Statistical Inference Course Project

Part 1: Simulation Exercise

Artem M.

November 2, 2017

Overview:

In this task we investigate the exponential distribution and compare it with Central Limit Theorem (CLT). The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$.

We will perform a 1000 simluations, where $\lambda = 0.2$ and number of observations in each simulations will be set to 40 ($n = 40$).

Note: For the code of the figures in this project see the Appendix.

Simulations:

We use R function **rexp** to generate the 1000 random samples of 40 exponentials. To make the results reproducible we will set random seed equal to 10. We will use matrix **exp_dis** with size $n \times 1000$ to store the results of the simulations:

```
set.seed(10)
exp_dis=NULL
for(i in 1:1000) exp_dis<-rbind(exp_dis,rexp(40,0.2))

dim(exp_dis)
```

```
## [1] 1000   40
```

Our simulation generated a matrix of 1000 rows(samples) by 40 columns(observations).

Sample Mean versus Theoretical Mean

We calculated the sample mean ($\bar{X}$) for our simulated data:

```
smeans<-as.data.frame(apply(exp_dis,1,mean))
names(smeans)<-"value"
xbar<-round(mean(smeans[,1]),3)
xbar
```

```
## [1] 5.045
```

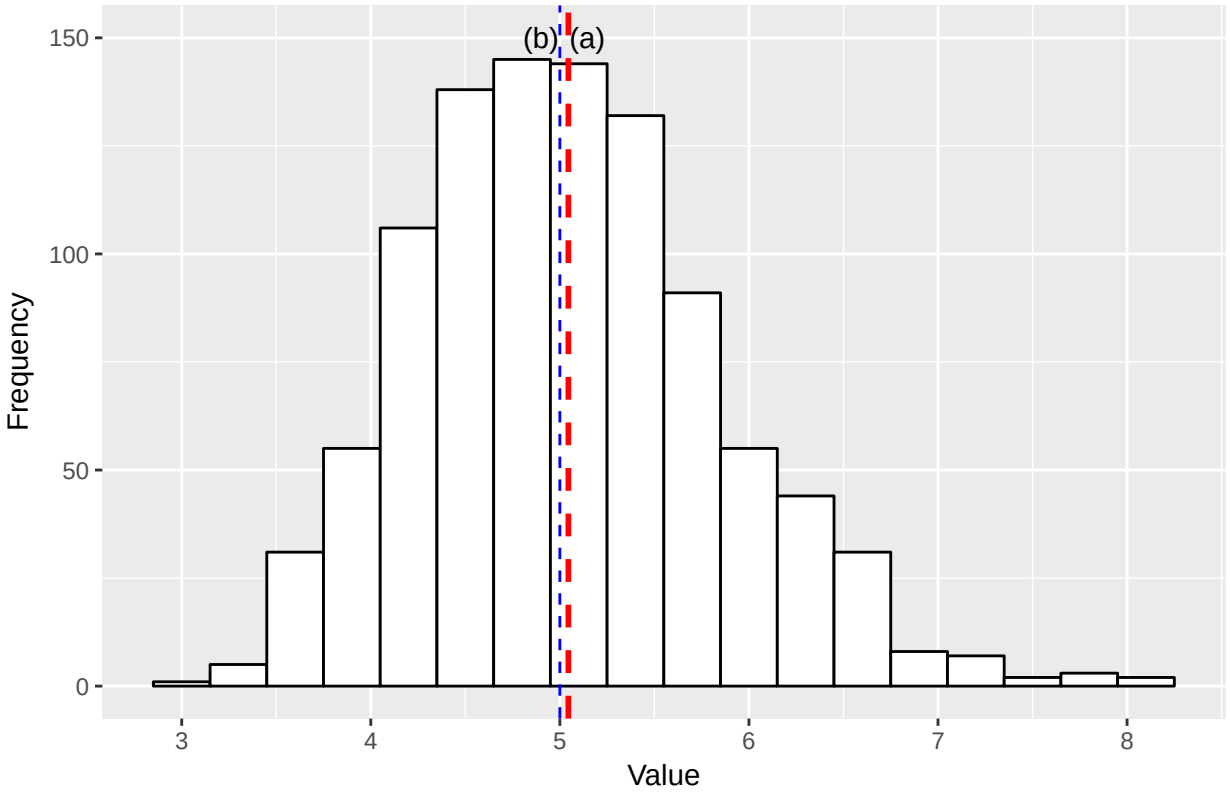
And then found the theoretical mean of the population ($E[X] = \mu = \frac{1}{0.2}$)

```
ex<-1/0.2
ex
```

```
## [1] 5
```

