Distribution of Avarages of Samples of 40 Exponential Randoms

Michal Siwek

Overview

This is part 1 of the Project Assignement for Statistical Inference course by JHU at Coursera. It illustrates via simulation and explarotory graphs the properties of the distribution of the mean of 40 exponentials. In particular it shows:

- mean of sample means and compare it with the theoretical mean
- distribution of means of samples and compare it with theoretical distibution assuming CLT
- mean of sample variances and compare it with theoretical variance
- distribution of variances of samples
- variance of sample means and compare it with theoretical variance
- distributions of exponential randoms and of means of samples of explential randoms and compare them with the normal distribution

Simulations

```
lambda <- .2; n <- 40; nosim <- 1000; set.seed(1234)
sim <- rexp(n*nosim, lambda)
mtx <- matrix(sim, nosim)
```

mtx is a matrix containing 1000 samples of 40 exponential randoms. Samples are represented by rows.

Sample Mean versus Theoretical Mean

Let's calculate mean of sample means and the theoretical mean and standard deviation.

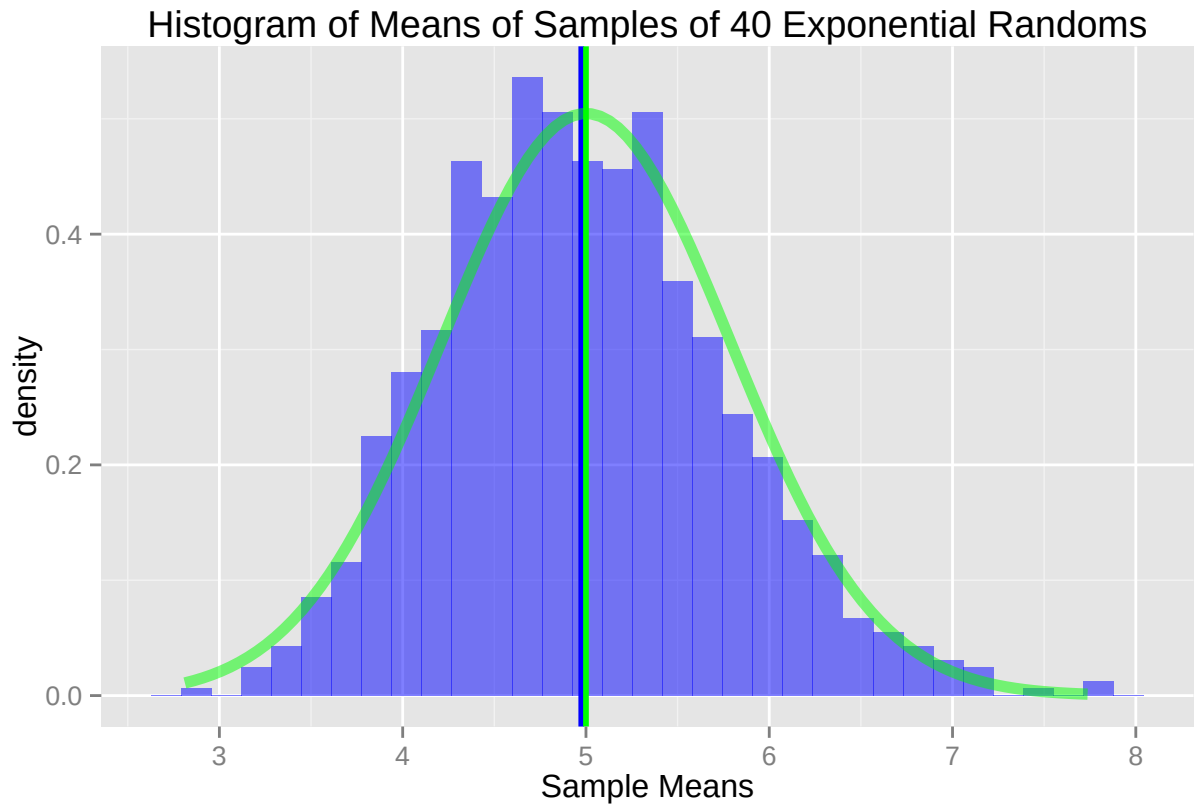
```
library(scales) # to show % in the commenting text
mns <- apply(mtx, 1, mean) # vector of sample means
mn <- mean(mns)
t_mn <- 1/lambda
```

The mean of sample means is 4.9742388 while the theoretical mean is 5, the difference is 0.515%.

