SI - Part 1: Simulation

Pascal P

11 August 2018

1 Overview

The aim of this project is to investigate the exponential distribution in `R` and compare it against the Central Limit Theorem (abbreviated as `CLT` in this document).
After running the simulation (`1000` samples of `40` exponential) and computing the relevant statistics (sample mean and sample variance) I will define two plots (a histogram of sample mean, and a cumulative variance) to show how sample mean and variance approximate their theoretical counterparts. Then I will assess the normality of our distribution by showing (i) a histogram with the normal standard overlayed and (ii) a quantile to quantile plot with reference line.

2 Simulations:

In `R`, we can simulate the exponential distribution with `rexp(n, lambda)`. Lambda (λ) is the rate parameter.
Both the theoretical mean and the theoretical standard deviation of the exponential distribution equal to $1/\lambda$.
In this simulation, λ is set to `0.2` and we use a distribution of averages of `40` exponentials for a total of `1000` simulations.
The following table summarized the results on our simulation.

Table 1: Summary

stats	experimental	theoretical
Mean	5.019	5.000
Variance	0.642	0.625

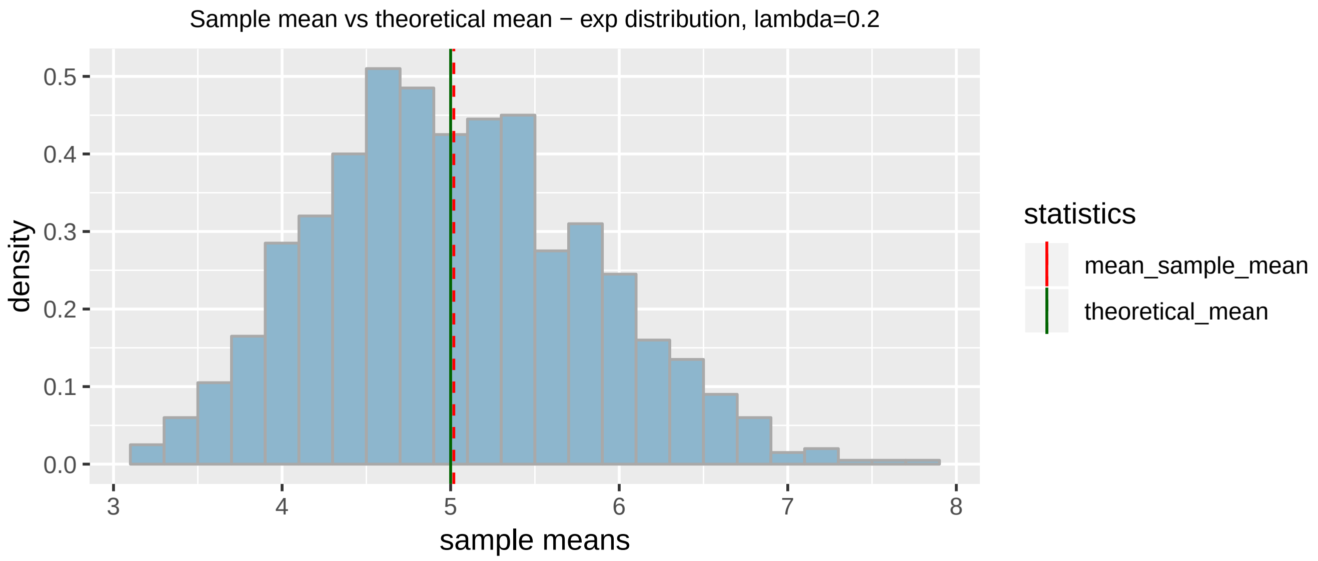
2.1 Sample Mean versus Theoretical Mean:

Our sample mean is as follows (it is also stated rounded to 3 decimal in the summary table, above).

```
mean_sample_mean
```

```
## [1] 5.018775
```

The sample mean (cf. above) is closed to the theoretical one ($1/\lambda = 5$), as expected from the *Law of Large Numbers*. We can visualize this with the following plot:



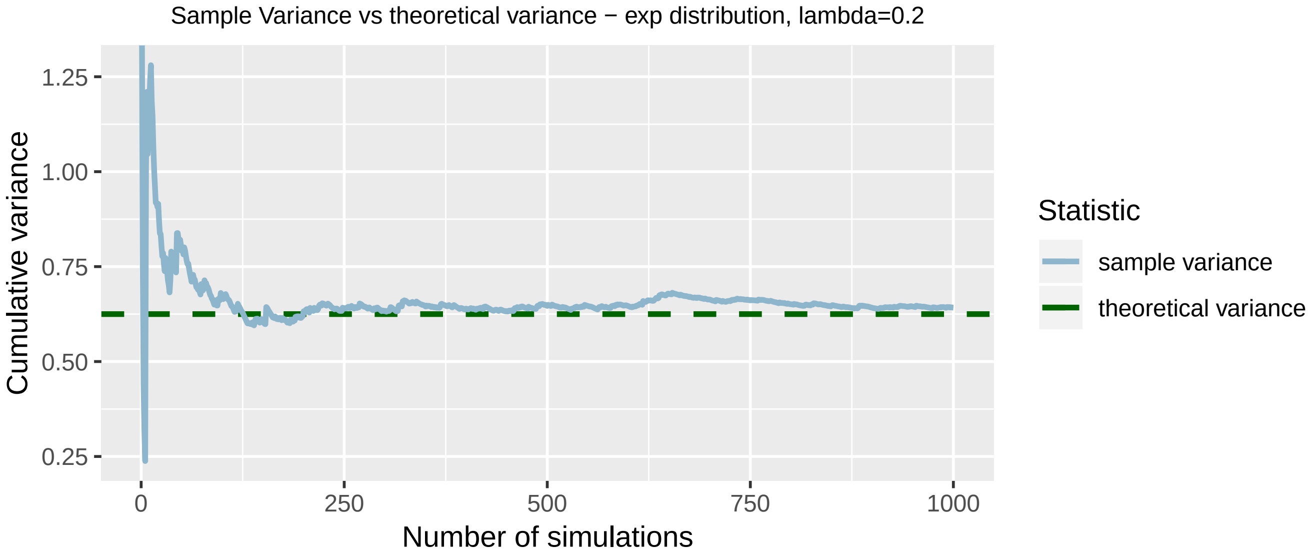
2.2 Sample Variance versus Theoretical Variance

We expect our sample variance to be closed to the theoretical one as predicted by `CLT`.
Our summary table (cf. paragraph 2 Simulation) shows that this is the case. The value is as follows:

```
var_sample_mean
```

```
## [1] 0.6419993
```

We can visualize this with the following plot, notice how as the number of iteration grows, the sample variance gets closer to the theoretical one.

