HW 4

Brandon Keck

2024-10-19

```
library(ggplot2)
```

1

$$E(\bar{X}) = \mu$$

$$Let\ \bar{X} = X_1 + X_2 + ... + X_n$$

$$E(\bar{X}) = \frac{1}{n}E(X_1 + X_2 + ... + X_n)$$

$$= \frac{1}{n}E(X_1) + E(X_2) + ... + E(X_n)$$

$$\frac{1}{n} * n\mu$$
$$= \mu$$

$$Var(\bar{X}) = \frac{1}{n^2}Var(X_1 + X_2 + ... + X_n)$$

$$= \frac{1}{n^2}Var(X_1) + Var(X_2) + ... + Var(X_n)$$

$$= \frac{1}{n^2} * n\sigma^2$$

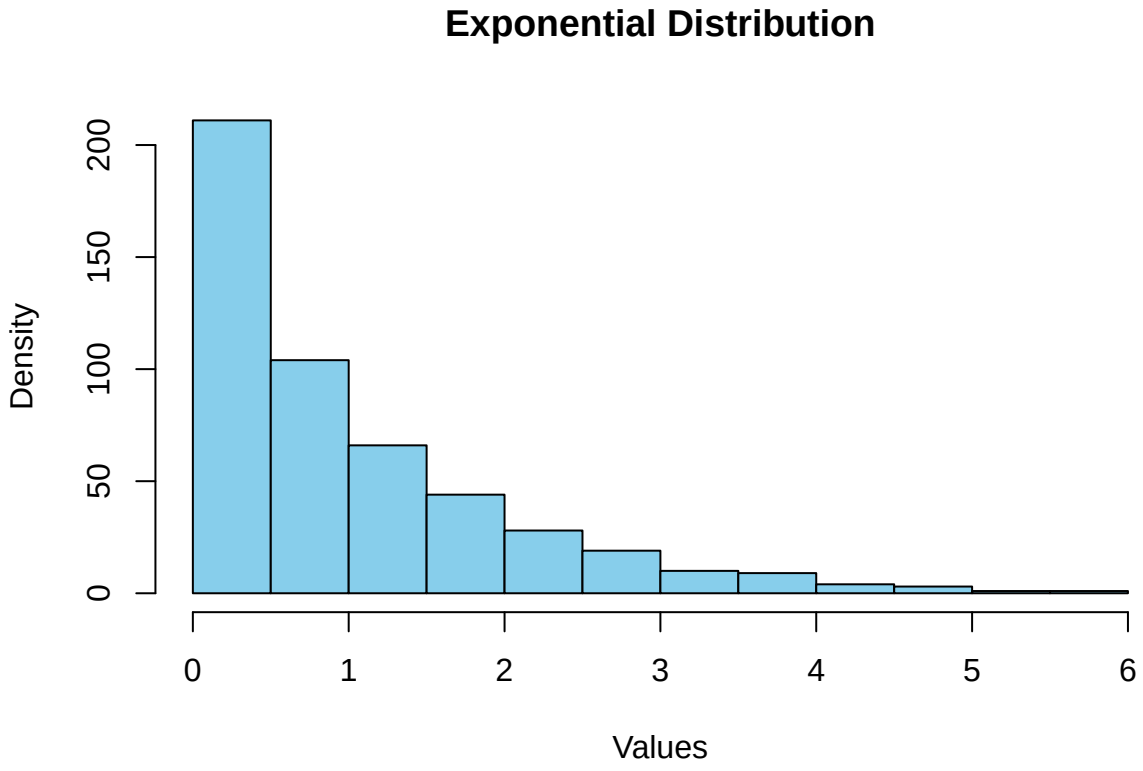
$$= \frac{\sigma^2}{n}$$

$$= \frac{\sigma}{\sqrt{n}}$$

$$E(\bar{X}) = \mu\ and\ variance\ SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

2.

```r
set.seed(11101987) # scorpio
values <- rexp(n = 500, rate = 1)

hist(
  values,
  main = "Exponential Distribution",
  xlab = "Values",
  ylab = "Density",
  col = "skyblue"
)
```




```r
# I found a great example of how to do this
# Here https://makemeanalyst.com/statistics-with-r/exponential-distribution-in-r-programming/#
```

