Comparing exponential distribution versus CLT

Overview

In this report we will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. The explanation will use the properties of Central Limit Theorem.

Simulations

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 $1/lambda$. We set $lambda = 0.2$ for all of the simulations. We will investigate the distribution of averages of **40** exponentials. We will perform **1000** simulations. The following code performs such simulations and calculates mean and standard deviation for each simulation:

```
# setup
library(ggplot2)
set.seed(1337)
lambda = 0.2 ; exp_mn = 1/lambda ; exp_sd = 1/lambda
nosim = 1000 ; noavg = 40
# simulations
sims = matrix(rexp(noavg*nosim, lambda), ncol=nosim)
# calc. statistics
mns = apply(sims, 2, mean)
sds = apply(sims, 2, sd)
# arrangement for display purposes
df = data.frame(means=mns, variances=sds^2)
```

Sample Mean versus Theoretical Mean

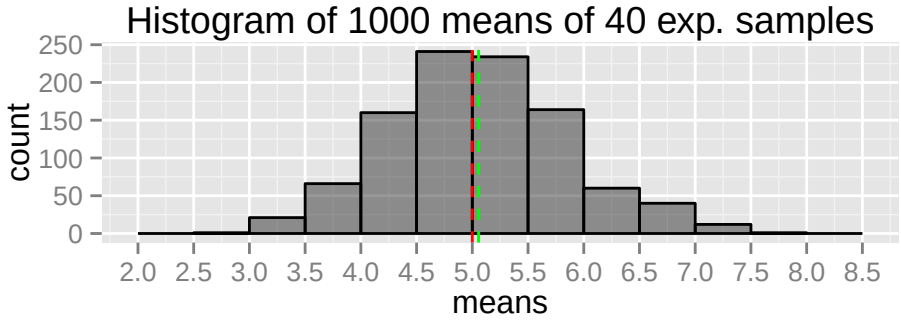
The following code presents a mean of simulations means, a mean of all simulations (the same as previous) and theoretical mean. As one can notice, they are nearly the same:

```
print(c(mean(mns), mean(sims), exp_mn))
```

```
## [1] 5.055995 5.055995 5.000000
```

The following figure presents a histogram of means. On the histogram are two lines. The red indicates a value of theoretical mean. The green indicates a value of a simulation mean. As one can notice, these two values are very close to each other:

```
g = ggplot(df, aes(x = means)) +
    geom_histogram(alpha = .50, binwidth=0.5, colour = "black" ) +
    scale_x_continuous(breaks = round(seq(min(df$means)-1, max(df$means)+1, by = 0.5),1)) +
    labs(title = paste("Histogram of", nosim, "means of", noavg, "exp. samples")) +
    geom_vline(aes(xintercept=exp_mn), colour="red", linetype="dashed") +
    geom_vline(aes(xintercept=mean(mns)), colour="green", linetype="dashed")
print(g)
```



Sample Variance versus Theoretical Variance

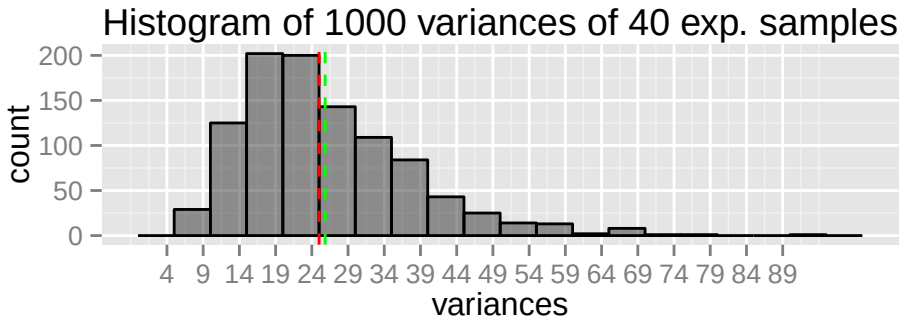
The following code presents a variance of simulations variances (for information purpose only), a variance of all simulations and theoretical variance. As one can notice, the theoretical and simulation variances are nearly the same:

```
print(c(sd(sds)^2, sd(sims)^2, (exp_sd)^2))
```

```
## [1]  1.200613 25.849170 25.000000
```