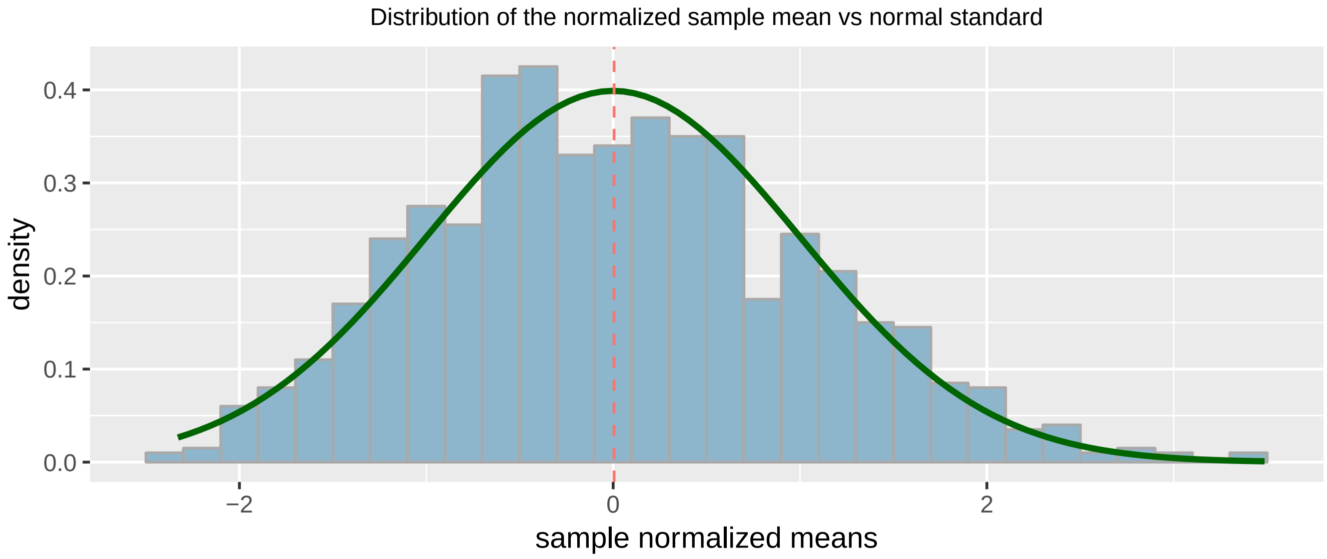


2.3 Distribution

The `CLT` states that the distribution of averages of independent identically distributed (`iid`) variables (provided these are properly normalized) has a distribution like that of a standard normal for large (enough) n. In order to show that, we are going to normalize each sample mean with the following formula:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n} \times (\bar{X}_n - \mu)}{\sigma}$$

Our normalized distribution should be a normal distribution with a mean close to zero.



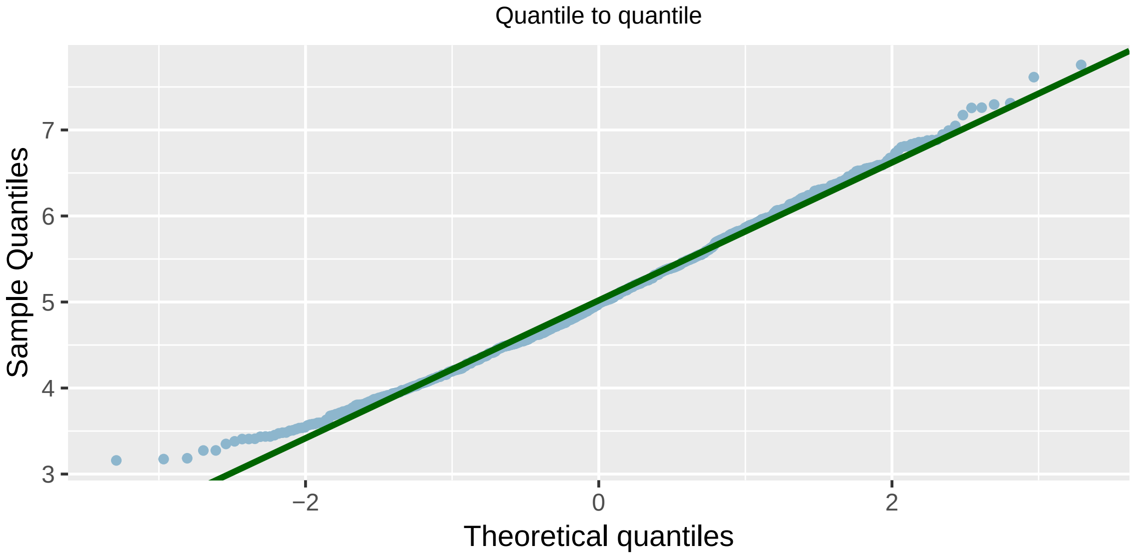
Our new normalized mean is:

```
norm_mean_sample_mean
```

```
## [1] 0.004749825
```

which is indeed close to `0`.

Another way to show that our distribution of average approximates the standard normal (our `n_sim = 1000` is large enough), is to use the quantile to quantile plot. The following figure shows how our sample distribution lies along the (straight diagonal) reference line from normal standard distribution. This indicates that our sample distribution is indeed normal.



3 Conclusion

From the four previous plots (cf section "Simulation:), we can see that:

- both our sample mean and sample variance approximate the theoretical mean and theoretical variance and
- the distribution of the `1000` means (of these `40` exponential) is quasi normal as stated by `CLT`.

4 Appendix

4.1 Code

All the supporting code is listed in this section.