3

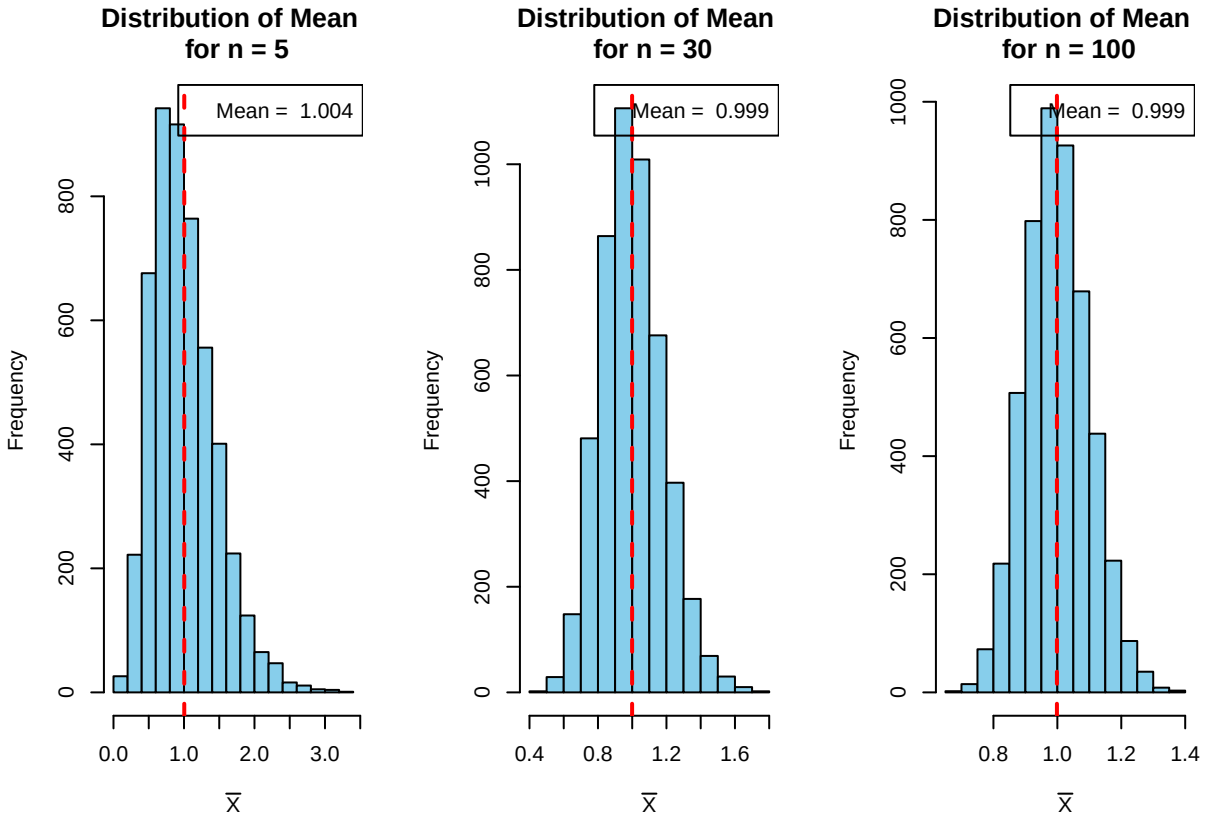
```r
mean(values)
```

```
## [1] 1.007818
```

The shape of the histogram is right skewed with a unimodal shape. Most of the values are near zero, with the center of the distribution around 1.007818. The spread of the distribution spans from 0 to about 6 with most of the values ranging from 0 to 3.

