Lecture 12: Central Limit Theorem

For each question below, write the R code you would use and record the result when appropriate. Answer in complete sentences when asked to explain.

Part 1: Populations, Samples, and Sampling Distributions

1. In your own words, explain the difference between a population and a sample.

2. In your own words, explain what a random variable is.

3. Suppose we take a random sample from a population and calculate the sample mean

   (a) What does the sample mean ($\bar{x}$) tell us?

   (b) If we take a different random sample from the same population, should we expect to get the exact same sample mean? Explain briefly.

4. In your own words, explain the difference between a sample and a sampling distribution.

Part 2: Looking at a Skewed Population

5. We will begin by generating values from a strongly right-skewed population. Specifically, we will use an Exponential distribution with rate = 1.

```
population <- rexp(10000, rate=1)
hist(population, main="Population Distribution", xlab="Value")

mean(population)
sd(population)
```

   (a) Describe the shape of the population distribution.

   (b) Record the mean and standard deviation of the population values.

   (c) Does the population distribution look approximately normal? Explain briefly.

6. For an Exponential distribution with rate = 1, the theoretical mean and theoretical standard deviation are both 1.

   (a) Is your simulated mean close to 1?

   (b) Is your simulated standard deviation close to 1?

   (c) Why do you think the simulated values are close to the theoretical values, but not exactly equal to them?

Part 3: Sampling Distributions from a Skewed Population

7. Now create 1,000 samples, each with sample size 5. For each sample, calculate the sample mean.

```
samples_5 <- matrix(rexp(5*1000, rate=1), nrow=5, ncol=1000)
means_5 <- apply(samples_5, MARGIN=2, FUN=mean)
hist(means_5, main="Sampling Distribution: n = 5", xlab="Sample Means")

mean(means_5)
sd(means_5)
```

   (a) What does one column in `samples_5` represent?

   (b) What does one value in `means_5` represent?

   (c) Describe the shape of the sampling distribution when $n = 5$.

   (d) Record the mean and standard deviation of the sampling distribution.

   (e) The theoretical standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$. Since $\sigma = 1$, calculate $\frac{1}{\sqrt{5}}$.

   (f) How does your simulated standard deviation compare to $\dfrac{1}{\sqrt{5}}$?

8. Repeat the same process with sample size 15.

   (a) Describe the shape of the sampling distribution when $n = 15$.

   (b) Record the mean and standard deviation of the sampling distribution.

   (c) Record the theoretical standard deviation, $\frac{1}{\sqrt{15}}$.

   (d) Compared to $n = 5$, does the sampling distribution look more normal, less normal, or about the same? Explain briefly.

9. Repeat the same process with sample size 30.

   (a) Describe the shape of the sampling distribution when $n = 30$.

   (b) Record the mean and standard deviation of the sampling distribution.

   (c) Record the theoretical standard deviation, $\frac{1}{\sqrt{30}}$.

   (d) How does this sampling distribution compare to the original skewed population distribution?

10. Repeat the same process with sample size 100.
    (a) Describe the shape of the sampling distribution when $n = 100$.
    (b) Record the mean and standard deviation of the sampling distribution.
    (c) Record the theoretical standard deviation, $\frac{1}{\sqrt{100}}$.
    (d) What happens to the spread of the sampling distribution as the sample size increases?

Part 4: Comparing the Sampling Distributions

11. Fill in the table below using your results from Part 3.

Sample Size	**Mean of Sample Means**	**SD of Sample Means**	**Theoretical SD**
5			
15			
30			
100			

12. As the sample size increases, what happens to the mean of the sampling distribution?

13. As the sample size increases, what happens to the standard deviation of the sampling distribution?

14. Which sampling distribution looked the most normal: $n = 5$, $n = 15$, $n = 30$, or $n = 100$? Explain briefly.

15. In your own words, explain why the sampling distribution can look approximately normal even when the original population distribution is strongly skewed.

Part 5: The Central Limit Theorem

16. The Central Limit Theorem tells us that, for a large enough sample size, the sampling distribution of the sample mean is approximately normal, regardless of the shape of the population distribution. In your own words, explain what this means.

17. The Central Limit Theorem also tells us that: $E(\bar{X}) = \mu$ and $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

    (a) What does $E(\bar{X}) = \mu$ tell us about the center of the sampling distribution?

(b) What does $SD(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$ tell us about the spread of the sampling distribution?

(c) Why does increasing n make the sampling distribution less spread out?

Part 6: Real Data Example with Baseball Home Runs

18. Now we will look at the `homeruns` variable from the `mlb_teams` dataset in the `openintro` library.

```
library(openintro)
data(mlb_teams)

hr <- na.omit(mlb_teams$homeruns)
hist(hr, main="MLB Team Home Runs", xlab="Home Runs")

mean(hr)
sd(hr)
length(hr)
```

(a) What does one value in `hr` represent?

(b) Describe the shape of the home runs distribution. Does it look normal?

(c) Record the mean and standard deviation of `hr`.

19. Create 1,000 samples, each with a different sample size. Then fill in the table below.

```
samples <- 1000
sample_size <- ???
hr_samples <- matrix(sample(hr, samples*sample_size, replace=TRUE),
                     nrow=sample_size)
hr_means <- apply(hr_samples, MARGIN=2, FUN=mean)

hist(hr_means, main="Sampling Dist: n = ???", xlab="Sample Means")
```

Sample Size	**Theoretical Mean**	**Theoretical SD**	**Mean of Sample Means**	**SD of Sample Means**
1				
3				
7				
10				
20				

(a) As the sample size increases from 1 to 20, what happens to the shape of the sampling distribution? Does it begin to look more normal? Explain briefly.

(b) Compare the simulated standard deviations to the theoretical standard deviations in the table. Are they reasonably close? What does this tell us about the Central Limit Theorem?