Evaluation of the Central Limit Theorem

A study of the distribution of means from simulated exponential distributions

Diego Rabatone Oliveira

Overview

We will do some simulations to investigate the exponencial distribution and compare these simulations to the expected theoretical exponencial distribution, using the **Central Limite Theorem** (CLT).

The exponential distribution can be simulated in R with `rexp(n,λ)` where `n` is the number of observations and 'λ' is the rate parameter. The mean of exponential distribution is $\mu = \frac{1}{\lambda}$ and the standard deviation is also $\sigma = \frac{1}{\lambda}$. For this project we will use $\lambda = 0.2$ on all simulations, and then we will investigate the distribution of the averages from 40 exponentials. This process we will be done a thousand times, in order to check the validity of the **CLT**.

The simulations

Firstly let's set our main variables:

```
# Setting parameters
set.seed(1985)  # setting a seed in order to be able to reproduce the results
nsim <- 1000  # number of simulations
lambda <- 0.2
n <- 40
mu <- 1/lambda  # Theoretical mean
sigma <- 1/lambda  # Theoretical standard deviation
variance <- sigma^2  # Theoretical variance
```

Now we create an empty dataframe with 42 'columns', being the first the variable `MEAN`, the second 'VAR' and the followings will receive the 40 observations of each simulation, in a way that our dataframe will have one line per simulation.

```
sims <- data.frame('MEAN'=NA, 'VAR'=NA, as.list(numeric(n)))
```

For our simulations, first we store the simulated values on the datafarme, passing NA as the "current mean" and variance, and then we calculate the effective mean and variance of each simulation.

```
for (sim in 1:nsim) {
  # Producing the simulations and storing each simulation as one line on the df
  sims[sim,] <- c(NA, NA, rexp(n, lambda))
  # Calculating the mean for the given simulation and storing it on the MEAN variable.
  sims[sim,'MEAN'] <- sims[sim,] %>% select(-MEAN,-VAR) %>% apply(1, FUN = mean)
  sims[sim,'VAR'] <- sims[sim,] %>% select(-MEAN,-VAR) %>% apply(1, FUN = var)
}
```

Let's see, as an example, the values from our first simulation.

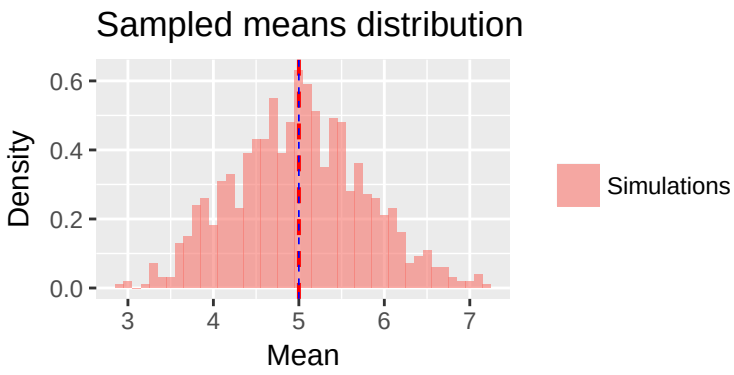
MEAN	VAR	X0	X0.1	X0.2	X0.3	X0.4	X0.5	X0.6	X0.7	X0.8	X0.9	X0.10
4.78	35.9	1.64	3.77	7.25	6.79	5.99	7.93	0.568	5.3	4.46	6.89	2.87

Sample Mean x Theoretical Mean

For a exponential distribution with $\lambda = 0.2$, the expected mean is $\mu = \frac{1}{\lambda} = 5$. From our simulations, the average mean found was $mean(sims\$MEAN) = 5.001$.

Let's see a plot with the distribution of the calculated means from the simulations:

```
ggplot(data=sims, aes(x=MEAN)) +
  ggtitle('Sampled means distribution') +
  geom_histogram(aes(y=..density.., fill='Simulations'), binwidth=0.1, alpha=0.6) +
  geom_vline(xintercept=mean(sims$MEAN), linetype="dashed", col='red', size=0.7) +
  geom_vline(xintercept=mu, linetype="dashed", col='blue', size=0.3) +
  scale_x_continuous(breaks=round(seq(min(sims$MEAN), max(sims$MEAN), by=1))) +
  xlab('Mean') + ylab('Density') + theme(legend.title=element_blank())
```



From the plot we can notice that the blue vertical line, that represents the theoretical mean, is right over the dashed red line, which is the mean from our simulations, showing how close those two means are.