4

```r
# Generate random values from an exponential distribution with rate = 1
set.seed(11101987) # scorpio
B <- 5000

x_bar5 <- rep(NA, B)

# for loop for samples size n = 5
for(i in 1:B){
  samplen5 <- rexp(n = 5, rate = 1)
  x_bar5[i] <- mean(samplen5) # mean n = 5
}

x_bar30 <- rep(NA, B)

# for loop for samples size n = 30
for(i in 1:B){
  samplen30 <- rexp(n = 30, rate = 1)
  x_bar30[i] <- mean(samplen30) # mean n = 30
}

x_bar100 <- rep(NA, B)

# for loop for samples size n = 100
for(i in 1:B){
  samplen100 <- rexp(n = 100, rate = 1)
  x_bar100[i] <- mean(samplen100) # mean n = 100
}

# from the bootstrap_ci lecture
par(mfrow = c(1,3)) # arrange plots in a grid
hist(x_bar5,
     xlab = expression(bar(X)),
     col = "skyblue",
     main = "Distribution of Mean \n for n = 5")
abline(v = mean(x_bar5), col = "red", lwd = 2, lty = "dashed")
legend("topright", legend = paste("Mean = ", round(mean(x_bar5), 3)))


hist(x_bar30,
     xlab = expression(bar(X)),
     col = "skyblue",
     main = "Distribution of Mean \n for n = 30")
abline(v = mean(x_bar30), col = "red", lwd = 2, lty = "dashed")
legend("topright", legend = paste("Mean = ", round(mean(x_bar30), 3)))

hist(x_bar100,
     xlab = expression(bar(X)),
     col = "skyblue",
     main = "Distribution of Mean \n for n = 100")
abline(v = mean(x_bar100), col = "red", lwd = 2, lty = "dashed")
legend("topright", legend = paste("Mean = ", round(mean(x_bar100), 3)))
```



5

As the sample size increases, the shape of the sampling distribution becomes more normally distributed. The Central Limit Theorem for sample means asserts that, as the sample size becomes large, the distribution of the sample means approaches a normal distribution, even if the population from which the samples are drawn is not normally distributed.

6

Sample Size	Mean (sd)	Theoretical mean (sd)
n = 5	1 (0.45)	1 (0.447)
n = 30	1 (0.18)	1 (0.183)
n = 100	1 (0.1)	1 (0.1)

7

The mean for all the samples are 1 which is equal to the theoretical mean. The sd when n = 5 is 0.45 which is slightly larger than the theoretical of 0.447. This suggests a small deviation across the spread of the distribution. When n = 30 the sd is 0.18 which is very close to the theoretical sd of 0.183. Lastly, for the sample size of n = 100 the sd and theoretical sd are both 1.

8

```
set.seed(11101987) # scorpio
sample10 <- rexp(n = 10, rate = 1)

mean(sample10) + c(-1,1) + qnorm(0.975) * 1/sqrt(10) # using the normal distribution (incorrect method)
```

```
## [1] 0.1826339 2.1826339
```

```
mean(sample10) + c(-1, 1) + qt(0.975, 9) * 1/sqrt(10) # using the t-distribution (correct method)
```

```
## [1] 0.2781958 2.2781958
```

9

The t-distribution is necessary when dealing with small sample sizes and an unknown population standard deviation. This is because it accounts for the added uncertainty that arises from using the sample standard deviation as an estimate. This is because the t-distribution has heavier tails, reflecting the increased variability. If we were to use the normal distribution, we would be underestimating the variability, resulting in a misleadingly narrow confidence interval.

10

The American Statistical Association (ASA) article on p-values emphasizes the correct interpretation and use of p-values in scientific research. It highlights how p-values, often misinterpreted, are not meant to be used in isolation to draw definitive conclusions. Many scientists have critiqued p-values, suggesting that over-reliance on them can lead to misinformed decisions, especially when they are treated as a strict threshold for significance. The ASA stresses that p-values can be valuable if contextualized within the study's design and paired with other forms of evidence. Importantly, the statement urges researchers to avoid binary decision-making based souley on p-values and instead adopt a more holistic view of statistical inference. This means integrating scientific reasoning, transparency, and the real-world context of the data and evaluating results. The ASA concludes that no single statistical measure should replace careful, thoughtful scientific reasoning.

11

I would say the most difficult part of this assignment was creating the for loops in order to generate random samples of varying sizes. I was essentially trying to reproduce what I had done in question 2. A classmate had suggested that I take a look at the bootstrap lecture though. There is a great example of this on the bootstrap lecture.

12

E - Excellent