630 HW 4

David Teng

Wednesday, October 16th by 11:59 pm

```
library(dplyr)
library(ggplot2)
library(knitr)
```

1.

The sample mean is defined as: X bar = 1/n * the sum of Xi from 1 to n Since Xi N (miu, sigma^2) E (X bar) = 1/n * the sum of E [Xi] from 1 to n = 1/n * n * miu = miu

```
cat("As my proof above, E[ X bar] = miu ")
```

```
## As my proof above, E[ X bar] = miu
```

The variance of Xi is sigma^2 so the variance of X bar is Var (X bar) = 1/ n^2 * the sum of Var (Xi) from 1 to n = 1/ n^2 * n * sigma^2 = sigma^2 / n Taking the square root gives the standard deviation: SD [X bar] = sigma / sqrt(n)

```
cat("As my proof above, SD [X bar] = sigma / sqrt(n) \n  ")
```

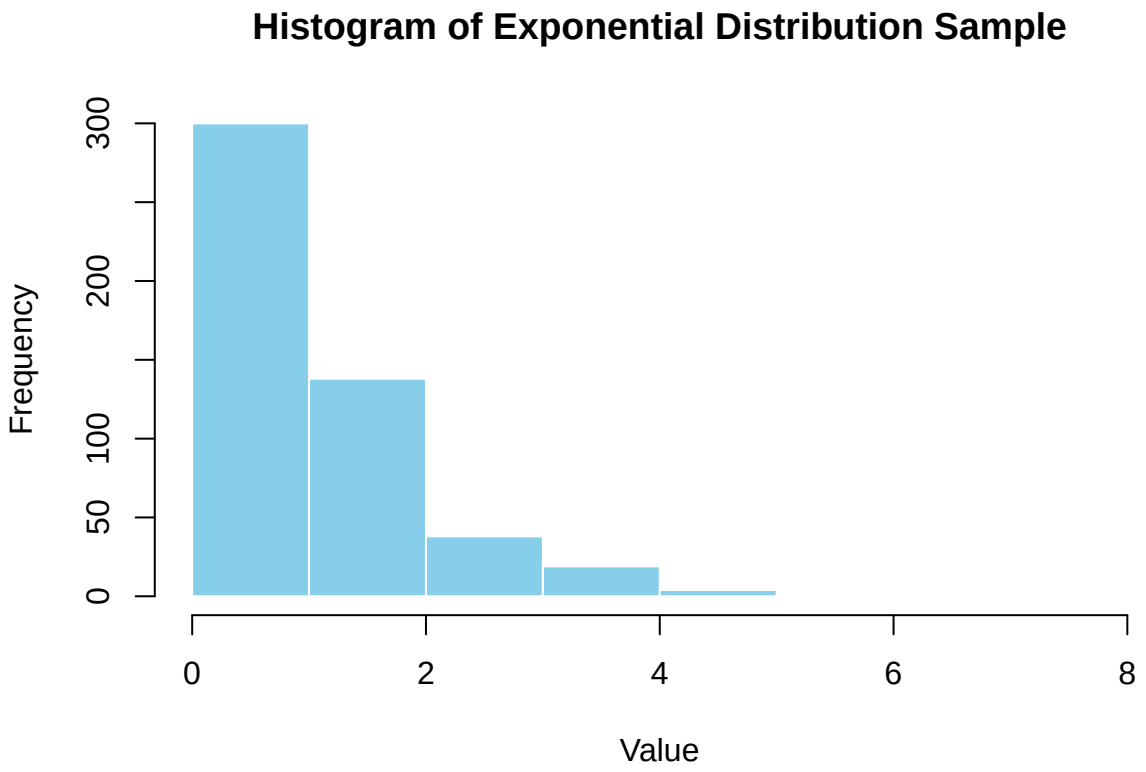
```
## As my proof above, SD [X bar] = sigma / sqrt(n)
##
```

```
cat("X bar ~ N (miu, sigma^2 / n ), and this proves the statement of the CLT. \n  ")
```

```
## X bar ~ N (miu, sigma^2 / n ), and this proves the statement of the CLT.
##
```

2.

```
set.seed(123)
samp <- rexp(500, rate = 1)
hist(samp,
     main = "Histogram of Exponential Distribution Sample",
     xlab = "Value",
     ylab = "Frequency",
     col = "skyblue",
     border = "white")
```