4.2 Preparation

```
n_samp  <- 40                      # num. of sample
n_sim   <- 1000                    # num. of simulation
lambda  <-  1 / 5                  # rate
th_mean <- 1 / lambda              # theoretical mean
th_var  <- 1 / (lambda^2 * n_samp) # theoretical variance

set.seed(20180810)

# define the matrix [1000, 40] containing all the simulations
mx <- matrix(rexp(n_sim * n_samp, rate=lambda), nrow=n_sim, ncol=n_samp)

df <- data.frame(sample_mean = apply(mx, 1, mean))
mean_sample_mean <- mean(df$sample_mean)
var_sample_mean <- var(df$sample_mean)

# summary table
stats    <- c("Mean", "Variance")
experimental <- c(round(mean_sample_mean, 3), round(var_sample_mean, 3))
theoretical  <- c(th_mean, th_var)
sum_df <- data.frame(stats, experimental, theoretical)
rownames(sum_df) <- NULL

knitr::kable(sum_df, caption='Summary')
```

4.3 Sample mean plot

```
library(ggplot2)

ggplot(df, aes(x=sample_mean, fill=n_samp)) +
  geom_histogram(binwidth=.2, color="darkgrey", aes(y=..density..), fill="lightskyblue3") +
  geom_vline(aes(xintercept=mean_sample_mean, color="mean_sample_mean"),
             linetype="dashed", size=0.5, show.legend=TRUE) +
  geom_vline(aes(xintercept=th_mean, color="theoretical_mean"), size=0.5, show.legend=TRUE) +
  xlab("sample means") +
  ggtitle(paste0("Sample mean vs theoretical mean - exp distribution, ",
                 expression(lambda), "=", lambda)) +
  scale_color_manual(name="statistics",
                      values=c(mean_sample_mean="red", theoretical_mean="darkgreen")) +
  theme(plot.title=element_text(hjust = 0.5, size=9))
```

4.4 Variance plot

```
cum_var <- cumsum((df$sample_mean - mean_sample_mean)^2) / (seq_along(df$sample_mean) - 1)

ggplot(data.frame(x=1:n_sim, y=cum_var)) +
  geom_hline(aes(yintercept=th_var, color="theoretical_variance"),
           size=1, linetype="dashed", show.legend=TRUE) +
  geom_line(aes(x=x, y=y, color="sample_variance"),
           size=1, show.legend=TRUE) +
  scale_color_manual(name="Statistic",
                     labels=c("sample variance", "theoretical variance"),
                     values=c(sample_variance="lightskyblue3", theoretical_variance="darkgreen")) +
  labs(x="Number of simulations",
       y="Cumulative variance") +
  ggtitle(paste0("Sample Variance vs theoretical variance - exp distribution, ",
                 expression(lambda), '=', lambda)) +
  theme(plot.title=element_text(hjust = 0.5, size=9)
```

4.5 Distribution

```
# normalization function
norm_fun <- function(xbar, n=n_sim, thmean=th_mean, thvar=th_var, lam=lambda) {
  sqrt(n) * (xbar - thmean) / (n * thvar * lam)
}

# create a new column in data  frame for normalized sample mean
df$norm_mx <- apply(df, 1, norm_fun, n_samp)
norm_mean_sample_mean <- norm_fun(mean_sample_mean)

ggplot(df, aes(x=norm_mx, fill=n_samp)) +
  geom_histogram(binwidth=.2, color="darkgrey", aes(y=..density..), fill="lightskyblue3") +
  stat_function(fun=dnorm, size=1.1, color="darkgreen") +
  geom_vline(aes(xintercept=norm_mean_sample_mean, color="norm_mean_sample_mean"),
             linetype="dashed", size=0.5, show.legend=FALSE) +
  xlab("sample normalized means") +
  ggtitle("Distribution of the normalized sample mean vs normal standard")  +
  theme(plot.title=element_text(hjust = 0.5, size=9))
```

```
ggplot(df, aes(sample=sample_mean)) +
  stat_qq(color="lightskyblue3", size=1.2) +
  # define reference line
  geom_abline(intercept=mean(df$sample_mean),
              slope=sd(df$sample_mean), color="darkgreen", size=1.1) +
  labs(x="Theoretical quantiles", y="Sample Quantiles") +
  ggtitle("Quantile to quantile") +
  theme(plot.title=element_text(hjust = 0.5, size=9))
```

4.6 References

- The code for the summary table was suggested by Len Greski, discussion: https://www.coursera.org/learn/statistical-inference/discussions/weeks/4/threads/eIsbPhVTEeaxsBIihWq87Q The related snippet is available from: https://github.com/lgreski/datasciencectacontent/blob/master/markdown/kableDataFrameTable.md
- The code for the reference line in the quantile to quantile plot (with ggplot2) is taken from the following: https://stackoverflow.com/questions/4357031/qqnorm-and-qqline-in-ggplot2
- Coursera “Statistical Inference” course, and in particular week 3.