close to the theoretical value of the mean, $1/\lambda = 5$.

Fig 2. Distribution of Mean Values



The mean of `sample_bar` is the same: 4.99. This is expected from the Central Limit Theorem.

Sample Variance vs Theoretical Variance

```
sample_Var <- var(c(sample))
sample_bar_var <- var(c(sample_bar))
```

The variance of the exponential distribution can be calculated directly with the `var()` function. The theoretical variance for an exponential distribution is $(1/\lambda)^2$, or 25 for our generating function. The calculated variance of the 40,000 samples was 25.1.

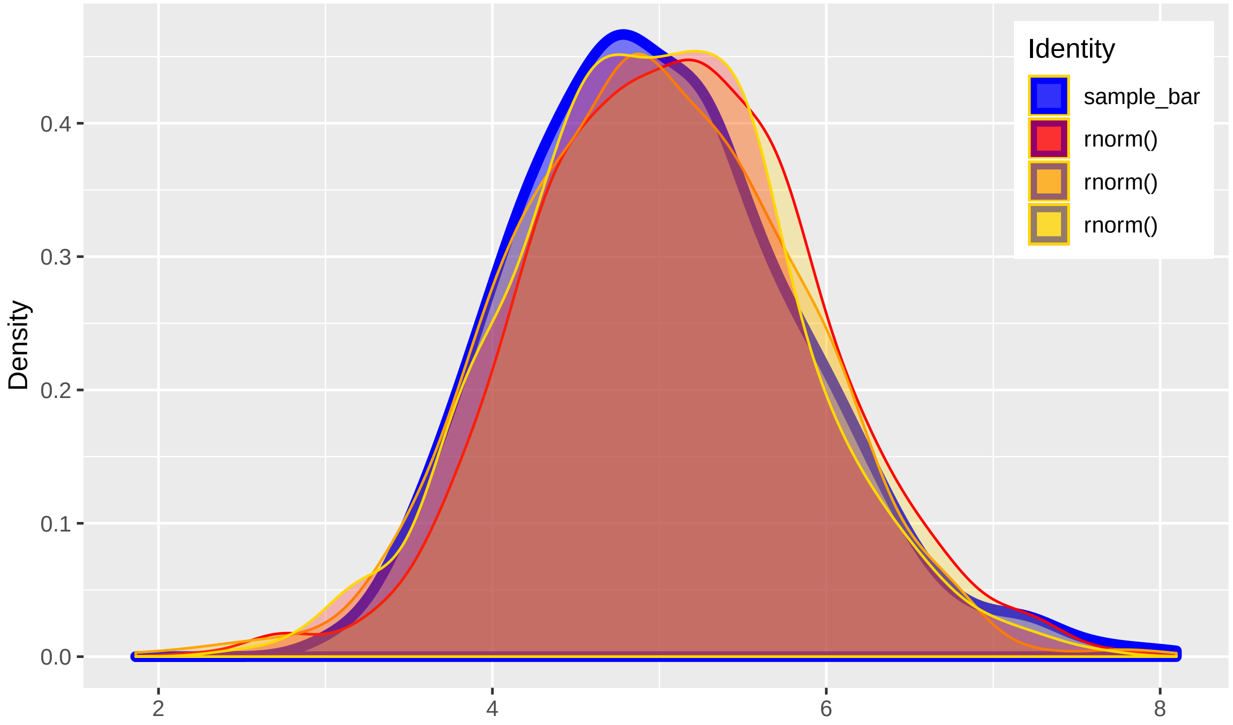
Taking the variance of `sample_bar` yields 0.708. For a distribution of means the variance is σ^2/n, where σ is the population standard deviation, and n is the sample size. For our generating function this would yield an expected variance of the sampling distribution of the mean of 0.625. This is less than the calculated variance, suggesting that the distribution is more variable than a standard normal distribution. This would be improved by increasing n.

Distribution

To examine the distribution's shape I compared it against three normal random distributions of equal size (n = 1000) and using the calculated standard deviation, 0.8412148 ($\sqrt{sample\_bar\_var}$), centered around 4.99 (`sample_bar_mean`).

Fig 3. Density plot of the Sampling Distribution of the mean

Plotted with 3 random normal distributions of the same variance,
sample size and mean



This visual comparison suggests that the distribution is very close to a normal distribution. It is possibly slightly skewed, but it is hard to tell if that is due to the large skewedness of the sampled from population (the exponential distribution) or the low number of means.

Appendix

Code for Figure 1

```
sample_gg <- as_tibble(c(sample))
ggplot(data = sample_gg) +
    geom_vline(xintercept = sample_Summ[[4]],
               color = 'red',
               size = 2) +
    geom_density(mapping = aes(x = value)) +
    labs(title = "Fig 1. Distribution of Sample Values",
         x = "Values",
         y = "Density")
```

Code for Figure 2

```
sample_bar_gg <- as_tibble(c(sample_bar))
ggplot(data = sample_bar_gg) +
    geom_vline(xintercept = sample_bar_Summ[[4]],
               color = 'blue',
               size = 2) +
    geom_density(mapping = aes(x = value)) +
    labs(title = "Fig 2. Distribution of Mean Values",
         x = "Mean Value",
         y = "Density")
```

Code for Figure 3

```
set.seed(121212)
ggplot(data = as.data.frame(sample_bar)) +
    geom_density(mapping = aes(sample_bar,
                               fill = 'blue'),
                               size = 2,
                               color = 'blue',
                               alpha = .5) +
    geom_density(mapping = aes(rnorm(1000,
                                     sd = sqrt(sample_bar_var),
                                     mean = sample_bar_mean),
                               fill = 'red'),
                               color = 'red',
                               alpha = .25) +
    geom_density(mapping = aes(rnorm(1000,
                                     sd = sqrt(sample_bar_var),
                                     mean = sample_bar_mean),
                               fill = 'orange'),
                               color = 'orange',
                               alpha = .25) +
    geom_density(mapping = aes(rnorm(1000,
                                     sd = sqrt(sample_bar_var),
                                     mean = sample_bar_mean),
                               fill = 'gold'),
                               color = 'gold',
                               alpha = .25) +
    labs(x = "",
         y = "Density",
         title = "Fig 3. Density plot of the Sampling Distribution of the mean",
         subtitle = "Plotted with 3 random normal distributions of the same variance,
         sample size and mean") +
    theme(legend.position = c(.9,.8)) +
    scale_fill_manual(name = 'Identity',
                       values = c('blue', 'red', 'orange', 'gold'),
                       labels = c('sample_bar', 'rnorm()', 'rnorm()', 'rnorm()'))
```