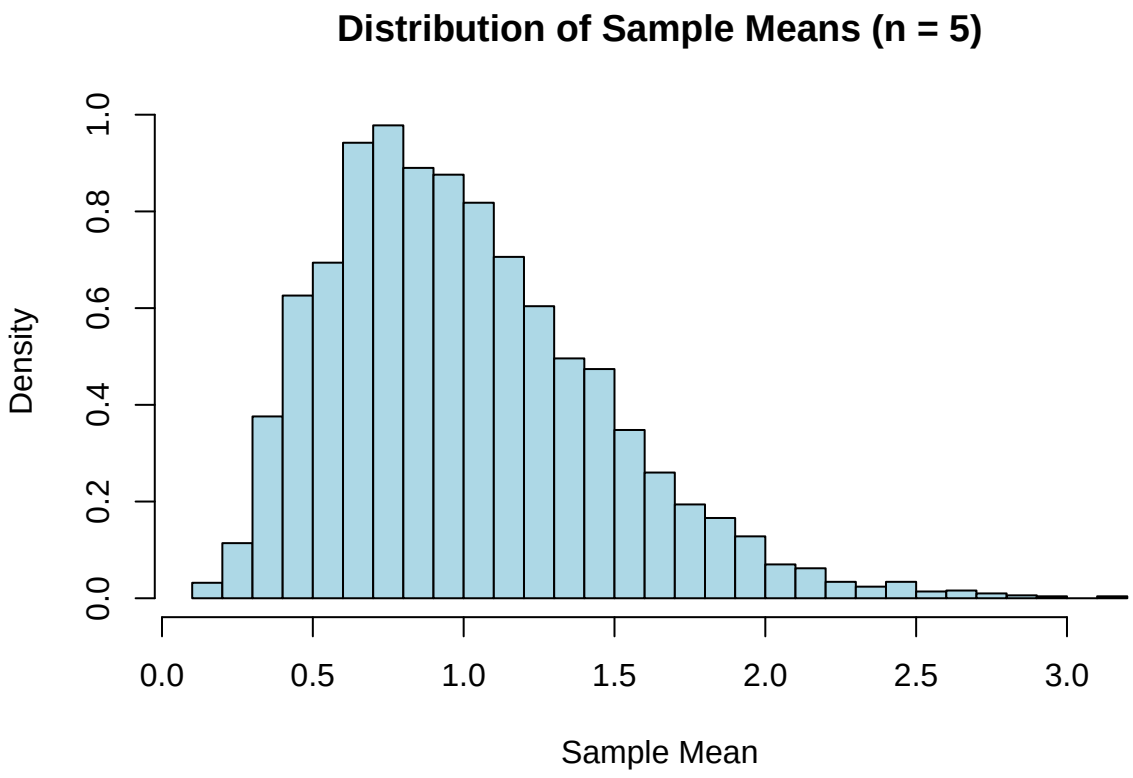
3. Comments of Histogram of Exponential Distribution Sample: The shape is right-skewed which is characteristic of the exponential distribution. The center is around 1, which is the mean of an exponential distribution with = 1. The spread is wide, and most values concentrated closer to the origin and the frequency decreasing while values increase.

4.

