Instructions: In the asynchronous material for this week, we created and made sense of sampling distributions by creating 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 first live exercise, we’ve used a few lines of code to set up another simulated population: scores on a test of achievement. Achievement tests like this one are calibrated so that the population mean is 100 and the population standard deviation is 10. The diagrams below depict a simulated population of N=100,000 test takers, each of whom scored somewhere between 60 and 140 on the test. Using the diagrams, answer the following questions about the simulated population:

1. Here’s a histogram of the population. What is the shape of this distribution and the formal statistical name for that shape:  
     
   *[Answer] – Shape is bell shaped/symmetrical distribution. Normal distribution.*
2. In this next histogram, I’ve marked the 1st, 2nd, and 3rd quartiles. Define the term quartile and estimate the values associated with these quartiles. The second quartile (also known as the 50th percentile) has a special name. What is it?  
   

*[Answer] – Quartile definition: Quartiles are either points in the data where 25%, 50% or 75% of observations lie below each quartile.*

*2nd quartile or 50th percentile is called the Median.*

*1st quartile = 93.18*

*3rd quartile = 106.711*

1. The 1st and 3rd quartiles in that histogram are very close to the middle. Can that be right? What percentage of cases fall below the 1st quartile? Above the 3rd quartile?   
   *[Answer] – Yes, the first and third quartiles can be close to middle.*

*25% values/obs are < 1st quartile*

*25% values/obs are > 3rd quartile*

1. By definition, this achievement test has a population standard deviation of 10. I’ve used abline() to mark points that are one and two standard deviations below the mean. Examine the histogram below and report what value is two standard deviations below the mean and what value is two standard deviations above the mean:  
     
   *[Answer] – 80 and 120*
2. Now I’ve marked some different quantiles, specifically the 0.025 quantile and the 0.975 quantile. Estimate where these quantiles fall on the X-axis. What percentage of cases occur between the 0.025 quantile and the 0.975 quantile?   
     
   

*[Answer] – 2.5% are below and above the 0.025 and 0.975 quantile (quantiles are expressed in decimal – so, 2.5% is 2.5/100 = 0.025)*

*> quantile(testPop, probs=0.025)*

*2.5%*

*80.38399*

*> quantile(testPop, probs=0.975)*

*97.5%*

*119.6717*

1. The area between the green lines is known as the “central region” and the two areas outside of the green lines are the “tails.” What percentage of cases fall in the lower tail, that is, below the 0.025 quantile in the histogram above? What percentage of cases fall in the upper tail, that is, above the 0.975 quantile in the histogram above?

*[Answer] – Below and above the 0.025 and 0.975 are 2.5% cases/obs.*

For your reference, here is the code that generated these histograms and associated markings:

set.seed(1234) # Control randomization

testPop <- rnorm(100000, mean=100, sd=10) # Create simulated population

testPopMean <- mean(testPop) # Calculate mean

sumSq <- sum((testPop-testPopMean)^2) # Sum of squares

testPopVar <- sumSq/length(testPop) # Variance

testPopSD <- sqrt(testPopVar) # Standard deviation

# Markings of quartiles

hist(testPop)

abline(v=quantile(testPop, probs=0.25),col="orange") # First quartile

abline(v=quantile(testPop, probs=0.5),col="orange") # Second quartile

abline(v=quantile(testPop, probs=0.75),col="orange") # Third quartile

# Show the numbers to verify the number of cases

quantile(testPop, probs=0.25) # The specific X value

quantile(testPop, probs=0.75) # The specific X value

length(testPop[testPop<quantile(testPop, probs=0.25)]) # Number of cases below

length(testPop[testPop>quantile(testPop, probs=0.75)]) # Number of cases above

# Markings of one and two standard deviations

hist(testPop)

abline(v=90, col="red")

abline(v=80, col="blue")

abline(v=110, col="red")

abline(v=120, col="blue")

# Markings of the central region and tails

hist(testPop)

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

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