For visualization purposes, the following figure presents a histogram of variances. On the **skewed** histogram are two lines. The red indicates a value of theoretical variance. The green indicates **the mean** value of a simulation variances. As one can notice, these two values are very close to each other, but the spread is bigger than in case of means:

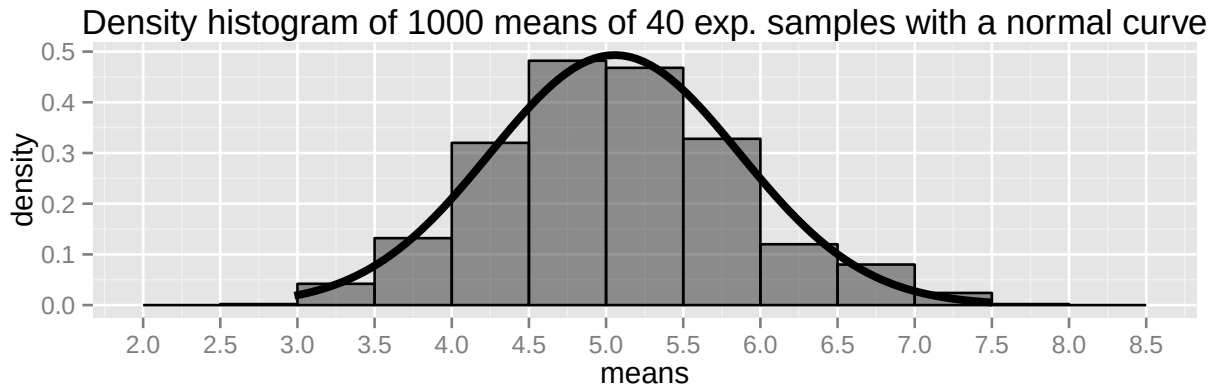
```
g = ggplot(df, aes(x = variances)) +
    geom_histogram(alpha = .50, binwidth=5, colour = "black" ) +
    scale_x_continuous(breaks = round(seq(round(min(df$variances))-1,
       round(max(df$variances))+1, by = 5),1)) +
    labs(title = paste("Histogram of", nosim, "variances of", noavg, "exp. samples")) +
    geom_vline(aes(xintercept=exp_sd^2), colour="red", linetype="dashed") +
    geom_vline(aes(xintercept=mean(sds^2)), colour="green", linetype="dashed")
print(g)
```



Distribution

The following figure presents the density histogram of means of the simulations. One can notice it has a normal *bell*-like shape:

```
g = ggplot(df, aes(x = means)) +
    geom_histogram(alpha = .50, binwidth=0.5, colour = "black", aes(y = ..density..)) +
    scale_x_continuous(breaks = round(seq(min(df$means)-1, max(df$means)+1, by = 0.5),1)) +
    labs(title = paste("Density histogram of", nosim, "means of",
        noavg, "exp. samples with a normal curve")) +
    stat_function(fun = dnorm, size=1.5, args=list(mean=mean(mns), sd=sd(mns)))
print(g)
```



The distribution of means nearly follows the **68–95–99.7 rule**, where:

- 68.0% of the data are within 1 standard deviation of the mean,
- 95.0% of the data are within 2 standard deviation of the mean,
- 99.7% of the data are within 3 standard deviation of the mean:

```
Mu = mean(df$means)
Sd = sd(df$means)

for (X in c(1,2,3))
{
    print(length(df[df$means < Mu + X*Sd & df$means > Mu - X*Sd, "means"])/nosim)
}
```

```
## [1] 0.688
## [1] 0.942
## [1] 0.999
```

For these reasons, one can tell the distribution is approximately normal.