# Homework (HW04) Sampling and Decisions

### General Instructions

For this homework you will upload 1 R file into blackboard.

We will be using sampling and replication:

* 1. Sampling is a process of drawing elements from a larger set. In data science, when analysts work with data, they often work with a sample of the data, rather than all of the data (which we call the population), because of the expense and inconvenience of obtaining all of the data. Imagine trying to collect data on the number of apps installed on every smart phone in the world. It would be much easier, and nearly as good, to work with just a sample of smart phones and calculate some statistics from that sample.
  2. One must be careful, however, because statistics from a sample usually fail to exactly match the characteristics of the population.

Reminder:

* All HW must start with an Identification Block like this sample…

################################################

# IST 387/687, Standard Homework Heading

#

# Student name:

# Homework number:

# Date due:

#

# Attribution statement: (choose the statements that are true)

# 1. I did this work by myself, with help from the book and the professor

# 2. I did this work with help from the book and the professor and these Internet sources: <provide the urls>

# 3. I did this work with coaching from <Name of another student> but did not cut and paste any code

# Run these three functions to get a clean test of homework code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list=ls()) # Clear all user objects from the environment!!!

# Set working directory

# Change to the folder containing your homework data files

setwd("~/MyDesktop/ISTX87/Homework")

### HW04

**Step 1: Cleaning up NAs from the airquality dataframe.**

A simple method of dealing with small amounts of missing data in a numeric variable is to substitute the mean of the variable in place of each missing datum.

1. The last line from your Prep Ex efforts should be myAQdata <- airquality followed by   
   View( myAQdata)
2. Execute the following command… myAQdata$Ozone[is.na(myAQdata$Ozone)]
3. Write a comment describing the results and how that statement works. You may want to run the command is.na(myAQdata$Ozone), or you may want to do a view of your dataframe to try to understand what the command does. Describe your understanding below.
4. Adapt that statement above to assign the mean of the Ozone variable to these missing values. First, you’ll want to calculate the mean of all of the Ozone instances.   
   *Reminder: you’ll have to do this by omitting the ones with an NA. (HINT: use na.rm=TRUE within the mean function). You should have both sides of the equation now (A <- C).*
5. Perform the same task on the one other variable that needs it (i.e., is there another column with NAs)?

**Step 2: using a package to replace NAs**

1. Install the package imputeTS by doing the following commands:

install.packages("imputeTS")

library(imputeTS)

1. Note that you only have to do the install.packages once, but you might have to library every time you are running the code in a clean environment.
2. Now, create myAQdata1, and then use the na\_interpolation( ) function, instead of mean substitution in your four variables. Don’t forget to rerun Step 1.A, because if you have already done mean substitution successfully there will not be any NAs left in the myAQdata data frame for na\_interpolation( ) to work on.
3. Compare the first 5 rows of myAQdata and myAQdata1. In a block comment, explain what you see.

**Step 3: Sampling**

1. Sample 10 observations from myAQdata$Wind and use printVecInfo( ) to display the characteristics of the resulting sample. Next display those 10 observations via a histogram. **In a comment, describe the purpose of an optional argument that you can supply to sample( ), replace=FALSE or replace=TRUE (and what was the value when you did the sample).**
2. Repeat what you did for Question 1.A three additional times. Each time that you create a new sample, run the resulting data through printVecInfo( ) and generate a histogram. Explain in a block comment why each result is different.

**Step 2: Replicating our samples**

1. Use the replicate function, to replicate the sampling of Wind (the sampling that was described above) combined with calculating the mean of the sampled Wind attribute. Replicate the sampling and mean calculation 200 times, then display the results of the 200 means as a histogram.  
   *If you need a hint, see Chapter 10 in the book.*
2. Repeat step 2A two more times.
3. Using a block comment, explain why the histograms generated in Step 1 are different than the histograms generated in Step 2

***You must submit all Homework to blackboard prior to the deadline specified for each assignment.***

Late HW assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*