Statistical Inference Course Project

Veasna Kheng

23/07/2020

Part 1: Means of 40 Exponential Distribution with $\lambda = 0.2$

This part reports the investigation of the distribution of averages of 40 exponentials and compare it with the Central Limit Theorem using R with a thousand of simulation. The probability density function of the exponential distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The following simulation is based on lambda = 0.2. Thus, the population mean and standard deviation are

$$\mu = \sigma = 1/0.2 = 5$$

.

Sample Mean Versus Theoretical Mean

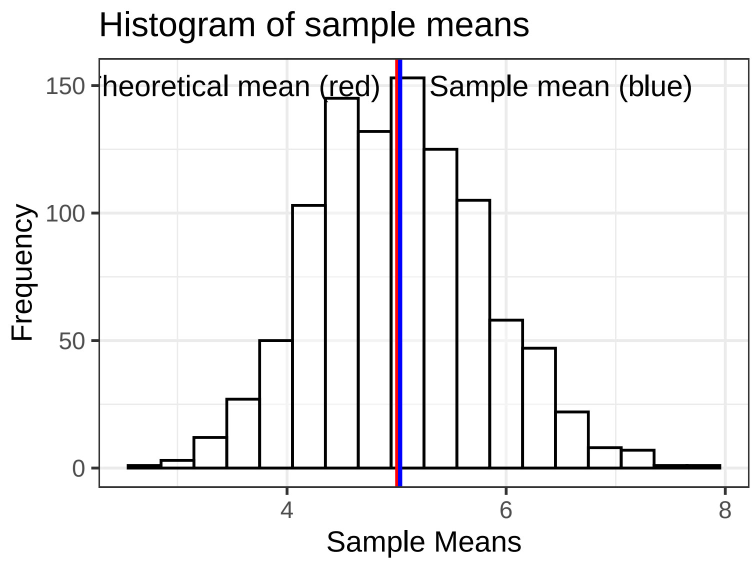
```r
set.seed(111)
n <- 1000 # number of simulations
lambda <- 0.2
s_means <- NULL
for (i in 1:n) s_means = c(s_means, mean(rexp(40,lambda))) # 1000 simulation of sample mean

# sample mean (expectation of sample mean)
mubar  <- mean(s_means)
mubar
```

```
## [1] 5.02562
```

```r
# theoretical mean
mu <- 1/lambda
mu
```

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

```r
library(ggplot2)
ggplot(data.frame(1:n, s_means), aes(x = s_means)) +
  geom_histogram(alpha =.40,binwidth=.3,col="black", fill = "white") +
  geom_vline(xintercept = mubar, col = "blue", size = 1) +
  geom_vline(xintercept = mu, col = "red") +
  labs(x ="Sample Means",y = "Frequency",title = "Histogram of sample means") +
  annotate("text",x = c(3.5,6.5) ,y= 150,
           label=c("Theoretical mean (red)","Sample mean (blue)")) +
  theme_bw()
```



Sample Variance Versus Theoretical Variance

```
# sample variance
var(s_means)
```

```
## [1] 0.6069798
```

```
# theoretical variance
1/lambda^2 / 40
```

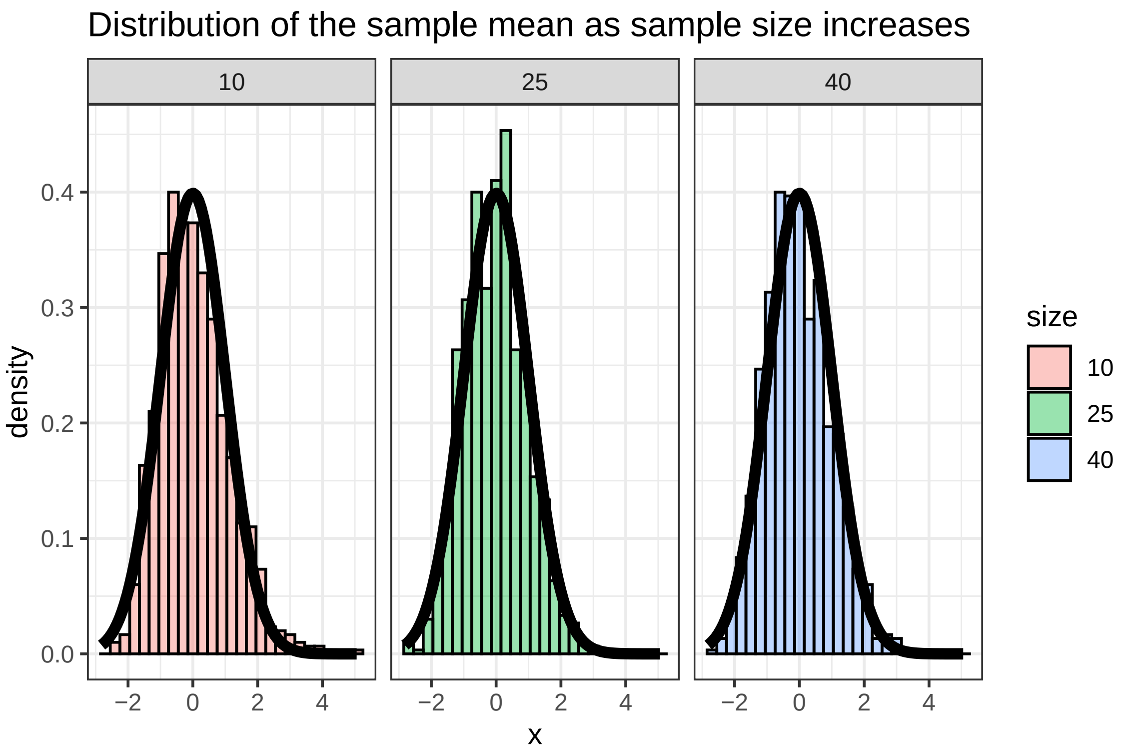
```
## [1] 0.625
```

The Central Limit Theorem (CLT)

The CLT states that the distribution of average of i.i.d variables becomes a normal distribution as the sample size increases. That is, the mean ($\bar{\mu}$) of sample size n is approximately $N(\mu, \sigma^2/n)$. The following simulation and figure prove this point.

Distribution of the sample mean

```
nosim <- 1000 # number of simulation
cfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5 # Z value
# create data with sample size 10, 25, and 40.
dat <- data.frame(
  x = c(apply(replicate(nosim,rexp(10, lambda)), 2, cfunc, 10),
        apply(replicate(nosim,rexp(25, lambda)), 2, cfunc, 25),
        apply(replicate(nosim,rexp(40, lambda)), 2, cfunc, 40)
        ),
  size = factor(rep(c(10, 25, 40), rep(nosim, 3))))
ggplot(dat, aes(x = x, fill = size)) +
  geom_histogram(alpha =.40,binwidth=.3,col="black",aes(y = ..density..)) +
  stat_function(fun = dnorm, size = 2) +
  facet_grid(. ~ size) +
  labs(title = "Distribution of the sample mean as sample size increases") +
  theme_bw()
```



We can see that as sample size increases, the distribution of sample mean becomes more like a bell shape, normal distribution.