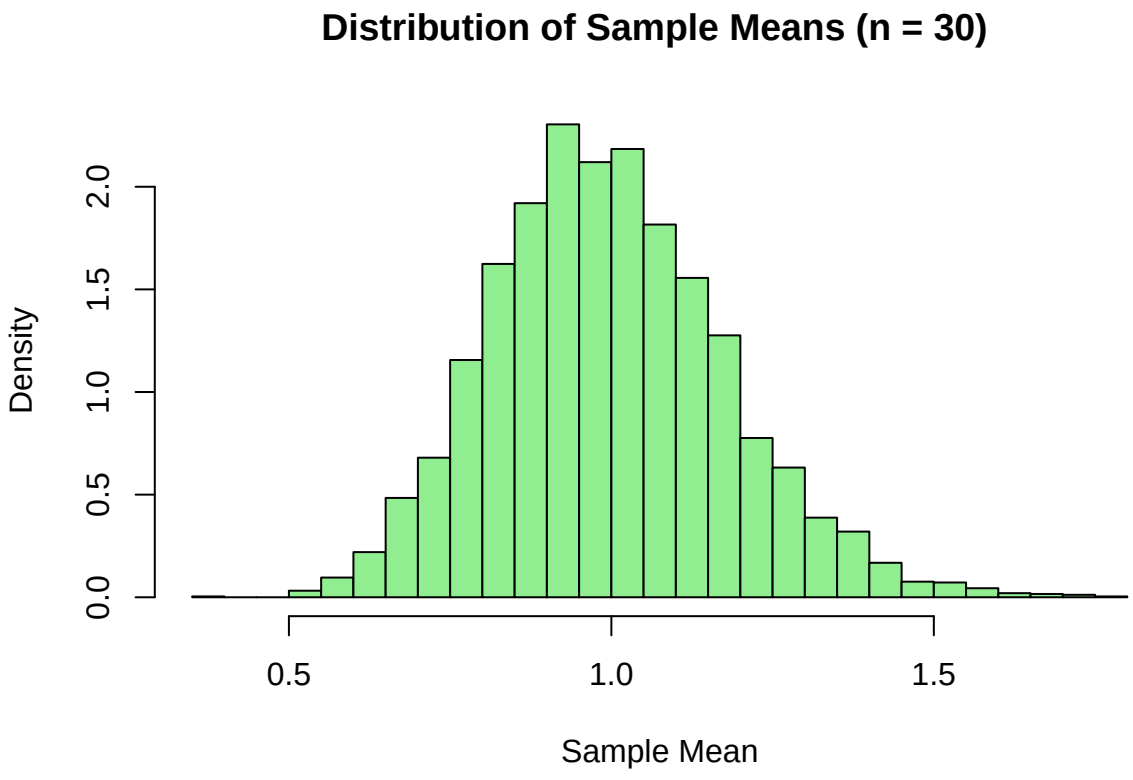
```
set.seed(123)
lambda <- 1
samp <- 5000

means <- function(n) replicate(samp, mean(rexp(n, rate = lambda)))
mean5 <- means(5)
mean30 <- means(30)
mean100 <- means(100)
par(mfrow = c(1, 3))
```

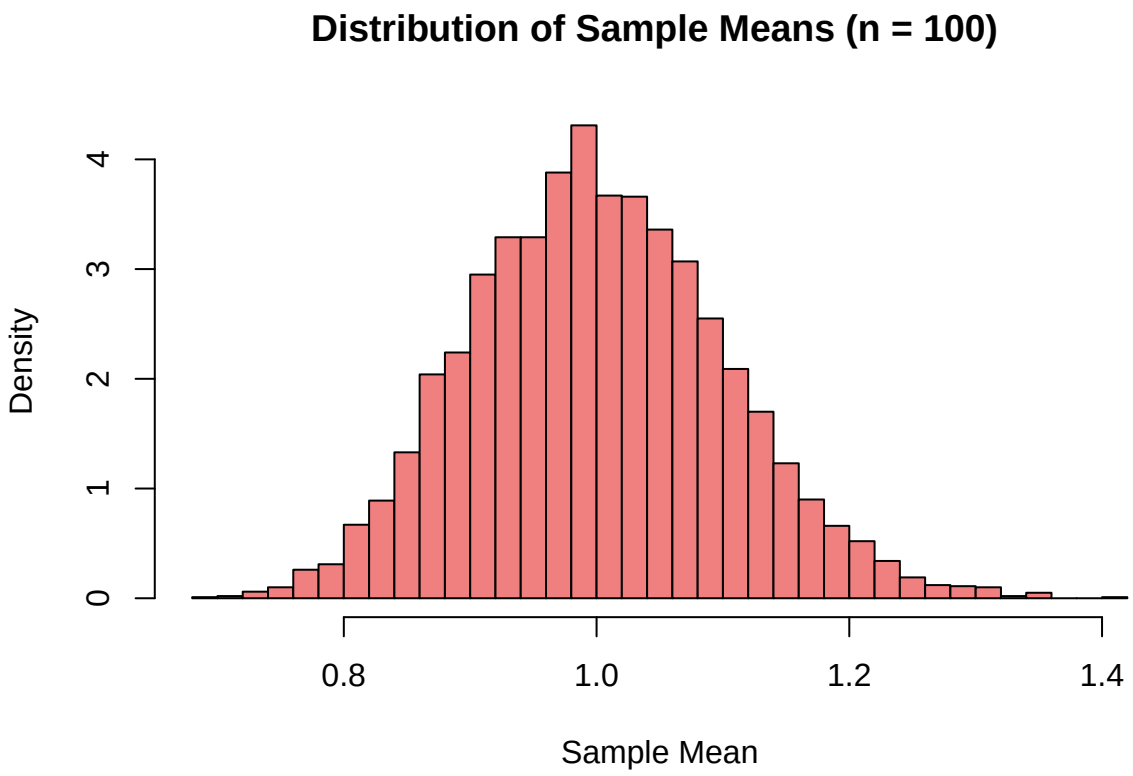
```
hist(mean5,xlab = "Sample Mean", breaks = 30, main = "Distribution of Sample Means (n = 5)", col = "lig
```




```
hist(mean30,xlab = "Sample Mean", breaks = 30, main = "Distribution of Sample Means (n = 30) ", col = "
```




```
hist(mean100,xlab = "Sample Mean", breaks = 30, main = "Distribution of Sample Means (n = 100) ", col =
```



