Course Project 1: Simulation Exercise

Xiaozhu Zhang

Jan. 13, 2021

In this part I will investigate the exponential distribution and compare it with CLT. If a random variable $X \sim \text{Exp}(\lambda)$, then

$$f(x) = \lambda e^{-\lambda}, \quad \mathbb{E}(X) = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

Now I will set $\lambda = 0.2$, and investigate the distribtuion of averages of 40 exponentials (namely $\bar{X} = (\sum_{i=1}^{40} X_i)/40$, for i.i.d. X_i) with 1000 simulations.

1. Simulation process

First, set the parameter λ, the sample size n, and the number of simulations N.

```
lambda = 0.2
n = 40
N = 1000
```

Then sample $40 * 1000$ values from the Exp(0.2) density, store them in a matrix, and use the `apply` function to obtain the `Sample` of length 1000.

```
set.seed(1234)
Sample <- matrix(rexp(n = n * N, lambda), N, n)
Sample <- apply(Sample, 1, mean)
```

2. Mean comparison

We can draw the histogram of `Sample`, where the blue line, $\mathbb{E}(\bar{X}) = 1/\lambda$, is the theoretical mean, while the red line is the sample mean of the distribution. It is clear from figure 1 that they are very close to each other. The exact numbers shown below give the same result.

```
# Theoretical mean
1/lambda
```

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

```
# Sample mean
mean(Sample)
```

```
## [1] 4.974239
```

3. Variance Comparison

The theoretical variance is $\text{Var}(\bar{X}) = \text{Var}(X)/n = 1/(n\lambda^2)$.

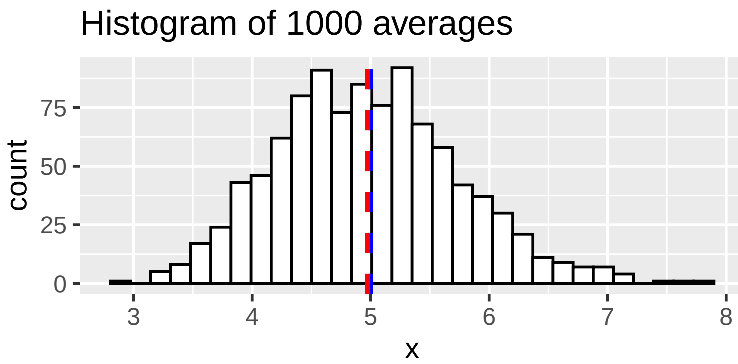

Figure 1: Histogram of 1000 averages over 40 exponentials

```r
# Theoretical variance
1/(n * lambda^2)
```

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

```r
# Sample variance
var(Sample)
```

```
## [1] 0.5949702
```

Still, they are close. But in a way the difference between theoretical and sample variance is beyond nuance, implying that $n = 40$ may not be large enough since CLT requires a large sample size. For example, if we increase n to 100 or 150, then the gap between theoretical and sample variance will decrease, and from the figure below we conclude that the distribution will be more centered.

```
##       theory_variance sample_variance
## n=40        0.6250000       0.5949702
## n=100       0.2500000       0.2357349
## n=150       0.1666667       0.1665783
```

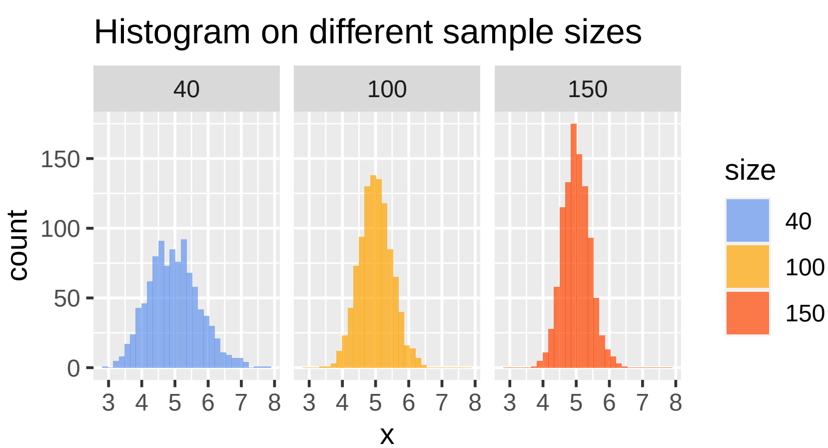

Figure 2: Histogram on different sample sizes

4. Distribution

The sampling distribution of $\bar{X}_n$ is shown in figure 1, while the population distribution is shown in figure 3. The distribution from population to sampling is becoming more symmetric and more like normal.

