Statistical Inference Course Project

Marko Intihar

Part 1: Simulation

We are investigating exponential distribution and we are comparing it with the Central Limit Theorem (CLT). Random variable X follows exponential distribution $exp(\lambda)$, where parameter $lambda$ is rate parameter. The probability density function (PDF) of exponential distribution is:

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

From the theory we know that mean or expected value of exponential distribution is $E[X] = \lambda$ and variance is $Var[X] = \frac{1}{\lambda^2}$.

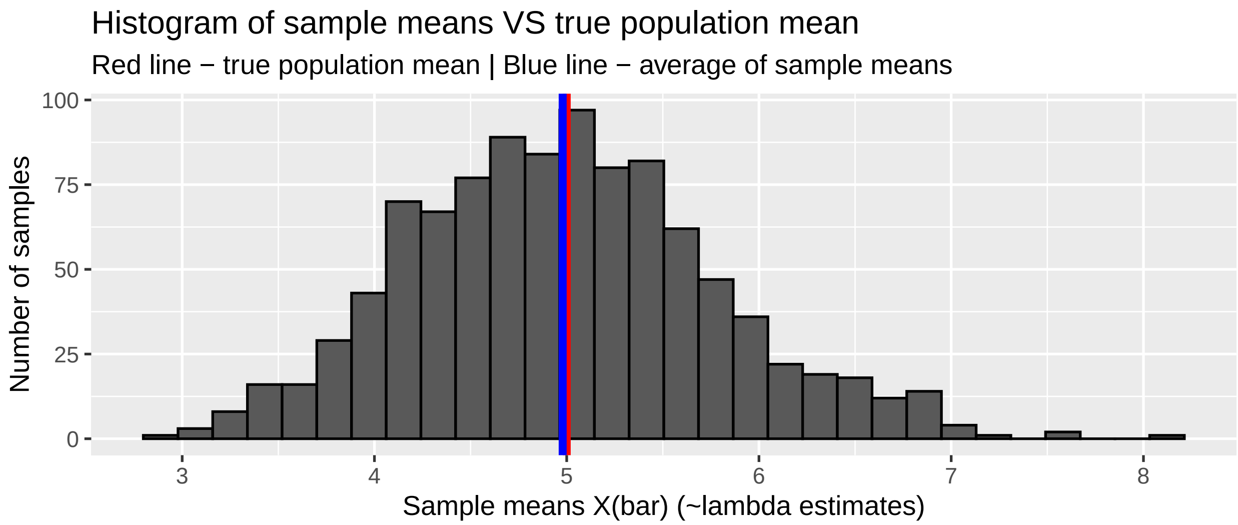
In our simulation we will simulate random samples from exponential distribution (where $\lambda = 0.2$). Each random sample will include $n = 40$ random numbers, and we will generate $B = 1000$ different random samples. Let's simulate our samples:

```
set.seed(11235)
lambda <- 0.2 # lambda of population distribution
n <- 40 # sample size
B <- 1000 # samples generated in simulation
samples <- matrix(rexp(n = n * B, rate = lambda),  # generate samples - save to matrix
                  nrow = B, ncol = n)
```

Now we would like to show the sample mean and compare it to the theoretical mean of the distribution. So let's draw a histogram of sample means VS theoretical mean of the distribution:

```
sample.means <- apply(samples, 1, mean) %>% # calculate sample means
  data.frame(X.bar = .)

# draw histogram sample means VS theoretical distribution mean
sample.means %>%
  ggplot(aes(x = X.bar)) +
  geom_histogram(color = "black", bins = 30) +
  geom_vline(xintercept = 1/lambda, color = "red", size = 1.5) +
  geom_vline(xintercept = mean(sample.means$X.bar), color = "blue", size = 1.5) +
  scale_x_continuous(breaks = seq(0,10)) +
  labs(title = "Histogram of sample means VS true population mean",
       subtitle = "Red line - true population mean | Blue line - average of sample means",
       x = "Sample means X(bar) (~lambda estimates)",
       y = "Number of samples")
```

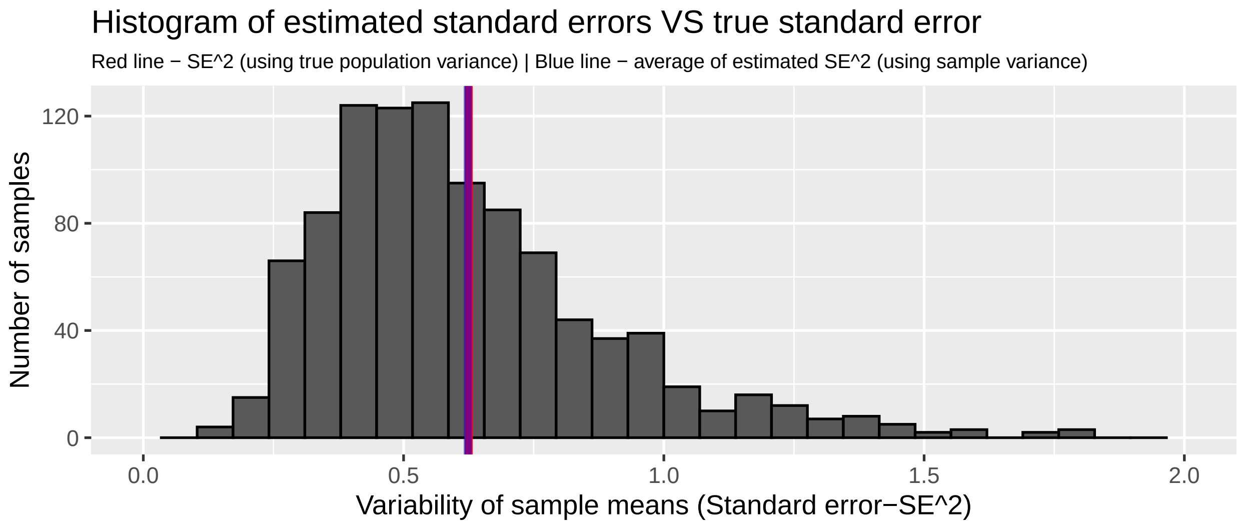


From the figure above we can see that estimated mean of all sample means (blue line), which has value of 4.98 is very close to the true mean of the distribution, which has value 5 (red line). This indicates that CLT applies to given distribution, and mean value for distribution of sample means is equal to mean of true population distribution (from which samples were generated).