5. When the sample size increases, the shape of the sampling distribution becomes more like normal distribution. This happens according to CLT, which states that the distribution of sample means approaches a normal distribution as the sample size becomes large regardless of the original distribution.

6.

```r
set.seed(123)
lambda <- 1
samp_sizes <- c(5, 30, 100)
samp <- 5000

stats <- function(n) {
  means <- replicate(samp, mean(rexp(n, rate = lambda)))
  c(mean = mean(means), sd = sd(means), se = (1 / lambda) / sqrt(n))
}
stats <- sapply(samp_sizes, stats)

table <- data.frame(
  "Sample Size" = samp_sizes,
  "Sample Means" = stats["mean", ],
  "Sample SD" = stats["sd", ],
  "Theoretical Mean" = 1 / lambda,
  "Theoretical SE" = stats["se", ]
)

kable(table, caption = "Sample and Theoretical Statistics for Different Sizes")
```

Table 1: Sample and Theoretical Statistics for Different Sizes

Sample.Size	Sample.Means	Sample.SD	Theoretical.Mean	Theoretical.SE
5	1.0051203	0.4508480	1	0.4472136
30	0.9956953	0.1819653	1	0.1825742
100	1.0009798	0.1003727	1	0.1000000

7. For a sample size of 5 : The sample mean is 1.005 while the theoretical mean is 1 . The sample standard deviation is 0.451 compared to the theoretical standard error of 0.447 .

For a sample size of 30 : The sample mean is 0.996 while the theoretical mean is 1 . The sample standard deviation is 0.182 compared to the theoretical standard error of 0.183 .

For a sample size of 100 : The sample mean is 1.001 while the theoretical mean is 1 . The sample standard deviation is 0.1 compared to the theoretical standard error of 0.1 .

8.

```
set.seed(123)
samp <- rexp(10, rate = 1)
n <- length(samp)
mean <- mean(samp)
sd <- sd(samp)

ci_norm <- mean + c(-1, 1) * qnorm(0.975) * (sd / sqrt(n))
ci_t <- mean + c(-1, 1) * qt(0.975, df = n - 1) * (sd / sqrt(n))
cat("The normal distribution CI (incorrect):", ci_norm, "\n")
```

```
## The normal distribution CI (incorrect): 0.1136765 1.159983
```

```
cat("The t-distribution CI (correct):", ci_t, "\n")
```

```
## The t-distribution CI (correct): 0.03301519 1.240644
```

9. The t-distribution is used for small samples with an unknown standard deviation because it has more variability, giving us wider and more reliable confidence intervals than the normal distribution.

10.

ASA has pointed out that p-values are often misunderstood and misused. Researchers often treat p = 0.05 as a strict threshold (bright line) for rejecting the null hypothesis, despite p-values close to 0.05 offering only weak evidence. This reliance on p-values ignores important factors like context and effect size, contributing to the reproducibility crisis—issues with replicating scientific findings. Despite these issues, the overemphasis on p-values continues, basically because it's how researchers have been taught, creating a "cycle" that's hard to break.

11.

Simulating the datasets was pretty straightforward, but connecting the results to the Central Limit Theorem was a bit tough. I need more practice with understanding how sampling distributions actually work in theory.

12.

E - Excellent