Then we constructed a sampled means histogram (see Fig.1) that shows that it is centered at 5.045 (a), which is close to the theoretical population mean 5 (b).

Fig.1: Sample histogram: (a) − $\overline{X}$=5.045; (b) − μ=5



Comparison of the results of the calculation demonstrates that the sample mean is close to the theoretical mean, as such clearly shows its applicability as a good population mean estimate.

Variability of the sample

By CLT $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, where σ^2 is the variation of the population being sampled from and the variance of the sample mean is (1)$Var(\bar{X}) = \frac{\sigma^2}{n}$

We can find the empirical sampling mean variance via R function var:

```
smvar<-var(smeans$value)
round(smvar,3)
```

```
## [1] 0.637
```

To find theoretical mean variance we will use variance equation (1) above and replace σ^2 with squared standard deviation for the exponential distribution $\left(\frac{1}{\lambda}\right)^2$. So the variance equation will take the form $Var(\bar{X}) = \frac{\left(\frac{1}{\lambda}\right)^2}{n}$. Variance value is calculated as follows:

```
tmvar<-(1/.2)^2/40
tmvar
```

```
## [1] 0.625
```

We will use standard deviation for graphically demonstrate mean variability ($\sqrt{Var(\bar{X})}$) where **sdemp** - is empirical standard deviation and **sdter** - theoretical SD.

```
sdemp<-round(sqrt(smvar),3)
sdemp
```

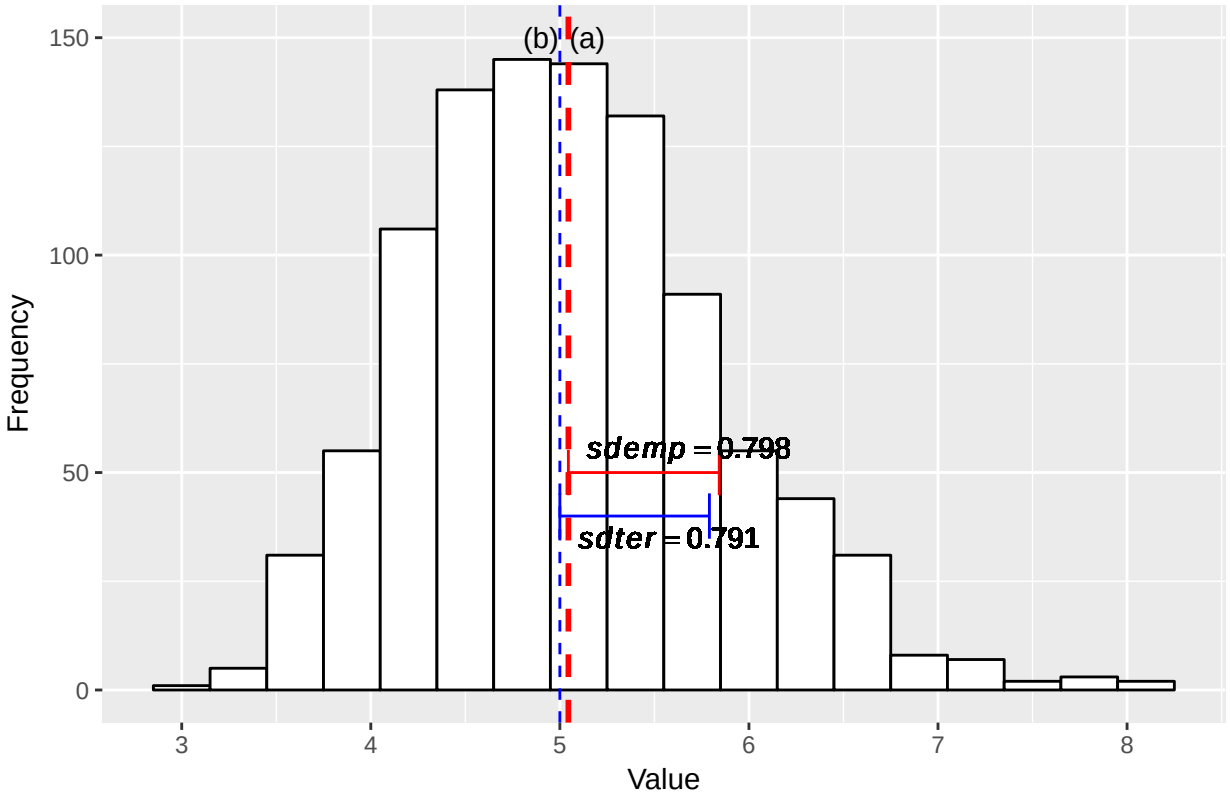
```
## [1] 0.798
```

```
sdter<-round(sqrt(tmvar),3)
sdter
```

```
## [1] 0.791
```

