Analyse de données d'action utilisateur

Yannick Martel

15/10/2014

We first initalize the base parameters:

```
nosim <- 2000
n <- 40
lambda <- 0.2
precision <- 1/sqrt(n)
```

We work on the $\overline{X}_n$ distribution of n averages of the X exponential distribution, with $n = 40$ and $\lambda = 0.2$. We compute 2000 samples of the distribution, plus the standard error for each sample.

```
xSamples <- matrix (rexp(nosim*n, lambda), nrow=nosim)
means <- rowMeans(xSamples)
sds <- apply(xSamples, 1, sd)
```

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution

We can compute the mean over the population of $\overline{X}_n$ samples, and the theoretical mean of the $\overline{X}_n$ distribution:

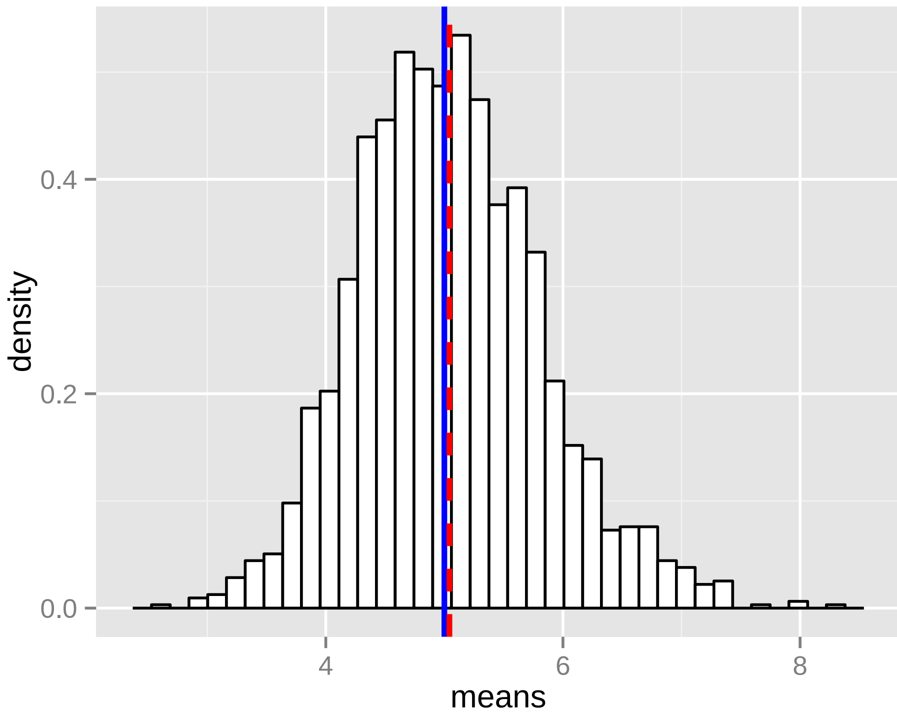
```
distrib_mean <- mean(means)
th_mean <- 1/lambda
```

We compute that the mean of the samples is 5.0425974 vs the theoretical center of distribution at 5, which gives the computed mean at 0.8519471% of the theoretical center.

We then plot the density distribution of the 2000 samples, with an overlay of a blue line for the theoretical center of distribution, a red dashed line for the computed mean of samples.

```
library(ggplot2)
library(grid)

g <- ggplot(data.frame(i = 1 : nosim, means = means), aes(x = means)) +
        geom_histogram(binwidth=precision, colour='black', fill='white',
                       aes(y = ..density..)) +
        geom_vline(aes(xintercept=th_mean), color='blue', size=1) +
        geom_vline(aes(xintercept=distrib_mean), color='red', linetype='dashed', size=1)
print(g)
```



We see, both graphically and in figures, that the distribution of the population of 2000 samples is nearly centered on the theoretical center of the distribution.