By CLT the distribution of sample means should be approximately normal. Let's plot the actual and the asymptotic theoretical distribution.

```
library(ggplot2)
t_sd <- 1/lambda # standard deviation of the population
data <- data.frame(Sample_Means = mns)
g <- ggplot(data, aes(x=Sample_Means), )
```

```
g <- g + geom_histogram(alpha=.5, fill="blue", aes(y = ..density..,))
g <- g + geom_vline(aes(xintercept=mean(mns)), color="blue", size=1)
g <- g + stat_function(fun=dnorm, args=list(mean=t_mn, sd=t_sd/sqrt(n)),
                       alpha=.5, size = 2, color="green")
g <- g + geom_vline(aes(xintercept=t_mn), size=1, color="green")
g <- g +  scale_x_continuous(name="Sample Means", breaks=seq(2,8,1))
g <- g + ggtitle("Histogram of Means of Samples of 40 Exponential Randoms")
g
```



The actual distribution of sample means is shown in blue, the vertical blue line indicates the actual mean of sample means.

The asymptotic theoretical distribution is shown by the overlaying green line. The vertical green line shows the theoretical mean of sample means.

Sample Variance versus Theoretical Variance

Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.

Empirical Sample Variance versus Theoretical Variance

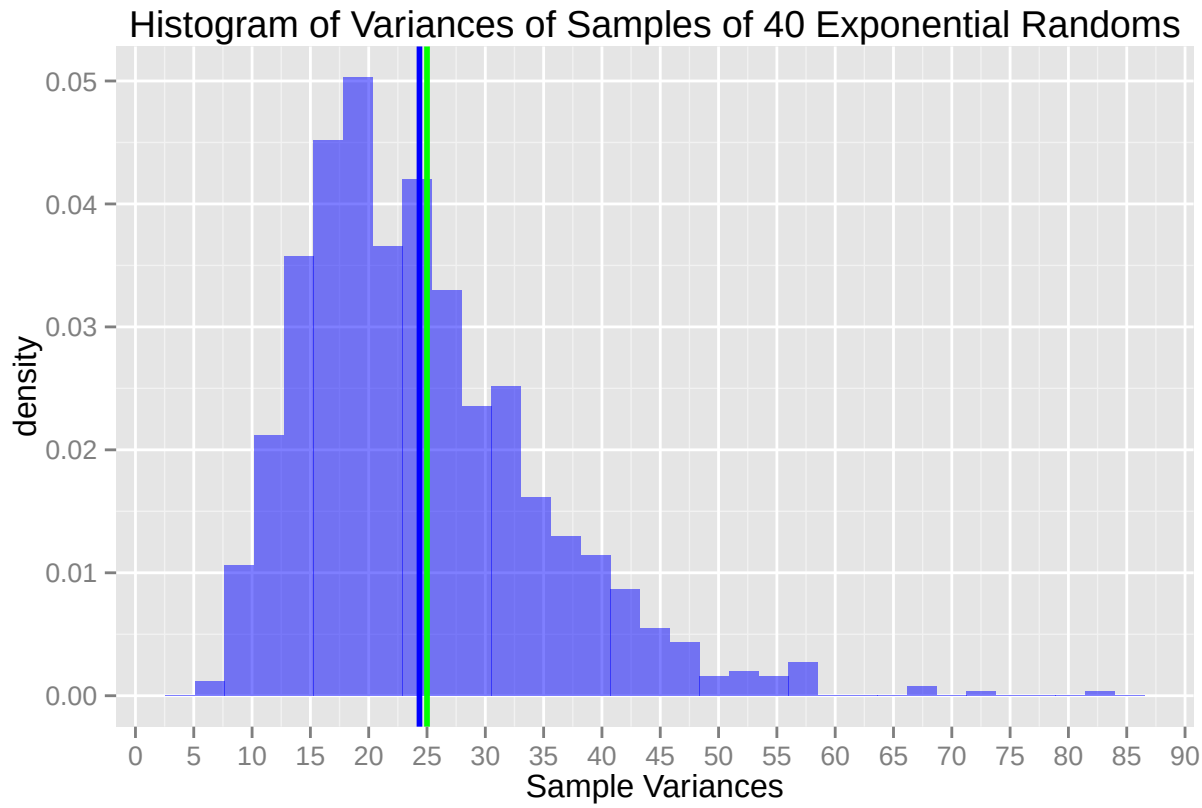
Let's calculate the empirical sample variance and the theoretical variance:

```
vars <- apply(mtx, 1, var) # vector of variances of samples
var_mean <- mean(vars)
t_var <- (t_sd)^2 # theoretical population variance
```

The mean of empirical variances is 24.3531011 while the theoretical variance is 25, the difference is 2.59%.