Now check what is the variability of sample means, based on the theory (CLT) the variability (variance of distribution) of sample means should be close to $SE^2 = \sigma^2/n$, where SE is standard error, σ^2 is variance of population distribution (exponential distribution: $1/\lambda^2$). Using samples we can estimate standard error $SE^2 = s^2/n$, where s^2 is estimated variance from sample.

```
sample.var <- apply(samples, 1, var) %>% # calculate sample variance
  data.frame(s2.bar = .) %>%
  mutate(SE2 = s2.bar / n)

# draw histogram SE2 (sample variance) VS SE (population variance)
sample.var %>%
  ggplot(aes(x = SE2)) +
  geom_histogram(color = "black", bins = 30) +
  geom_vline(xintercept = (1/lambda^2)/n, color = "red", size = 1.5) +
  geom_vline(xintercept = mean(sample.var$SE2), color = "blue", size = 1.5, alpha = 1/2) +
  scale_x_continuous(breaks = seq(0,10, 0.5), limits = c(0,2)) +
  labs(title = "Histogram of estimated standard errors VS true standard error",
       subtitle = "Red line - SE^2 (using true population variance) | Blue line - average of estimated SE^2 (using sample variance)",
       x = "Variability of sample means (Standard error-SE^2)",
       y = "Number of samples") +
  theme(plot.subtitle = element_text(size = 8))
```

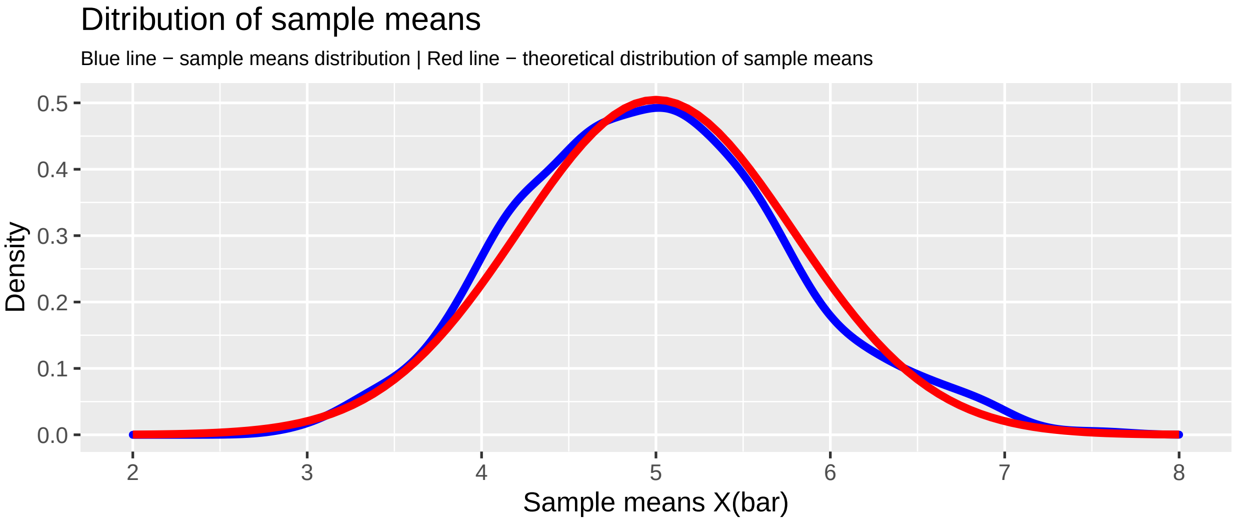


From the figure above we can see that mean value of estimated SE^2 (using samples) - blue line is 0.62, and is very close to true SE^2 (using true population variance) with value 0.62 - red line. This goes with the given

CLT theory.

Now let's show that distribution of sample means follow normal distribution with given parameters $\bar{X} \sim N(\mu, \sigma^2/n)$, where μ is population distribution mean, σ^2 is population distribution variance, and n is sample size. The figure below shows that our distribution is approximately normal and is very similar to theoretical normal curve.

```r
# draw density plot of distribution of smaple means VS theoretical normal curve
sample.means %>%
  ggplot(aes(x = X.bar)) +
  geom_density(color = "blue", size = 1.5, ) +
  stat_function(fun = dnorm, args = list(mean = 1/lambda, sd = sqrt((1/lambda^2)/n)), color = "red", size = 1.5) +
  scale_x_continuous(breaks = seq(0,10, 1), limits = c(2,8)) +
  labs(title = "Ditribution of sample means",
       subtitle = "Blue line - sample means distribution | Red line - theoretical distribution of sample means",
       x = "Sample means X(bar)",
       y = "Density") +
  theme(plot.subtitle = element_text(size = 8))
```


Ditribution of sample means
Blue line − sample means distribution | Red line − theoretical distribution of sample means
Density
0.5
0.4
0.3
0.2
0.1
0.0
2
3
4
5
6
7
8
Sample means X(bar)