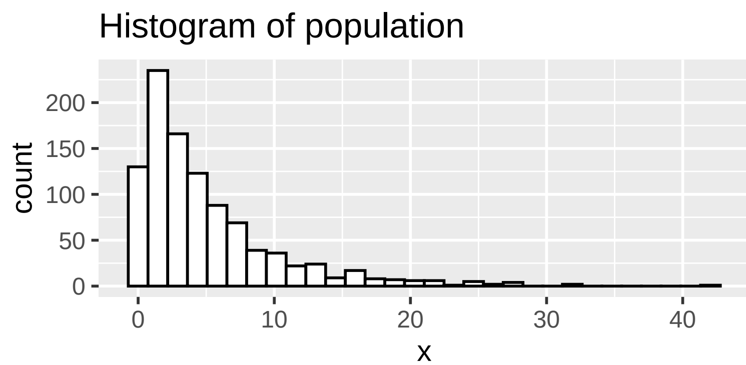

Figure 3: Histogram of population

In order to investigate more about the normality, we can draw the qq-plot as figure 4. We conclude that the sampling distribution is *approximately normal* although with few outlier points indicating a little skewness.

```
library("ggpubr")
ggqqplot(Sample)
```

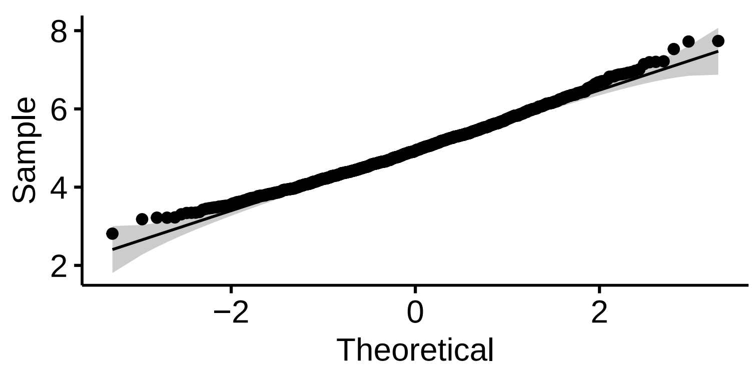

Figure 4: qq-plot of sampling distribution

In a nutshell, as n goes to infinity, we have

$$\bar{X} \to N(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$$

in distribution.

Appendix

```r
# Codes for figure 1
library(ggplot2)
ggplot(data.frame(Sample), aes(x = Sample)) +
  geom_histogram(colour = "black", fill = "white", bins = 30) +
  geom_vline(aes(xintercept = 1/lambda),
             color = "blue", linetype = "dashed", size = 0.9) +
  geom_vline(aes(xintercept = mean(Sample)),
             color = "red", linetype = "dashed", size = 0.9) +
  labs(title = "Histogram of 1000 averages", x = "x")
```

```r
# Codes for figure 2 and the table
SD <- function(lambda, n, N){
  set.seed(1234)
  Sample <- matrix(rexp(n = n * N, lambda), N, n)
  Sample <- apply(Sample, 1, mean)
  return(list(Sample = Sample))
}
Sp0 <- Sample
Sp1 <- SD(lambda, 100, N)$Sample
Sp2 <- SD(lambda, 150, N)$Sample
Sp <- c(Sp0, Sp1, Sp2)
label <- factor(rep(c("40", "100", "150"), each = 1000),
                levels = c("40" ,"100", "150"), ordered = TRUE)
# the table
theory_var <- 1/(c(40, 100, 150) * lambda^2)
sample_var <- c(var(Sp0), var(Sp1), var(Sp2))
dat <- data.frame(theory_var, sample_var)
rownames(dat) <- c("n=40", "n=100", "n=150")
colnames(dat) <- c("theory_variance", "sample_variance")
dat

# figure 2
ggplot(data.frame(Sp, size = label), aes(x = Sp)) +
  geom_histogram(aes(fill = size), bins = 30) +
  scale_fill_manual(values = alpha(c("#6495ED", "#FFA500", "#FF4500"), 0.7)) +
  facet_grid(.~ size) +
  labs(title = "Histogram on different sample sizes", x = "x", y = "count")
```

```r
# Codes for figure 3
dat <- data.frame(x = rexp(n = 1000, lambda))
ggplot(dat, aes(x = x)) +
  geom_histogram(colour = "black", fill = "white", bins = 30) +
  labs(title = "Histogram of population", x = "x", y = "count")
```