Week 3 In Class Exercises – Sampling Distributions

Instructions: In the lectures for this week, we created sampling distributions by building a simulated population from random data and drawing samples from that population. By examining large collections of samples, we can understand what to expect when we make inferences from samples of real data. In this exercise, we create another simulated population: scores on an achievement test. Achievement tests are often calibrated so that the population mean is 100 and the population standard deviation is 10. The code develops a simulated population of N=100,000 test takers, each of whom scored somewhere between 60 and 140.

1. Here's code to create a simulated population and show a histogram. Add a comment describing the shape of this distribution.

```
set.seed(1234)                             # Control randomization
testPop <- rnorm(100000, mean=100, sd=10)# Make simulated pop
hist(testPop)
```

2. This next code marks quantiles on the raw data histogram: the 0.025 quantile and the 0.975 quantile. Also write two lines of code that display the actual values.

```
hist(testPop)
abline(v=quantile(testPop, probs=0.025),col="green") # Lower tail
abline(v=quantile(testPop, probs=0.975),col="green") # Upper tail
```

3. The area between the green lines the "central region" and the areas outside of the green lines are the "tails." Add a comment to say what percentage of cases occur in the central region. What percentage falls in the lower tail, below the 0.025 quantile in the histogram? What percentage falls in the upper tail, above the 0.975 quantile?

4. Build a sampling distribution of means with these two lines of code:

```
# Custom function to pull one sample of size n
sampleTestScores <- function(n){sample(testPop,size=n,replace=TRUE)}

# Write a comment explaining this line
samplingDistribution <- replicate(1000, mean(sampleTestScores(100)))
```

5. Display histograms of the original distribution and the sampling distribution together:

```
par(mfrow=c(2,1)) # Divide the plot area horizontally
hist(testPop, xlim=c(50,140))
hist(samplingDistribution, xlim=c(50,140))
par(mfrow=c(1,1)) # Restore the plot area
```

The takeaway message is that the sampling distribution of means (lower histogram) converges on the same mean as the population (upper histogram) but the sampling distribution (lower histogram) is much less dispersed. The smaller dispersion results from the corrective influence of having many sampled observations contribute to each sample mean.

Week 3 In Class Exercises – Sampling Distributions

In the case studies below, we will <u>only</u> be examining sampling distributions of means. For each case study, create a new histogram of the sampling distribution and mark it like this:

```
hist(samplingDistribution)
abline(v=quantile(samplingDistribution, probs=c(0.025,0.975)))
```

For each case study, mark the specified sample mean, using a different colored line. Each of the sampling distributions has a story that goes with it, where researchers worked with a new group of test takers <u>who may or may not be similar to the calibration population</u>.

Case Study A:
A sample of 100 students from Syracuse, NY took a standardized test. The sample mean for this group of students was $\bar{X} = 101$. Add code/comments to answer these questions.

1. What is the mean of the sampling distribution?
2. What is the lower bound of the central region of the sampling distribution?
3. What is the upper bound of the central region of the sampling distribution?
4. Where does the new mean of $\bar{X} = 101$ fall on the sampling distribution?
5. Was the new sample drawn from the population that created the sampling distribution?

Case Study B:
A sample of 100 students from Utica, NY took a standardized test. The sample mean for this group of students was $\bar{X} = 96$. Add code/comments to answer these questions.

1. Where does the new mean of $\bar{X} = 96$ fall on the sampling distribution?
2. Was the new mean drawn from the population that created the sampling distribution?

Case Study C:
In the final stages of an FDA trial, researchers gave dietary supplements to a sample of **n=500** students to improve their ability to memorize words. **Note the change in sample size. This means that you will have to recreate samplingDistribution with the new larger sample size**. Modify the second line of code from problem 4 to accomplish this.

The n=500 students then took the same standardized test as in the previous two case studies and the sample mean for these students was $\bar{X} = 101$.

1. Mean of the sampling distribution?
2. Lower bound of the central region?
3. Upper bound of the central region?
4. Where does the new mean fall?
5. Was the new mean drawn from the population that created the sampling distribution?
6. Why is this result different from Case Study A?