**Random Variables Exercises**

**Exercises**

For these exercises, we will be using the following dataset:

library(downloader)

url <- "https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/femaleControlsPopulation.csv"

filename <- basename(url)

download(url, destfile=filename)

x <- unlist( read.csv(filename) )

Here x represents the weights for the entire population.

1. What is the average of these weights?

x <- unlist(population)

mean(x) = 23.89338

1. After setting the seed at 1, set.seed(1) take a random sample of size 5. What is the absolute value (use abs) of the difference between the average of the sample and the average of all the values?

diff <- mean(s) - mean(x)

diff = -0.2753778

abs(diff) = 0.2753778

1. After setting the seed at 5, set.seed(5) take a random sample of size 5. What is the absolute value of the difference between the average of the sample and the average of all the values?

# Repeat with set seed to 5

set.seed(5)

s2 <- sample(x, 5, replace = TRUE)

s2

diff2 <- mean(s2) - mean(x)

abs(diff2) = 0.4113778

1. Why are the answers from 2 and 3 different?
   * A) Because we made a coding mistake.
   * B) Because the average of the x is random.
   * C) Because the average of the samples is a random variable.
   * D) All of the above.
2. Set the seed at 1, then using a for-loop take a random sample of 5 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of x ?
3. We are now going to increase the number of times we redo the sample from 1,000 to 10,000. Set the seed at 1, then using a for-loop take a random sample of 5 mice 10,000 times. Save these averages. What percent of these 10,000 averages are more than 1 ounce away from the average of x ?
4. Note that the answers to 4 and 5 barely changed. This is expected. The way we think about the random value distributions is as the distribution of the list of values obtained if we repeated the experiment an infinite number of times. On a computer, we can’t perform an infinite number of iterations so instead, for our examples, we consider 1,000 to be large enough, thus 10,000 is as well. Now if instead we change the sample size, then we change the random variable and thus its distribution.

Set the seed at 1, then using a for-loop take a random sample of 50 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of x ?

1. Use a histogram to “look” at the distribution of averages we get with a sample size of 5 and a sample size of 50. How would you say they differ?
   * A) They are actually the same.
   * B) They both look roughly normal, but with a sample size of 50 the spread is smaller.
   * C) They both look roughly normal, but with a sample size of 50 the spread is larger.
   * D) The second distribution does not look normal at all.
2. For the last set of averages, the ones obtained from a sample size of 50, what percent are between 23 and 25?
3. Now ask the same question of a normal distribution with average 23.9 and standard deviation 0.43.

The answer to 9 and 10 were very similar. This is because we can approximate the distribution of the sample average with a normal distribution. We will learn more about the reason for this next.