Homework 4 of STAT 540
Section 001, Fall 2024
Due: Wednesday Sep 25 (before class)
Total Points: 70

Please hand in a hard copy of your homework (compiled pdf file from R markdown) in class and email your R code to Kaniz Fatema (KFATEMA@email.sc.edu). Please use the R markdown Homework template (HWtemplate.Rmd) to write your homework solutions. Work on the homework independently.

Problem 1. This exercise is for practicing central limit theorem.

(a) Draw n=5 samples from uniform distribution and calculate sample means. Repeat this experiment 200 times, plot the distribution of sample means. (10 points) [Hint: To simulate n samples from uniform distribution, use runif(n). Use plot(density(x)), where x is the vector contains the sample means from these 200 experiments.]

(b) Repeat (a) but use n=100 (3 points).

(c) Compare the sample distributions obtained in (a) and (b), what do you observe? (10 points)

Problem 2. Use the murders data in the dslabs package and generate boxplots of the state populations by region. (4 points)

Problem 3. Examine the built-in dataset ChickWeight. Which of the following is true: (3 points)

a. ChickWeight is not tidy: each chick has more than one row.
b. ChickWeight is tidy: each observation (a weight) is represented by one row. The chick from which this measurement came is one of the variables.
c. ChickWeight is not tidy: we are missing the year column.
d. ChickWeight is tidy: it is stored in a data frame.

Problem 4. We will be using the data from the survey collected by the United States National Center for Health Statistics (NCHS). This center has conducted a series of health and nutrition surveys since the 1960's. Starting in 1999, about 5,000 individuals of all ages have been interviewed every year and they complete the health examination component of the survey. Part of the data is made available

via the NHANES package. Once you install the NHANES package, you can load the data like this**:**

>library(NHANES)

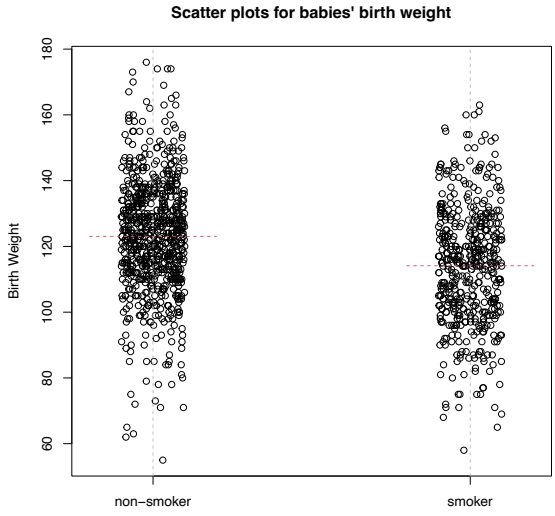
We will provide some basic facts about blood pressure. First let's select a group to set the standard. We will use 20-to-29-year-old females. AgeDecade is a categorical variable with these ages. Note that the category is coded like " 20-29", with a space in front! What is the average and standard deviation of systolic blood pressure as saved in the BPSysAve variable? Save it to a variable called ref. (5 points)

Problem 5. Permutation Test

We will be using the following dataset for this problem.

>url <- "https://people.stat.sc.edu/hoyen/STAT540/Data/baby.txt"
>babies <- read.table(file=url, header=TRUE)
We are going to perform a permutation test to compare babies' birth weight (**bwt**) between mothers who smoke (**smoke=1**) versus who do not smoke (**smoke=0**).

(a) Reproduce the following plot, what do you observe in this plot? (10 points)



[Hint: Use the function stripchart(y~x, method= "jitter", jitter=0.2, vertical=T, ylab=…, main=…) where y is the baby's birth weight (**bwt**) and x is mom's smoking status (**smoke**). Use ylab to label y-axis correctly and main to create a

main title. For the blue lines indicating means in each group, use lines(c(x1,x2), c(y1,y2), col=4) where x1, x2, y1, y2 are the locations of the line in x-axis and y-axis respectively.]

(b) Perform a student-t test for babies' birth weight between mothers who smoke versus who do not smoke. State your hypothesis and report your results. (3 points)

(c) Create a permuted dataset by randomly shuffling the group label (**mom's smoking status, smoke**), then perform t test using the permuted dataset. Report the t-statistic using the permuted dataset (named the permuted t-statistic). (3 points)

(d) Repeat (b) 1000 times, plot the density of the permuted t-statistics. Calculate a p value by counting how many permuted t-statistics (c) exceed the observed t-statistics in (b). (10 points)

(e) Plot the corresponding t-distribution on the plot generated in (d). (4 points)