STI-Project 1. Simulation to compare the exponential distribution in R with the Central Limit Theorem

Gerrit Versteeg

27 November 2016

1. Synopsis(Overview)

In this report we aim to compare the exponential distribution (in R) with the Central Limit Theorem by simulating 1000 samples of 40 exponentials using R and researched the resulting distribution, that should look like a normal distribution. The resulting sample mean was 5.011 which is pretty close to the theoretical mean of $1/\lambda = 5$, while the resulting variance was 0.622 also very close to the theoretical variance of $1/\lambda^2/n = 25/40 = 0.625$. Checks on the 'normality' of the sample distribution (SW-test p-value: 0.00045) ffound that the sample mean distribution was indeed approximately 'normal'

2. Simulations

First step is to setup a vector 'smeans' containing 1000 sample means of 40 exponentials. The sample means are derived by calculating the mean of each sample consisting of 40 random variables drawn from the R-exponential distribution (where lambda = 0.2). All sample means are collected in the vector 'smeans'. To be able to compare the exponential distribution with the hopefully normal distribution of means, a second vector 'exps' is used to store the 1000 exponentials with the same seed. The two vectors are combined into a data frame because ggplot works with data frames.

```r
library("ggplot2", warn.conflicts=FALSE)  ## for the graphs
n <- 40                         ## setting sample size
nrsim <- 1000                   ## setting number of samples
lambda <- 0.2                   ## setting lambda of exp. function
set.seed(140958)                ## to enable reproduction
exps <- rexp(nrsim, lambda)     ## fill exps for comparison
smeans <- NULL                  ## prepare vector for sample means
for (i in 1:1000)               ## generate 1000 means of 40 exp's
        smeans = c(smeans, mean(rexp(n,lambda)))
dfsim <- as.data.frame(cbind(exps, smeans))
                                ## combine exponentials and sample
                                ## means into a single data frame
aM <- mean(smeans)              ## store mean of the sample means
aV <- var(smeans)               ## store variance of sample means
```

Just to get an impression of the resulting distribution, we depict the resulting sample distribution and the original exponential distribution of 1000 exponentials (with the same seed). The R-code and figures are listed in the appendix (figure 1 & 2). Looking at both histograms, clearly the sample means are distributed rather 'normally' especially compared to the exponential distribution, that is heavily skewed to the right.

3. Sample Mean versus Theoretical Mean

The Central Limit Theorem states the mean of the sample means should be equal to the mean of the underlying dsitribution (population) drawn samples from, when the sample size (n) is large enough. The mean of the 1000 sample means should therefor be theoretically equal to $1/\lambda = 1/0.2 = 5$.

Lets look at the sample distribution again and for comparison draw vertical lines for the theoretical mean and the mean actually found in the sample. The R-code and resulting figures can be found in the appendix (figure 3). The actual means of the 1000 sample means is represented by the green solid line, while the theoretical means based on CLT is shown by the dashed, red line.

Clearly the actual observed sample mean (green line) is almost at the same position as the theoretical mean ($1/\lambda = 5$) shown by the red line.

The actual value of the mean of sample means = 5.0108359, which is pretty close to the theoretical mean of $1/\lambda = 5$

5. Sample Variance versus Theoretical Variance

The Central Limit Theory also claims that the variance of sample means should be the same as the variance of the underlying distribution divided by the sample size. In other words: $Var_{sample} = Var_{pop}/n$. The variance of the exponential distribution function is its standard deviation ($\sigma = 1/\lambda$) squared. So what we expect from the CLT is that for our simulation the theoretical variance would be:

$Var_{sample} = 1/\lambda^2/n = 25/40 = 0.625$.

The actual variance of the 1000 sample means is 0.6219065. So this result is also pretty close.

In the appendix a plot of the means is provided (figure 4), with vertical lines for the actual variance (in green) as well as the theoretical variance (in red). The lines represent the area: sample mean +/- (variance * sample mean). Please note, we are looking at variances instead of the usual standard deviations (standard errors of the mean).