Let's plot the distribution of the empirical variance.

```
data <- data.frame(Sample_Variances=vars)
g <- ggplot(data, aes(x=Sample_Variances))
g <- g + geom_histogram(alpha = .50, fill="blue", aes(y = ..density..,))
g <- g + geom_vline(aes(xintercept=mean(vars)), color="blue", size=1)
g <- g + geom_vline(aes(xintercept=t_var), size=1, color="green")
g <- g + scale_x_continuous(name="Sample Variances", breaks=seq(0,100,5))
g <- g + ggtitle("Histogram of Variances of Samples of 40 Exponential Randoms")
g
```



The empirical distribution of variances is shown in blue. The blue verical line is the mean of the distribution, while the green vertical line is the theoretical variance of the population.

Variance of the Sample Means versus Theoretical Variance

Let's calculate the variance of the sample means and the theoretical variance.

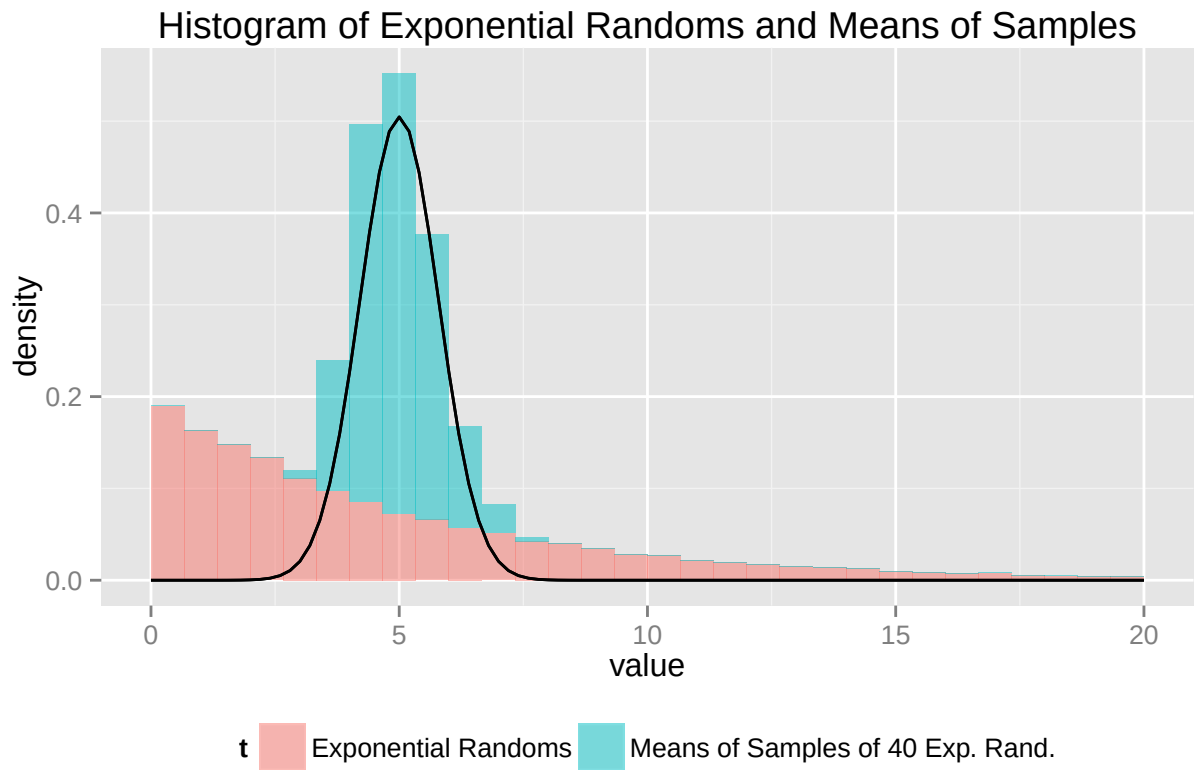
```
var_of_means <- var(mns) # the variance of sample means
t_var_of_means <- (t_sd)^2/n # the theoretical variance of sample means
```

The variance of sample means is 0.5949702 while the theoretical variance is 0.625, the difference is 4.8%.

Distribution

Let's plot the distributions of the exponential randoms and of the sample means of exponential randoms.

```
data <- data.frame(
    x <- c(sim, mns),
    t <- factor(rep(c("Exponential Randoms","Means of Samples of 40 Exp. Rand."),
                    c(nosim*n,nosim))))
g <- ggplot(data, aes(x=x, fill=t))
g <- g + geom_histogram(alpha=.5, aes(y = ..density..,))
g <- g + stat_function(fun=dnorm, args=list(mean=t_mn, sd=t_sd/sqrt(n)))
g <- g + scale_x_continuous(name="value", limits=c(0,20), breaks=seq(0,100,5))
g <- g + theme(legend.position="bottom")
g <- g + ggtitle("Histogram of Exponential Randoms and Means of Samples")
g
```



We can see that in contrast to the exponential randoms the means of samples of exponential randoms have an approximatelly norma distribution. An assymptotic theoretical distribution of sample means given by CLT is overlaid on the plot.