**Module 7/8 Lab Activity: Statistical inference in SLR & Multivariate correlations**

**PSY 652 Research Methods**

**Oct 23, 2019**

**Activity #1: Inferences in SLR”**

The Cherry Blossom 10 Mile Run is a road race held in Washington, D.C. in April each year. (The name comes from the famous cherry trees that are in bloom in April in Washington.) The results of this race are published. This data frame contains the results from the 2005 race.

The dataframe, called TenMileRace, represents all 8636 runners and includes the following variables:

* *state*: State of residence of runner.
* *time*: Official time (in seconds) from starting gun to finish line.
* *net*: The recorded time (in seconds) from when the runner crossed the starting line to when the runner crossed the finish line. This is generally less than the official time because of the large number of runners in the race: it takes time to reach the starting line after the gun has gone off.
* *age*: Age of runner in years.
* *sex*: A factor with levels F M.

1. Download the “TenMileRace.csv” dataset from the Module 7 Lab dropbox folder and save it into your project folder.
2. Create a new R notebook from your project file and name it “Ten\_Mile\_Race”
3. Create a new R chunk with a first level header: “Load Libraries”
   1. load the tidyverse & psych packages in this R chunk
4. Create a new R chunk with a first level header: “Import Data”
   1. read in the “TenMileRace.csv” dataset, assign it to an object named “population”
5. Write a first level header: “Describe Data”
   1. Use your preferred method to describe the population dataset
6. Create a first level header: “Explore the principles of the sampling distribution”
7. Create a second level header: “Store population mean”
   1. Create an R chunk: Calculate the mean of the variable *net* and save this to a new object named “mean\_net\_time”
8. Create a second level header: “Create a histogram of net time across all people in the population”
   1. Create an R chunk: Create a histogram of the variable *net*. Set the binwidth to 60. Title your histogram “Net time for race completion in seconds across participants.” Title your x axis “Time in seconds".
9. Create a second level header: “Draw 5 random samples from the population”
   1. Create an R chunk: Pull 5 random samples of 50 (with replacement) from the population dataframe and save them each in a separate object.

Hint: To pull a random sample, use the following code sample:

*object\_name* <- sample\_n(*dataframe\_name*, *sample\_size,* replace=TRUE)

Do this 5 times, once for each random sample.

1. Create a second level header: “Calculate the mean in each random sample”
   1. Create an R chunk. Calculate the mean of *net* in each random sample you pulled (do this 5 times, once for each random sample).
   2. In the white space, write a sentence describing how the means vary across random subsamples of the original population.How much do these means differ from the original population mean? And by how many standard deviations (from the population dataset) do they differ?
      1. *Note:* Because these are random samples, your answers will vary.
2. Create a second level header: “Save all means to one dataframe”
   1. Create an R chunk: Create an object named means, which is a list (vector) of the five means you calculated. Then, convert *means* to a dataframe called *sampling\_dist* using the data.frame function.

Hint: use the following template for two lines of code.

*object\_name* <- c(*mean1, mean2, mean3, mean4, mean5)*

*Dataframe\_name <-* data.frame(object\_name)

1. Create a second level header: “Make a histogram of the random sample means”
   1. Create a histogram of the *means* variable from the *sampling\_dist* dataframe. Set the binwidth to 30. Title your histogram “Net time for race completion in seconds for 5 random subsamples of 50.” Title your x axis “Time in seconds".
   2. In the whitespace, write a sentence comparing the range of this histogram to the range of the original population histogram created step 8.
2. Create a second level header: “Draw 5000 random samples, obtain the mean in each, and plot.
   1. Create an R chunk. Copy and paste the following code (yes, really!) and click run.

**rep\_sample\_n <- function(tbl, size, replace = FALSE, reps = 1)**

**{**

**n <- nrow(tbl)**

**i <- unlist(replicate(reps, sample.int(n, size, replace = replace), simplify = FALSE))**

**rep\_tbl <- cbind(replicate = rep(1:reps,rep(size,reps)), tbl[i, , drop=FALSE])**

**dplyr::group\_by(rep\_tbl, replicate)**

**}**

**sample\_means <- population %>%**

**rep\_sample\_n(size = 50, reps = 5000, replace = TRUE) %>%**

**summarize(mean\_net = mean(net))**

**m <- as.matrix(summarise(sample\_means, m = mean(mean\_net)))**

**ggplot(sample\_means, aes(x = mean\_net)) +**

**geom\_histogram(binwidth = 5) +**

**geom\_vline(xintercept = mean\_net\_time, lwd = 1, color="deeppink") +**

**geom\_vline(xintercept = m, lwd = 0.5, color="gold") +**

**labs(title = "Average time across 5000 samples", subtitle = "pink line = population mean from original sample\ngold line = mean of means from 5000 random samples", x = "average time in the sample")**

*This code pulls 5000 random samples of 50 from the population dataframe. Then, we calculate the mean for net from each of the 5000 random subsamples, and plot the distribution of these means. The overall population mean for net is marked on the histogram with a pink line and the mean-of-means for the 5000 random samples is marked with a gold line.*

1. In the white space, write 1-2 sentences comparing the sampling distribution for the means you just created to the distribution of the variable in the full sample you plotted in step 8. Include consideration of how the mean, range, and shape of the curves differ between plots.
2. What does this mean practically for researchers? In the white space, write 1-2 sentences answering this question

**Activity #2: Multivariate correlations**

Answer the following questions in 2-3 sentences. Use the Module 8 lecture slides and Murphy, *In Press* reading for reference.

1. What are the pros and cons of conducting multivariate correlations that include multiple outcome variables? Why would you select this method over models that include univariate outcomes?
2. In your own words, why would you assign different weights to predictor and outcome variables in multivariate correlations? How might you adjust your weighting if some predictors are highly correlated?
3. Thinking about your own research interests, when would multivariate correlation and variable weighting be a good fit for answering your research question? Give an example of a research study you might use this for.