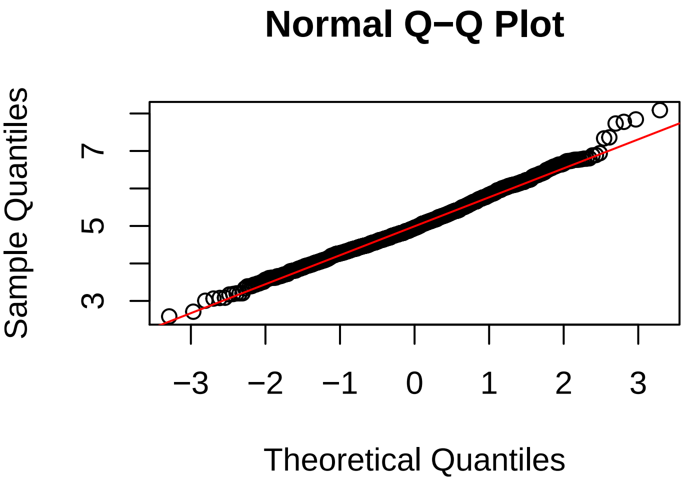
Looking at that graph, one can clearly see that the theoretical variance of a sample mean of 40 exponentials (red lines) is almost the same as the actual variance of the 1000 sample means of 40 exponentials (green lines).

The actual value of the variance of sample means = 0.6219065, which is pretty close to the theoretical variance of $1/\lambda^2/n = 0.625$

6. Distribution

We have seen that mean and variance of our 1000 sample means with size 40, are indeed almost the same as what we expected from their theoretical values. As a final step we need to check whether the distribution of the 1000 sample means approximates the normal distribution. We have already seen that the histogram of the 1000 sample means (figure 2 in the appendix) looks pretty 'normally' distributed. But to be more formal about it, we use a QQ-plot. This plot will compare the quantiles of our actual distribution ('sample quantiles') with the normal distribution ('theoretical quantiles'). If the quantiles are in a nice straight line, with limited outliers, the sample distribution approximates a normal distribution.

```r
qqnorm(smeans); qqline(smeans, col = 2) ## plot the quantiles
```



The resulting graph shows a pretty straight line through (0,5) representing the 50% quantile (0 for normal and 5 sample) and looks to be okay, no skewing except for the far ends.

To be more sure, there is a variety of tests we can use to measure the 'normality' of our sample means. Although there is much discussion on the usability of these test, the one most often used is the 'Shapiro-Wilk' test.

```
shapiro.test(smeans)      ## perform a Shapiro-Wilk test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smeans
## W = 0.99394, p-value = 0.0004492
```

The resulting p-value is very small. That means we consider the distribution to be normal (H0) within usual significance levels (like alpha = 0.05). Another indication that our sample distribution indeed apporxinates the normal distribution.

APPENDIX

Figure 1. Distribution of the exponentials

```
g <- ggplot(dfsim,
        aes(x=dfsim$exps))            ## setup graphic object
g + geom_histogram(                   ## plot histogram bars
        fill=I("White"),              ## fill with white
        col=I("Black"),               ## outline with black
        bins=20,                      ## appropriate bins
        aes(y=..density..)) +         ## set y to density
        xlab("rexp(1000, 0.2") +      ## label X-axis
        ylab("Frequency") +           ## label Y-axis
        ggtitle("Histogram of Exponential distribution") +
    geom_density (aes(y=..density..)) ## plot density-line
```