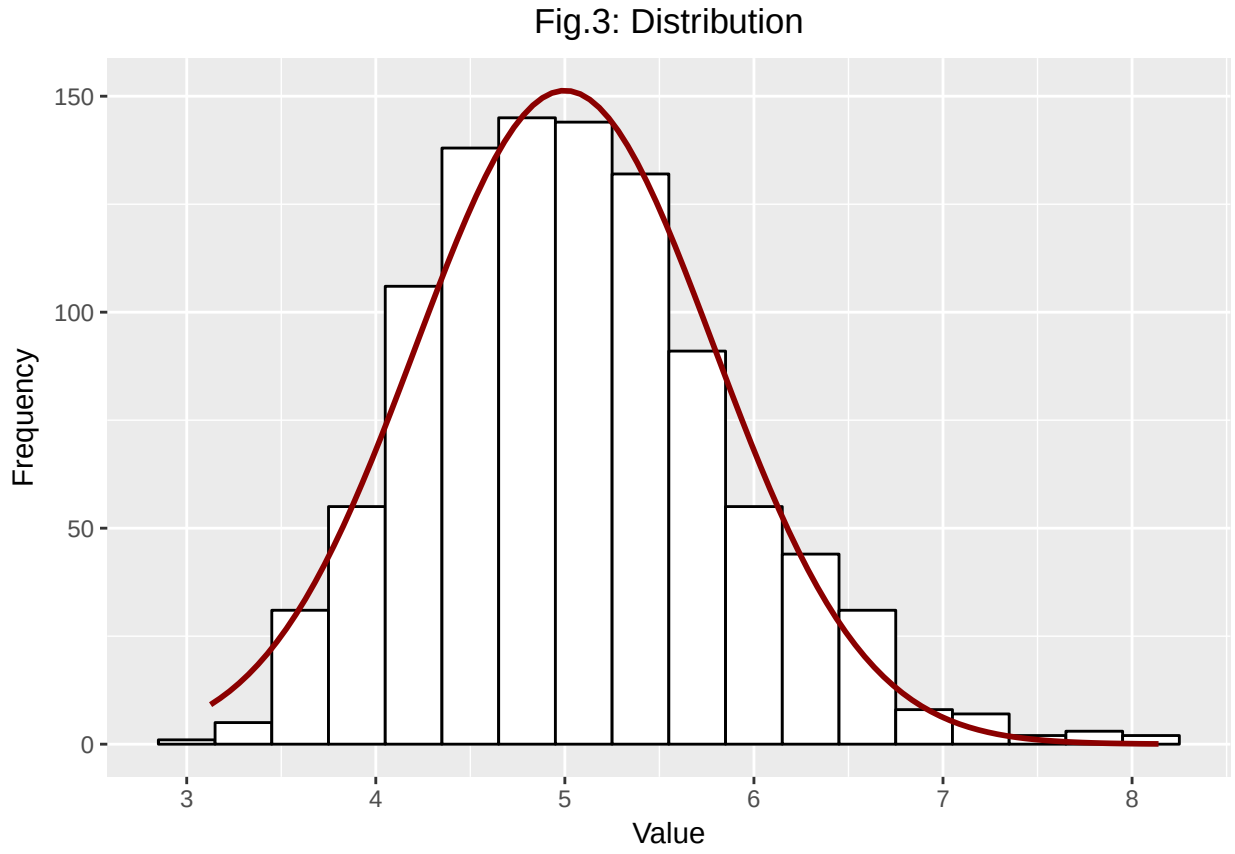
We see that on Fig.2 below, values of sdemp = 0.798 and sdter = 0.791 are very close, and consequently relevant variances.

