Johns Hopkins - Statistical inference

Benjamin Berhault

September 12, 2015

Simulation exercise - Peer Assessment 1/2

Synopsis

In this project we investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. We investigate the distribution of averages of 40 exponentials.

URL

- Coursera - Johns Hopkins : Statistical Inference (https://class.coursera.org/statinference-032/)
- Code source can be found here : github.com/benjamin-berhault/statistical-inference (https://github.com/benjamin-berhault/statistical-inference)

Mission

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Data preparation

We generate 1000×40 values according to the exponential distribution with a rate $\lambda = 0.2$ and store all those values in a table of 1000 rows and 40 columns. Then, we compute the mean for each row.

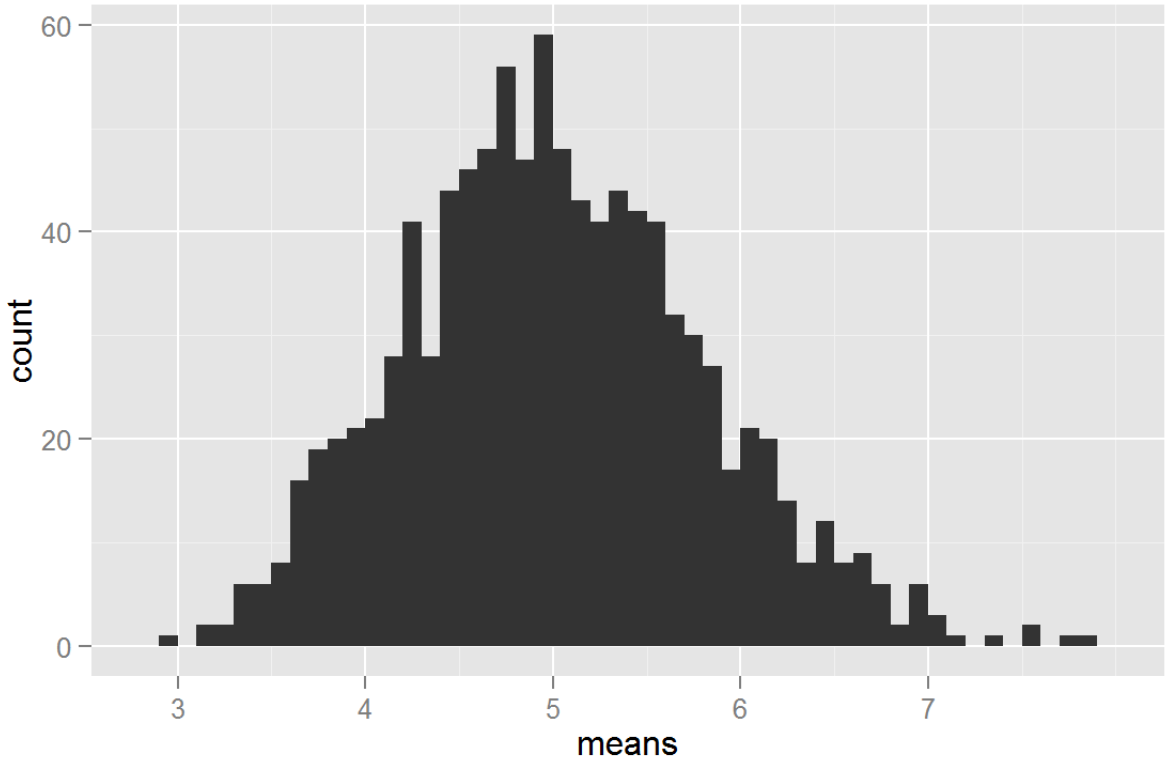
```
### set constants ###
# rate parameter for the exponential distribution function : rexp()
# It's the inverse of the expected duration
lambda <- 0.2 # lambda for rexp
n <- 40 # number of exponetials
numberOfSimulations <- 1000 # number of tests

# set the seed for reproducability
set.seed(1982)

# rexp() here generates 1000*40 values according to the exponential
# distribution with 0.2 as rate parameter
# matrix() put them in one table of 1000 rows (num_sim) and
# 40 columns (sample_size)
exponentialDistriMatrix <- matrix(data=rexp(n * numberOfSimulations, lambda), nrow=number
OfSimulations)
# compute the mean for each row
exponentialDistriMeans <- data.frame(means=apply(exponentialDistriMatrix, 1, mean))
```