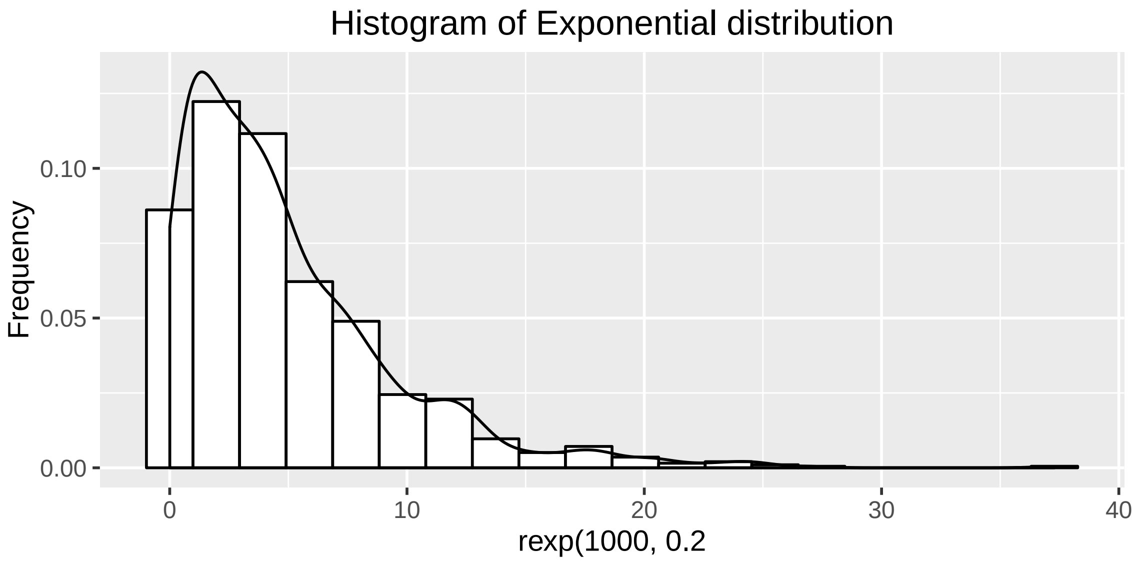


Figure 2. Distribution of the sample means

```
m <- ggplot(dfsim,
        aes(x=dfsim$smeans))             ## setup graphic object
m + geom_histogram(                      ## plot histogram bars
        fill=I("White"),                 ## fill with white
        col=I("Black"),                  ## outline with black
        bins=20,                         ## appropriate bins
        aes(y=..density..)) +            ## set y to density
        xlab("sample means") +           ## label X-axis
        ylab("Frequency") +              ## label Y-axis
        ggtitle("Histogram of sample means distribution") +
     geom_density (aes(y=..density..))   ## plot density-line
```

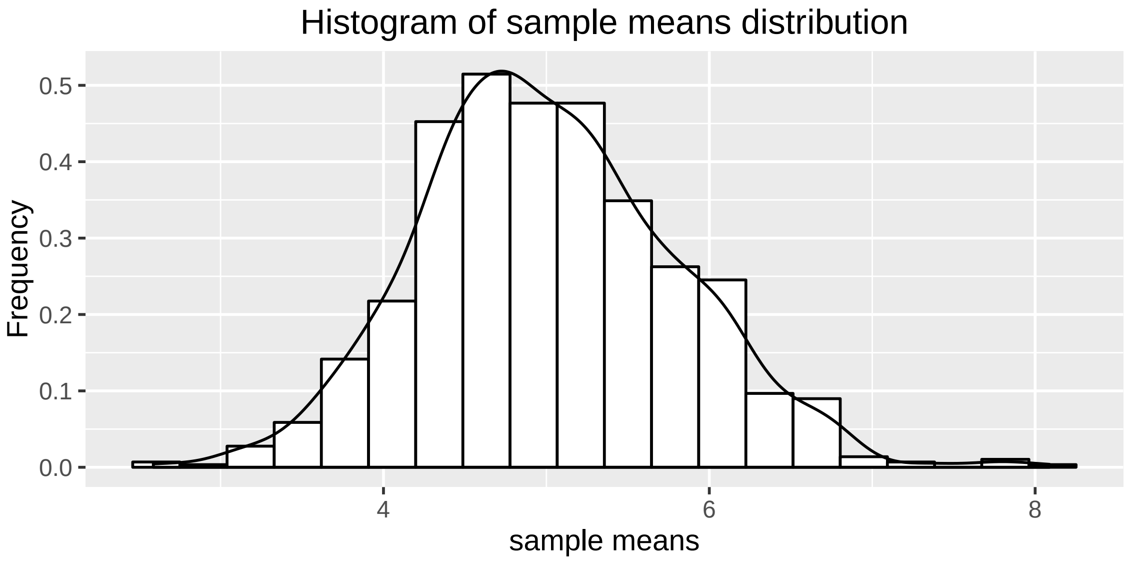


Figure 3. Comparing means: Actual sample (green) vs. Theoretical (red)

```
m <- ggplot(dfsim,
        aes(x=dfsim$smeans))            ## setup graphic object
m + geom_histogram(                     ## plot histogram bars
        fill=I("White"),                ## fill with white
        col=I("Black"),                 ## outline with black
        bins=20,                        ## appropriate bins
        aes(y=..density..)) +           ## set y to density
        xlab("sample means") +          ## label X-axis
        ylab("Frequency") +             ## label Y-axis
        ggtitle("Comparing means: actual sample (green) vs. theoretical (red)") +
    geom_density (aes(y=..density..)) + ## plot density-line
    geom_vline (xintercept = aM,        ## plot actual mean of samples
        color = "Green") +
    geom_vline (xintercept = 1/lambda,  ## plot theor. mean of samples
        color = "Red",
        linetype = "longdash")
```

