Exponential distribution and the Central Limit Theorem

Heli Pykälä

2016-05-25

Overview

The Central Limit Theorem states that averages of independent and identically distributed random variables should follow the standard normal distribution when properly normalized. We test this by simulation. We find that the Central Limit Theorem seems to hold for averages of 40 samples from the exponential distribution $Exp(0.2)$

Simulation

First we set the parameters for the simulation. Exponential distribution has one parameter: rate marked by λ. In this simulation we use $\lambda = 0.2$. We draw 40 samples from $Exp(\lambda)$ and repeat the simulation 1000 times. Call the number of samples and the number of simulations n and m respectively.

```
lambda <- .2
n <- 40
m <- 1000
```

Then set seed for reproducibility and draw $n * m$ samples from the distribution.

```
set.seed(136)
samples <- rexp(n*m, lambda)
```

The sample mean

Arrange the samples into a matrix with 1000 rows and 40 columns and calculate row-wise means to acquire 1000 samples of means of 40 exponentials.

```
mat <- matrix(samples, nrow = m, ncol = n)
means <- apply(mat, 1, mean)
```

Sample Mean versus Theoretical Mean

Theoretically the mean of the means should be the population mean, $\mu = \frac{1}{\lambda}$

```
data.frame(Theoretical = 1/lambda, Sample.mean = mean(means))

##   Theoretical Sample.mean
## 1           5    5.013405
```

The sample mean is close to the theoretical value.

Sample Variance versus Theoretical Variance

The standard deviation should be $\frac{\sigma}{\sqrt{n}} = \frac{1}{\lambda\sqrt{n}}$

```
data.frame(Theoretical = 1/lambda/sqrt(n), Sample.sd = sd(means))

##   Theoretical Sample.sd
## 1   0.7905694 0.8186127
```

The standard deviation of the sample is close to the theoretical value.

Distribution

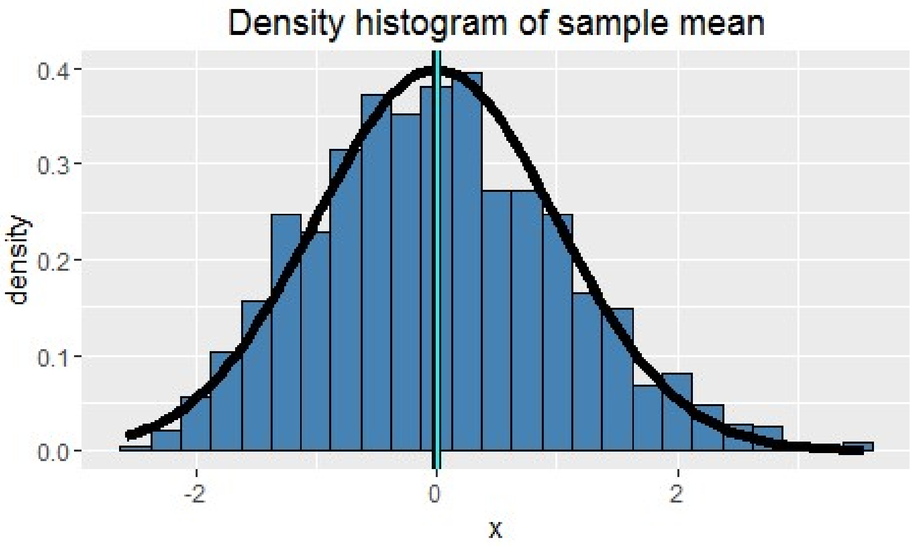
Normalize the values by subtracting the mean and dividing by the standard deviation

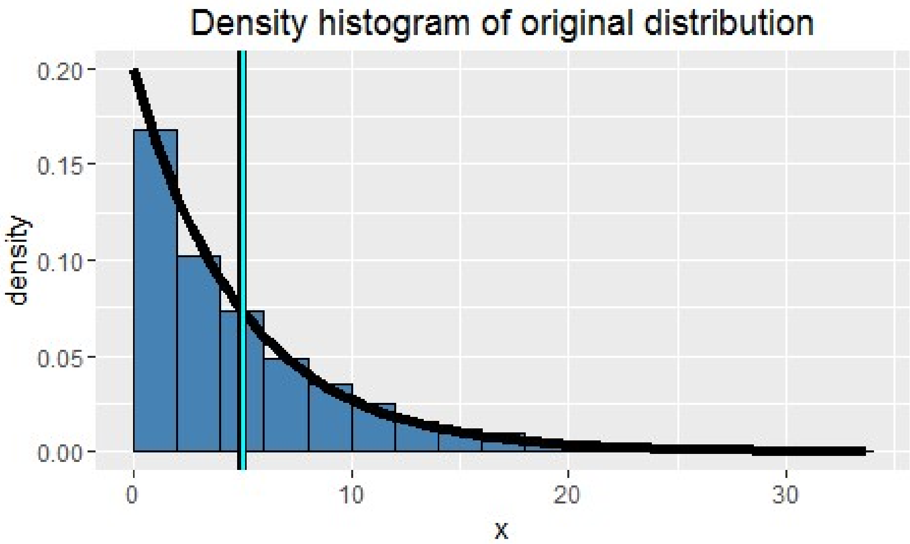
```
scaledMeans <- scale(means, center = 1/lambda, scale = 1/lambda/sqrt(n))
```

Plot the density histogram of the scaled averages along with the standard normal distribution. Also for comparison plot the density histogram of 1000 samples of the underlying exponential distribution. For both figures the thick black vertical line shows the theoretical mean and the cyan line the mean calculated from the data.

```
library(ggplot2)
dfMeans <- data.frame(x = scaledMeans)
ggplot(dfMeans,aes(x=x))+
    geom_histogram(colour = "black", fill = "steelblue",
                   binwidth = .25, aes(y = ..density..)) +
    stat_function(fun = dnorm, size = 2) +
    geom_vline(xintercept = 0, size = 2, color = "black") +
    geom_vline(xintercept = mean(scaledMeans), size = 1, color = "cyan") +
    ggtitle("Density histogram of sample mean")

dfSamples <- data.frame(x = samples[1:1000])
ggplot(dfSamples,aes(x=x))+
    geom_histogram(colour = "black", fill = "steelblue", binwidth = 2,
                   boundary = 0,  aes(y = ..density..)) +
    stat_function(fun = dexp, size = 2, args = list(rate = lambda)) +
    geom_vline(xintercept = 1/lambda, size = 2, color = "black") +
    geom_vline(xintercept = mean(dfSamples$x), size = 1, color =  "cyan") +
    ggtitle("Density histogram of original distribution")
```





The histogram looks similar to the theoretical normal distribution. The distribution isn't perfectly normal and has some deviations from expected. The normality of the result could be improved by increasing the sample size. The original exponential distribution has very different characteristics being always > 0, having the largest density at 0, and then exponentially decreasing. Compared to the original distribution the distribution of sample means is much more normal.

Conclusion

The Central Limit Theorem seems to hold reasonably well for averages of 40 exponentials. The original exponential distribution is really different from the normal distribution, but when we calculate the average of 40 random samples the resulting random variable seems to follow the normal distribution with expected parameters quite well.