Histogram of those 1000 means

```
# plot the means
ggplot(data = exponentialDistriMeans, aes(x = means)) +
  geom_histogram(binwidth=0.1) +
  scale_x_continuous(breaks=round(seq(min(exponentialDistriMeans$means), max(exponentialD
istriMeans$means), by=1)))
```



1. Show the sample mean and compare it to the theoretical mean of the distribution.

The expected mean μ of an exponential distribution of rate λ is $\mu = \frac{1}{\lambda} = \frac{1}{0.2}$

```
mu <- 1/lambda
mu
```

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

Average sample mean of our 1000 samples from exponentially distributed values :

```
meanOfMeans <- mean(exponentialDistriMeans$means)
meanOfMeans
```

```
## [1] 5.014947
```

- Average sample mean expected : **5**
- Average sample mean computed : **5.0149473**

As we can see, theoritical and observed means are very close.

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

The expected standard deviation σ of an exponential distribution of rate λ is $\sigma = \frac{1/\lambda}{\sqrt{n}}$

```
standard_deviation_theory <- 1/lambda/sqrt(n)
standard_deviation_theory
```

```
## [1] 0.7905694
```

Standard deviation of the samples means :

```
standard_deviation_dist <- sd(exponentialDistriMeans$means)
standard_deviation_dist
```

```
## [1] 0.7957905
```

- Standard deviation expected : **0.7905694**
- Standard deviation of the samples means : **0.7957905**

The theorical variance Var of the standard deviation σ is $Var = \sigma^2$

```
variance_theory <- standard_deviation_theory^2
variance_theory
```

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

Variance of the average samples means :

```
variance_dist <- var(exponentialDistriMeans$means)
variance_dist
```

```
## [1] 0.6332826
```

- Variance expected : **0.625**
- Variance of the samples means : **0.6332826**

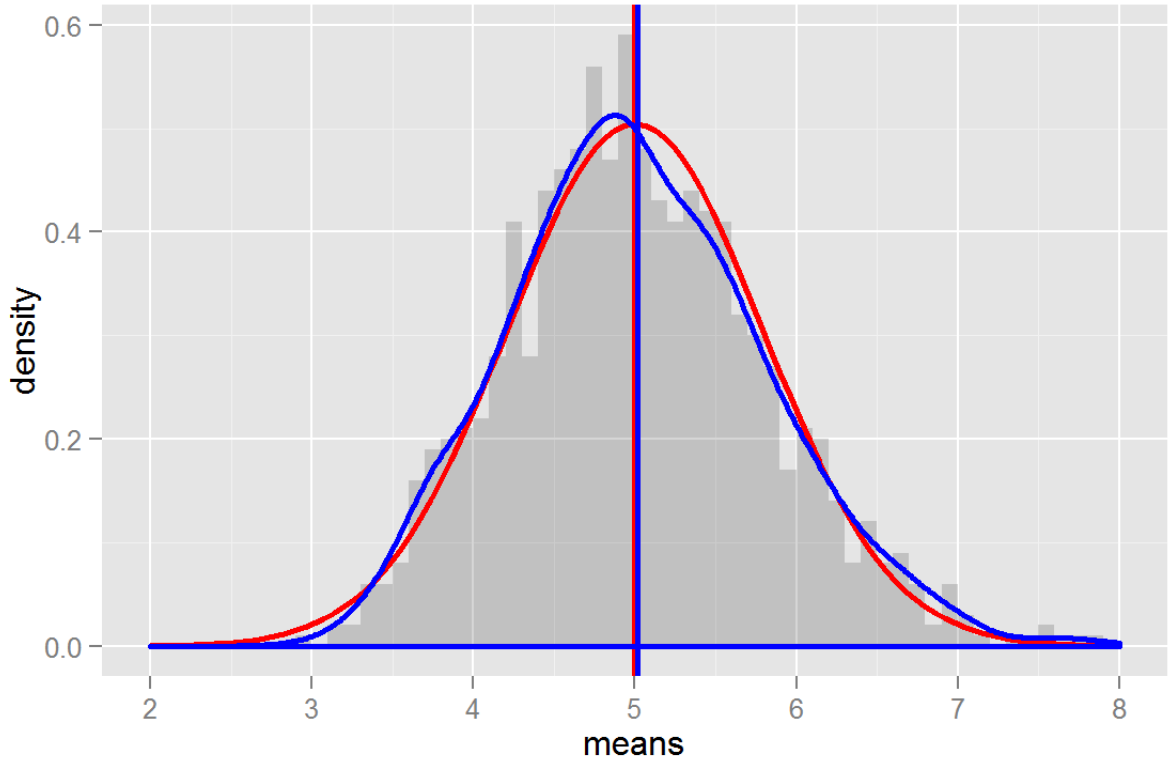
Theoritical and observed variances are also very close.