Fig.2: Empirical SD: sdemp =0.798 vs. theoretical SD sdter=0.791



Distribution

To demonstrate that the samlping mean distribution is approximately normal we superimposed a normal curve over the Sample histogram, see Fig.3. To keep the frequency count of the y-axis we scaled the normal curve by the factor of the simulations(1000) and bin width(0.3). As we see on Fig.3 the original Sample histogram closely follows the normal curve, as such we can conclude that the sampling means distirbution is approximately normal.



Appendix

Fig.1 code

```
tit1<-substitute(
        paste("Fig.1: Sample histogram: (a) - ",bar(X),"=",m,"; ", "(b) - ",mu,"=",n),
                                list(m=xbar,n=5))
gmeans<-ggplot(smeans, aes(x=value)) +
        geom_histogram(binwidth=.3, colour="black", fill="white") +
        labs(title=tit1 )+

        labs(x="Value", y="Frequency") +
        theme(plot.title = element_text(hjust = 0.5)) +
        geom_vline(aes(xintercept=xbar),
               color="red", linetype="dashed", size=1)+
        annotate('text', x = xbar+.1, y = 150, label ="(a)")+
        geom_vline(aes(xintercept=5), color="blue", linetype="dashed", size=.5)+
        annotate('text', x = 5-.1, y = 150, label ="(b)")
gmeans
```

Fig.2 code

```
library(grid)
sd1 <- sprintf("italic(sdemp) == %.3f",sdemp )
sd2 <- sprintf("italic(sdter) == %.3f",sdter )
```

```
tit2<-substitute(paste("Fig.2: Empirical SD: sdemp =",m," vs. theoretical SD sdter=",n),
                                 list(m=sdemp,n=sdter))
vars<-ggplot(smeans, aes(x=value)) +
        geom_histogram(binwidth=.3, colour="black", fill="white") +
        labs(title=tit2 )+

        labs(x="Value", y="Frequency") +
        theme(plot.title = element_text(hjust = 0.5)) +
        geom_vline(aes(xintercept=mean(value)), color="red", linetype="dashed", size=1)+
        annotate("segment", x=xbar, xend=xbar+sdemp, y=50, yend=50, colour="red",
                 arrow=arrow(ends="both", angle=90, length=unit(.3,"cm")))+
        geom_text(x=xbar+.1, y=55, aes(label=sd1), parse=TRUE, hjust=0)+
        annotate('text', x = xbar+.1, y = 150, label ="(a)")+
        geom_vline(aes(xintercept=5), color="blue", linetype="dashed", size=.5)+
        annotate("segment", x=5, xend=5+sdter, y=40, yend=40, colour="blue",
                 arrow=arrow(ends="both", angle=90, length=unit(.3,"cm")))+
        geom_text(x=5+.1, y=35, aes(label=sd2), parse=TRUE, hjust=0)+
        annotate('text', x = 5-.1, y = 150, label ="(b)")
vars
```

Fig.3 code

```
dist<-ggplot(smeans, aes(x=value)) +
        geom_histogram(binwidth=.3, colour="black", fill="white") +
        labs(title="Fig.3: Distribution" )+
        labs(x="Value", y="Frequency") +
        theme(plot.title = element_text(hjust = 0.5))+
        stat_function(fun = function(x) dnorm(x, mean = 5, sd = .791)*1000*.3,
    color = "darkred", size = 1)

dist
```