

Statistical inference Assignment

By Bruno Berrehuel, 10th november 2016

Part 1 : Exponential distribution

Overview

In this part I'll do some investigations of the exponential distribution, with the R `rexp(n, λ)` function, and more precisely the distribution of averages of 40 exponentials distributions. In accordance with the Central Limit Theorem, I'll investigate if the mean μ is equal to $\frac{1}{\lambda}$ and the variance of the distribution is equal to $\frac{\sigma^2}{n}$. For numerical purposes, λ is choosen as $\lambda = \frac{1}{5}$, so $\mu = \sigma = \frac{1}{\lambda} = 5$.

Then I'll look at the distribution of 10 000 means of 40 exponential distributions, and verify that they are normal distribution $\mathcal{N}(\frac{1}{\lambda}, \frac{\sigma^2}{n})$, as expected with the CLT.

Simulations

Calculate 10 000 means of 40 exponentials distributions with $\lambda = \frac{1}{5}$: for each i, calculate the mean of 40 exponentials distributions, then add it to the variables `rexpMean` and `rexpVar`. The `rexpMean` and `rexpVar` variables will contain 10 000 values each.

```
set.seed(12345)
rexpMean=NULL
for (i in 1:10000) {
  rexpMean = c(rexpMean, mean(rexp(40,0.2)))
  rexpVar = c(rexpVar, var(rexp(40,0.2)))
}
```

Sample mean and variance vs. theoreticals

Calculate the means of the mean and the variance of the sample with R, and compare with mean and variance expected. Recall that the expected sample mean $\mu = \frac{1}{\lambda} = 5$ and the expected sample variance $s = \frac{\sigma^2}{n} = \frac{5^2}{40} \simeq 0.625$. The expected distribution variance is $\sigma^2 = 25$.

```
data.frame(mean=mean(rexpMean), sampleVariance=var(rexpMean), variance=mean(rexpVar))
```

```
##      mean sampleVariance variance
## 1 5.003857      0.6083614 24.92046
```

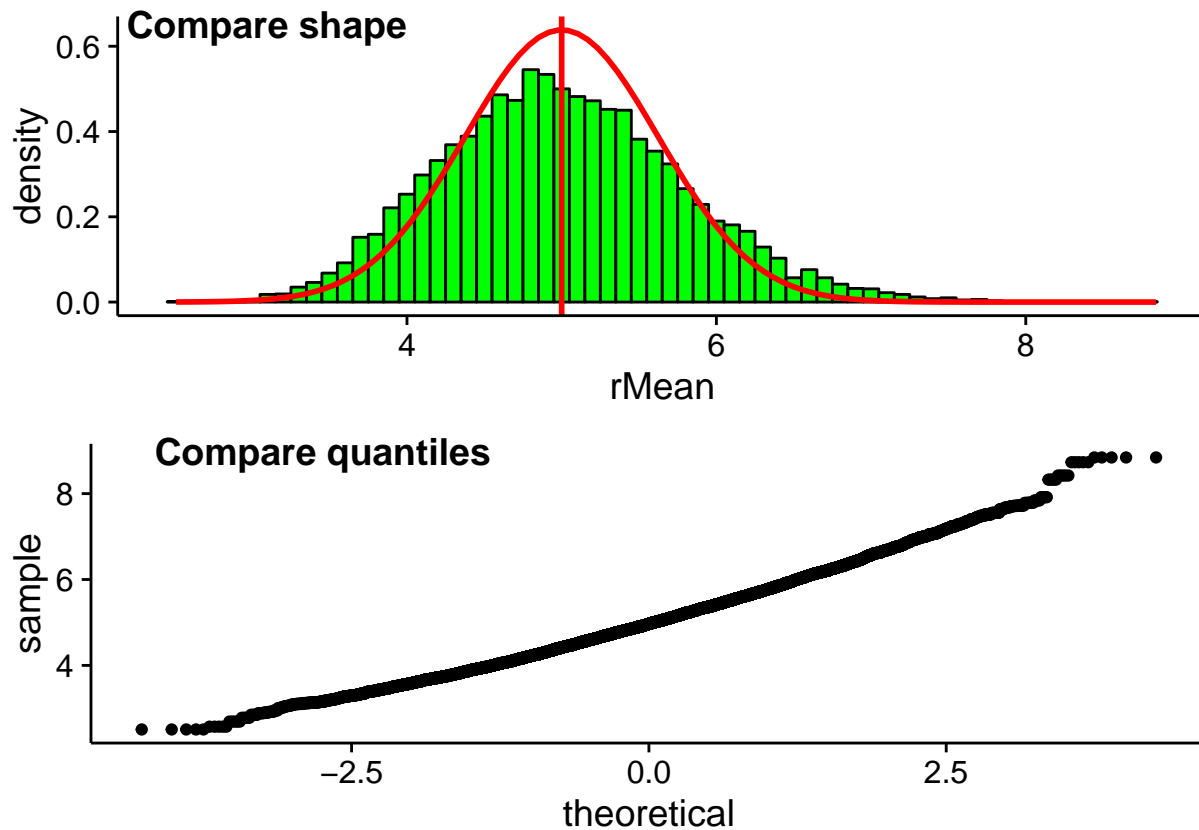
The values are very close to the expected values : 0.08% for the sample mean, 2.66% for the sample variance, and 0.32% for the distribution variance.

Distribution

Plot an histogram with the 10 000 previous calculation, and compare the shape with a normal distribution $\mathcal{N}(5, 0.625)$, then compare the distributions using a quantile-quantile diagram :

```
library(ggplot2)
library(cowplot)
dfRexp <- data.frame(rMean=rexpMean, rVar=rexpVar)
par(mfrow=c(1,2))
histoMean <- ggplot(dfRexp, aes(x=rMean)) +
  geom_histogram(aes(y=..density..), color="black", fill="green", binwidth=0.1) +
  geom_vline(aes(xintercept=5), color="red", size=1) +
  stat_function(fun=dnorm, color="red", size=1,
    args=list(mean=5, sd=0.625))
qqMean <- ggplot(dfRexp, aes(sample=rMean)) + stat_qq()

plot_grid(histoMean, qqMean, labels=c("Compare shape", "Compare quantiles"), ncol=1, nrow=2)
```



Conclusion

As points are aligned in the qqnorm diagram and with the shape of the histogram, the distribution of the mean follows the normal distribution $\mathcal{N}(5, 0.625)$, as expected with the CLT.

Part 2 : ToothGrowth Data Analysis