3. Show that the distribution is approximately normal.

To highlight that we :

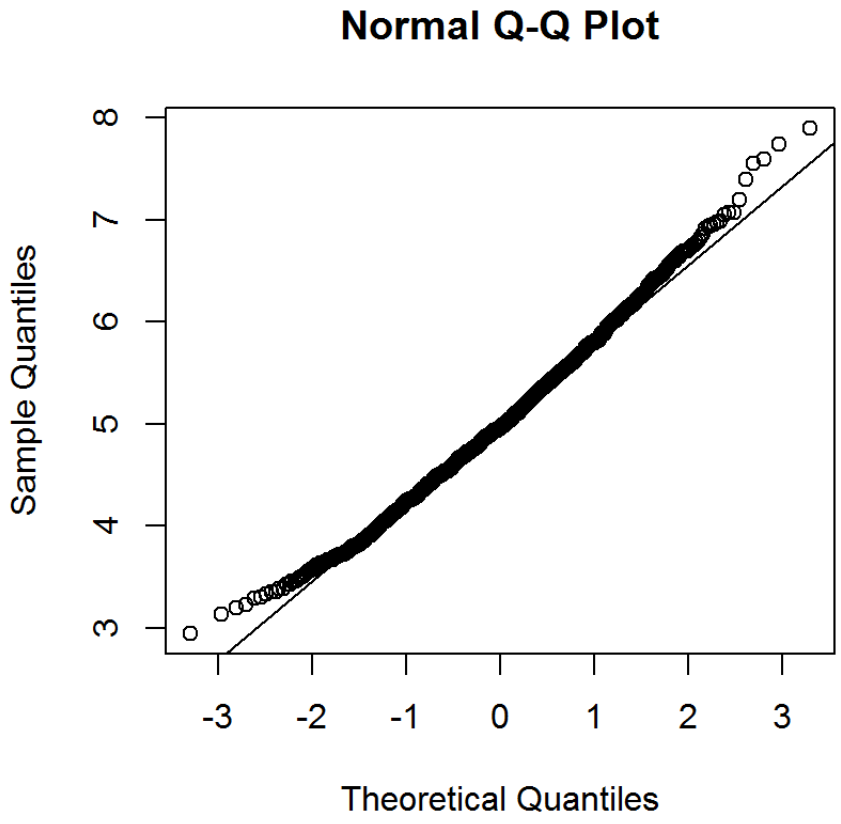
- Create an approximate normal distribution and see how sample data aligns with it.
- Make the QQ-plot for quantiles

```
# plot the means
ggplot(data = exponentialDistriMeans, aes(x = means)) +
  geom_histogram(binwidth=0.1, aes(y=..density..), alpha=0.2) +
  stat_function(fun = dnorm, arg = list(mean = mu , sd = standard_deviation_theory), colo
ur = "red", size=1) +
  geom_vline(xintercept = mu, size=1, colour="#FF0000") +
  geom_density(colour="blue", size=1) +
  geom_vline(xintercept = meanOfMeans, size=1, colour="#0000FF") +
  scale_x_continuous(breaks=seq(mu-3,mu+3,1), limits=c(mu-3,mu+3))
```



QQ-plot below suggests also the normality.

```
qqnorm(exponentialDistriMeans$means)
qqline(exponentialDistriMeans$means)
```



As you can see from figures, the calculated distribution of means of random sampled exponantial distributions overlaps quite nice with the normal distribution of expected values based on the given λ.

Appendix

Additionnal resources

- jbstatistics video - Introduction to the Central Limit Theorem (YouTube) (https://www.youtube.com/watch?v=PujoI1yC1_A)

Tips

To disable warnings and messages for Knit PDF rendering :

```
library(knitr)
opts_chunk$set(fig.width=7, fig.height=4, warning=FALSE, message=FALSE)
```