**IST772 Week 3 Breakout – Sampling Distributions**

In Chapter 3 you learned how to create sampling distributions by building a simulated population from random data and then 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 a test of achievement. Achievement tests like this one are often calibrated so that the population mean is 100 and the population standard deviation is 10.

Note: You will see the terms cases and observations used in the Breakout. In general, these are interchangeable and refer to data values in the distribution.

1. The code below develops this simulated population of N=100,000 test takers, each of whom scored somewhere between 60 and 140 on the test.

set.seed(1234) # Control randomization

testPop <- rnorm(100000, mean=100, sd=10) # Make simulated pop

hist(testPop)

**Paste the histogram and add a comment describing the shape of this distribution.**

Chart, histogram

Description automatically generated

* The histogram has a bell shape, with most values centered around 100, which is the mean.

1. These next lines of code will mark the 1st, 2nd, and 3rd quartiles.

hist(testPop)

abline(v=quantile(testPop, probs=c(0.25,0.5,0.75)), col="blue")

**a. Paste the histogram below.**

Chart, histogram

Description automatically generated

**a. The second quartile has a special name. What is it?**

The second quartile is also known as Median of the dataset.

**b. What does this code give you?** quantile(testPop,c(0.25,0.5,0.75))

Graphical user interface, text, application

Description automatically generated

* This code splits the population into three quartiles – 25th, 50th and 75th.

**c.** **What percentage of cases fall below the 1st quartile? (No work needed.)**

* This implies that 25% of the values falls below the first quartile (93.2941)

**d. What percentage of cases fall above the 3rd quartile? (No work needed.)**

> 25% of the values falls above the 3rd quartile (106.7399).

**e. Add code that calculates the number of observations in each of the 4 sections of the distribution. Report how many are in each section.**

Graphical user interface, text, application

Description automatically generated

We can observe that each quartile has 25000 values.

1. The 1st and 3rd quartiles in that histogram seem very close to the middle! **Can that be correct?**

* Yes, the 1st and 3rd quartiles being close to the middle means that most of the values (more than half) are centered around the mean, i.e., most values fall near the second quartile.

1. Now we consider two additional quantiles, specifically the 0.025 quantile and the 0.975 quantile.

**a. Use the code below to generate the testPop distribution with these quantile lines and paste the graph below.**

hist(testPop)

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

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

The area between these quantiles is known as the “**central region**” and the two areas outside of the green lines are the “**tails**.”

Chart, histogram

Description automatically generated

**b. What percentage of cases occur in the central region.**

* **95%** cases occur in central region.

**c. What percentage of cases fall in the lower tail?**

* **2.5%** cases fall in the lower tail.

**d. What percentage of cases fall in the upper tail?**

* **2.5%** cases fall in the upper tail.

1. Run the following code:

hist(testPop) #generates population histogram

testPopMean <- mean(testPop) #finds the mean of the distribution

sumSq <- sum((testPop-testPopMean)^2) #finding the sum of the sq. deviations

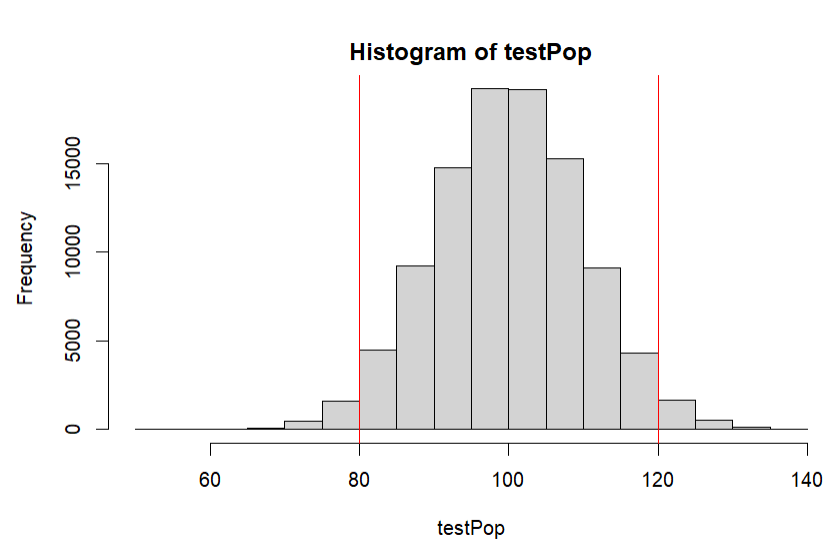
testPopVar <- sumSq/length(testPop) #finding the variance

testPopSD <- sqrt(testPopVar) #finding the standard deviation

abline(v=(testPopMean - 2\*testPopSD), col="red") #red line 2 sd’s below mean

abline(v=(testPopMean + 2\*testPopSD), col="red") #red line 2 sd’s above mean

1. **Add a comment to each line explaining what the code is finding. Paste the histogram below.**

****

**b. How do these red quantile lines and the green lines from #4 compare?**

* The upper and lower 2.5% quantiles nearly aligns with the 2 standard deviations lower than and higher than the mean.

**6. Share your answers to the following in the Week 3 Discussion.**

* 1. **#3**
  2. **#4b**
  3. **#5b**