Comparing the distribution of the exponential distribution sample mean and the exponential distribution itself

Sergio Martínez Tornell

Overview

Simulations

The first thing we have to do is to simulate. Define the seed and simulate 40 exponential samples 1000 times and store them in a matrix (each row is a time).

```
set.seed(0)
nsim = 10^3
n = 40
lambda = 0.2

simulations <- matrix(rexp(nsim * n, lambda), nsim)
```

Calculate the mean and the variance of each simulation.

```
means <- apply(simulations, 1, mean)
vars <- apply(simulations, 1, var)
```

Sample Mean versus Theoretical Mean

From theory we know that the sample mean, S, is a random variable normally distributed with $E[S] = \bar{X}$ and $Var[S] = \frac{Var[X]}{n}$. Therefore, the expected averaged value of means is $\frac{1}{\lambda} = 5$. The average mean of our samples is 4.99 which is quite close to the expected value. The theoretical variance,

$$Var[S] = \frac{\sigma}{n} = \frac{1}{\lambda^2/n} =$$

is 0.625, if we calculate the sampled variance we obtain 0.6638725 , which is also close to the theoretical value.

```
1/lambda
```

```
## [1] 5
```

```
mean(means)
```

```
## [1] 4.989678
```

```
1/lambda^2/n
```

```
## [1] 0.625
```

```
var(means)
```

```
## [1] 0.6638725
```

Sample Variance versus theoretical Variance

From theory, we know that the sampled variance is a good estimator of the actual variance of our population. The expected variance is:

$$1/\lambda^2 = 25$$

Our sampled average variance is 25.2377, which is close to the theoretical value, but it is not so close as the previously calculated estimator for the mean. It seems that for accurately estimating the variance, a second order statistic, we need much more repetitions.

```
1/lambda^2
```

```
## [1] 25
```

```
mean(vars)
```

```
## [1] 25.23771
```

Distribution

To plot the histogram of the sampled mean and the histogram of the exponential distribution in the same figure we need to put both in the same data frame.

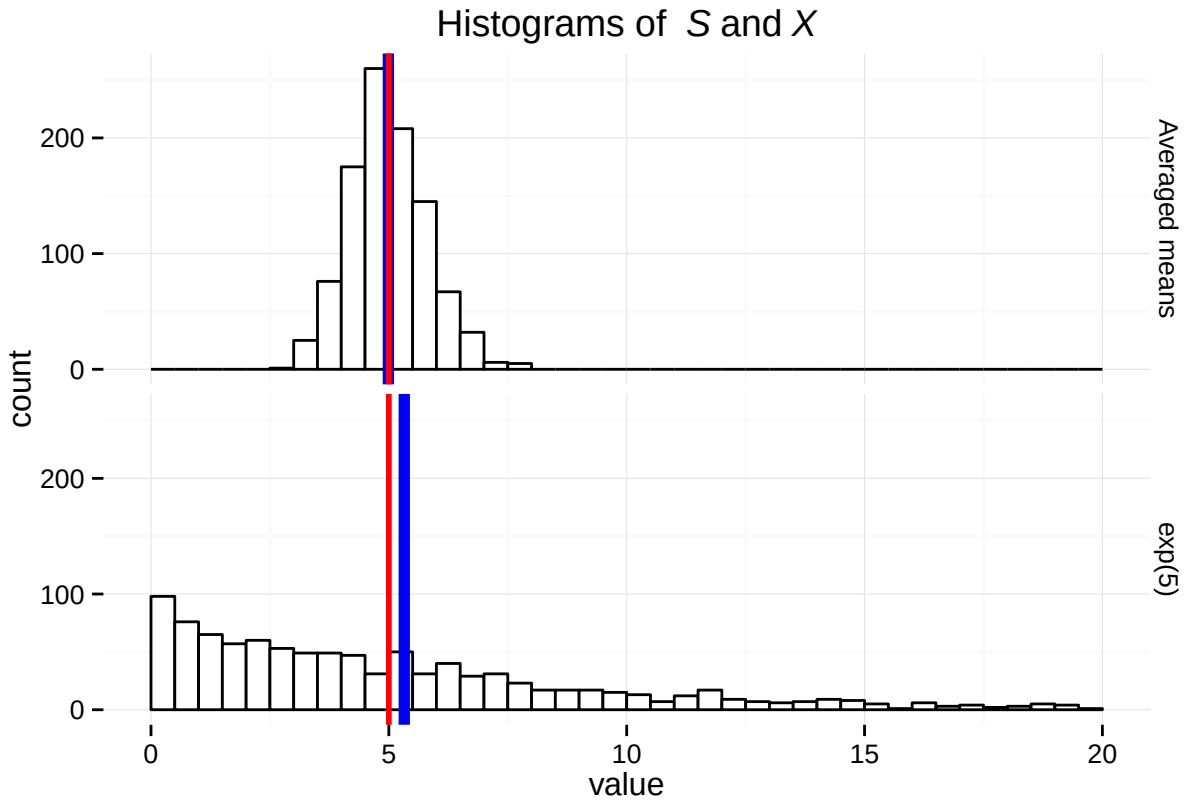
```
library(ggplot2)
library(reshape2)
library(plyr)
comparison <- data.frame(means = means, samples = rexp(1000, lambda));
comparison <- melt(comparison);
```

```
## No id variables; using all as measure variables
```

Then, we can plot them, I have added a red line for the theoretical mean and a blue line for the estimated mean in each case.

```
levels(comparison$variable) <- c("Averaged means", expression(exp(5)))
ggplot(comparison, aes(x=value)) +
  geom_histogram(binwidth = 0.5, fill = "white", colour = "black") +
  geom_vline(data = ddply(comparison, ~ variable, summarise, mean = mean(value)),
             aes(xintercept = mean),
```

```
            colour = "blue", size = 2) +
facet_grid(variable ~ . ) +
geom_vline(xintercept = 1/lambda, colour = "red", size = 1) +
xlim(c(0, 20)) +
ggtitle(expression("Histograms of " ~ italic(S) ~ and ~ italic(X) ))+
theme_minimal()
```



We can see that, while the sampled means are normally distributed around the theoretical mean, the histogram of the exponential distribution is, obviously, close to an exponential function. Moreover, we can see that the averaged sample means are a better estimator of the population mean that the mean of 1000 samples from the distribution.