2. Show how variable it is and compare it to the theoretical variance of the distribution

To get the variability of the population, we compute their standard error of the population of $\overline{X}_n$, and we compare it to the theoretical standard deviation of $\overline{X}_n$, which is $\sigma/\sqrt{n}$, where $\sigma = 1/\lambda$ is the standard deviation of the exponential distribution.

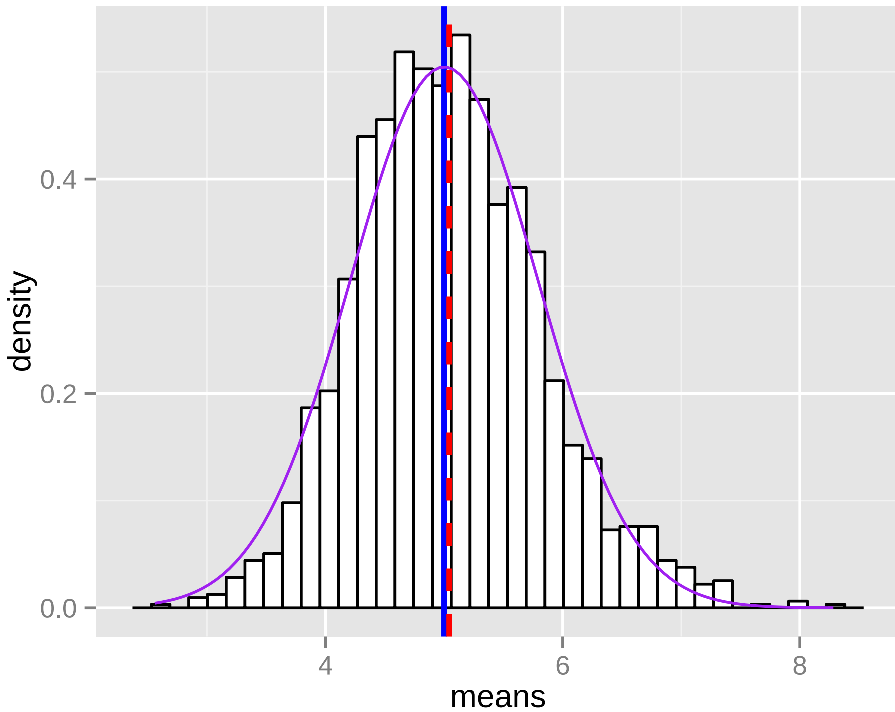
```
distrib_sd <- sd(means)
th_sd <- 1/lambda/sqrt(n)
```

We compute that the standard error of the population is 0.7787855 vs the theoretical standard deviation at 0.7905694, which gives the computed standard deviation at -1.4905569% of the theoretical standard deviation. We can also compare the squared standard error at 0.6065069 to the theoretical variance at 0.625.

3. Show that the distribution is approximately normal

We overlay a (purple) normal distribution of mean $1/\lambda$ and standard deviation $1/(\lambda \times \sqrt{n})$ over the computed density distribution to compare them.

```
print (g + stat_function (fun = function (x) { dnorm(x, mean=th_mean, sd=th_sd) },
                          color='purple'))
```



We can see that the distribution of $\overline{X}_n$ is approximately normal.

4. Evaluate the coverage of the confidence interval for $1/\lambda$: $\overline{X} \pm 1.96 \times S/\sqrt{n}$

We evaluate the confidence interval for each sample with the formula $\overline{X} \pm 1.96 \times S/\sqrt{n}$, where S is the standard error of the sample. We then compute the 2000 samples for which the theoretical mean $1/(\lambda \times \sqrt{n})$ is within this interval.

```
prop_in_confidence_interval <- mean(((means - 1.96 * sds / sqrt(n)) <= th_mean &
                                      th_mean <= (means + 1.96 * sds / sqrt(n))))
```

We get a proportion of 93.3%, which we compare to the theoretical proportion of 95%, associated to the 1.96 coefficient. We found it to be not so close from the theoretical proportion. By experimenting, we found it got closer to the 95% proportion by increasing the n size parameter of the samples from 40 to a value in the range of 1000.