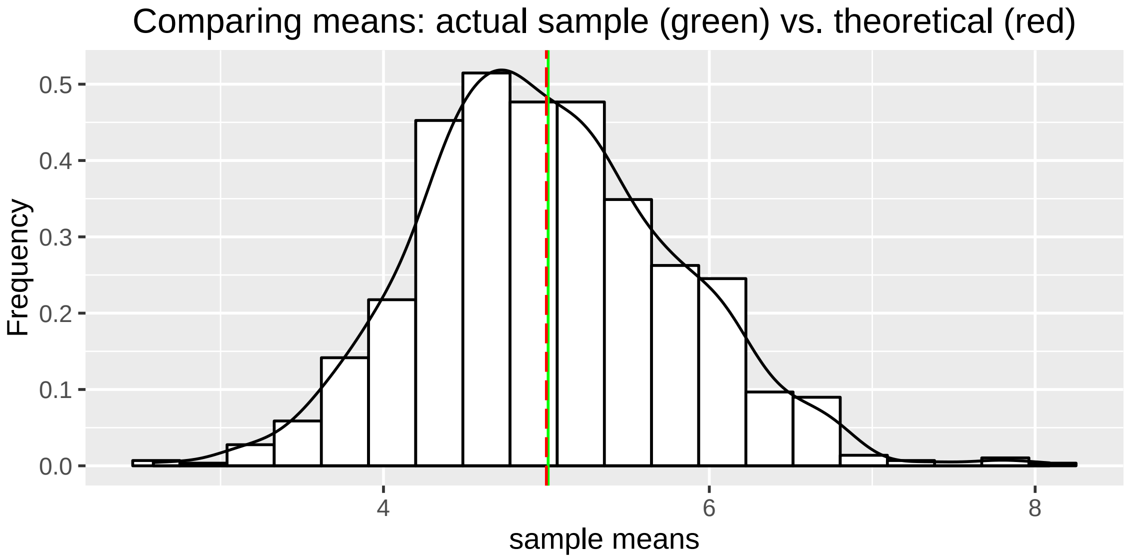
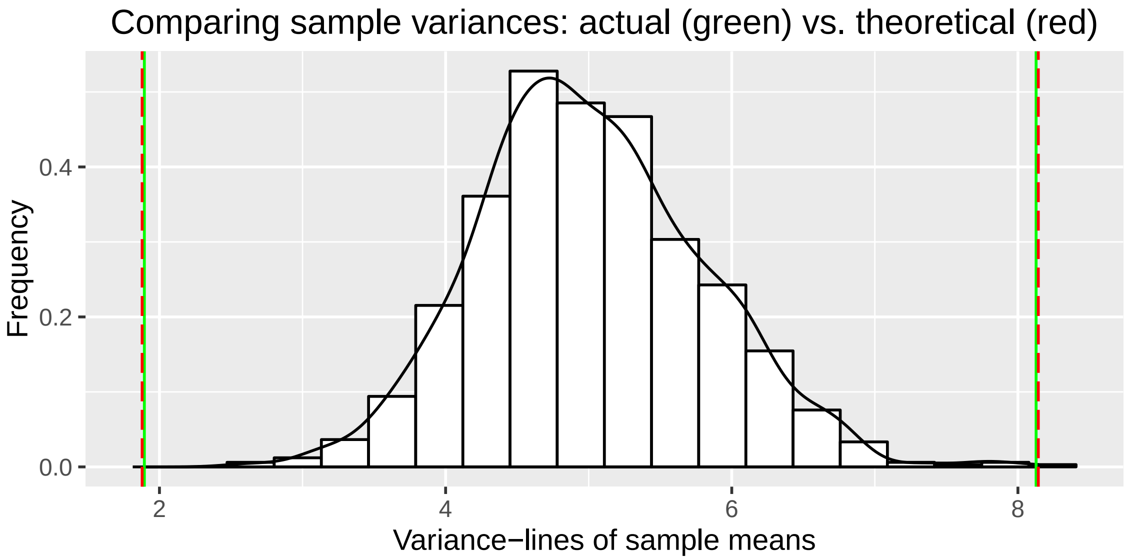


Figure 4. Comparing variances: Actual sample (green) vs. Theoretical (red)

```
m <- ggplot(dfsim,
        aes(x=dfsim$smeans))           ## setup graphic object
m + geom_histogram(                   ## plot histogram bars
        fill=I("White"),              ## fill with white
        col=I("Black"),               ## outline with black
        bins=20,                      ## appropriate bins
        aes(y=..density..)) +         ## set y to density
        xlab("Variance-lines of sample means") + ## label X-axis
        ylab("Frequency") +           ## label Y-axis
        ggtitle("Comparing sample variances: actual (green) vs. theoretical (red)") +
    geom_density (aes(y=..density..)) + ## plot density-line
    geom_vline (xintercept = aM+c(-1,1)*aV*aM, ## plot actual variance of sample means
          color = "Green") +
    geom_vline (xintercept = aM+c(-1,1)*(1/lambda^2/n)*aM,  ## plot theor. variance of sample means
          color = "Red",
          linetype = "longdash")
```


Comparing sample variances: actual (green) vs. theoretical (red)