Verifying the simulated means with a T Test:

```
t.test(sims$MEAN, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  sims$MEAN
## t = 200, df = 1000, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.95 5.05
## sample estimates:
## mean of x
##         5
```

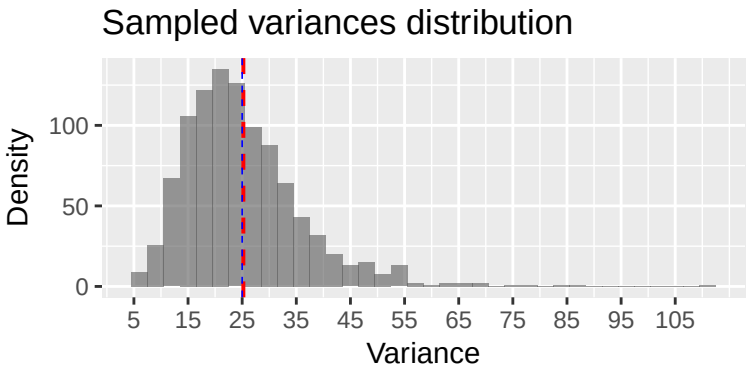
We can observe that, with a 95% confidence interval, our p-value is equal to 0 , so, we are able to state that our simulated means are equal to 5, which is our expected mean.

Sample Variance x Theoretical Variance

Our overall variance can be calculated as the mean of the variances, which is $mean(sims\$VAR) = 25.213$. The theoretical expected variance for the exponential distribution with $\lambda = 0.2$ is $(\frac{1}{\lambda})^2 = 25$. So, we can observe that both variances are quite close one to each other.

Now we can plot the variance for each simulation:

```
ggplot(data=sims, aes(x=VAR)) +
  ggtitle('Sampled variances distribution') +
  geom_histogram(binwidth=3, alpha=0.6) +
  geom_vline(xintercept=mean(sims$VAR), linetype="dashed", col='red', size=0.7) +
  geom_vline(xintercept=variance, linetype="dashed", col='blue', size=0.3) +
  scale_x_continuous(breaks=round(seq(min(sims$VAR), max(sims$VAR), by=10))) +
  xlab('Variance') + ylab('Density') + theme(legend.title=element_blank())
```



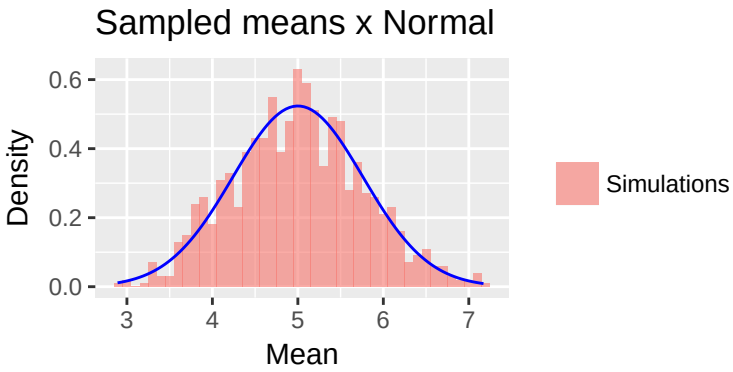
And, from the plot, we can observe how the variance is centered on the expected variance (blue dashed line), and also how close the mean variance of our trial is to the expected variance.

As we would expect from the **CLT**, if we did more simiulations our variance would get even closer to the expected one.

Distribution

Now we evaluate the behaviour of the distribution, specially when compared to the normal distribution:

```
ggplot(data=sims, aes(x=MEAN)) +
  ggtitle('Sampled means x Normal') +
  geom_histogram(aes(y=..density.., fill='Simulations'), binwidth=0.1, alpha=0.6) +
  stat_function(fun=dnorm, color="blue", args=list(mean=mean(sims$MEAN), sd=sd(sims$MEAN))) +
  scale_x_continuous(breaks=round(seq(min(sims$MEAN), max(sims$MEAN), by=1))) +
  xlab('Mean') + ylab('Density') + theme(legend.title=element_blank())
```



By plotting the distribution of the calculated means (red histogram) against the normal distribution (blue line) with mean μ and standard deviation σ, we can notice that the distribution of the calculated means is close to the normal distribution, as we would expect based on the **Central Limit Theorem**. If more simulations were done, then the simulated means distribution would look even more to the theoretical expected result.

This report can be found at: https://github.com/diraol/CourseraProjects