Homework Assignment 1

DS4043, Spring 2023

Due on March 15, 2023 at 11:59 pm

Instructions: You need to fully show your explanations, codes, and results to get full credit. You will need to submit your R markdown file and the generated pdf file. Missing the R markdown file, you will get a 10% penalty. Missing pdf file, you will have no grades (Your TA will not knit pdf for you). Late submission will not be accepted.

1. a) Generate a random sample $X_1, ., X_{100}$ which is from a normal distribution with mean $\mu = 5$ and standard deviation $\sigma = 3$. Use set.seed(99) before random number generation.

   b) Write an R function 'fx' in R to implement the function $y = (x - a)/b$, which will transform an input vector x and return the output y. However, the function should take three input arguments x, a and b.

   c) Generate the random sample y using the function in b), where $x =$ the random sample generated in a), a=5 and b=3. What is the distribution of y? And explain. Then calculate the sample mean and standard deviation of y and compare them with the population mean and standard deviation. Note, please show your derivation of the population mean and standard deviation.

   d) Display a probability histogram of the random sample y and add an estimated probability density function to your histogram

   e) Add the true probability density function to your histogram in d)

2. We will use the dataset called `hflights`. This dataset contains all flights departing from Houston airports IAH (George Bush Intercontinental) and HOU (Houston Hobby). The data comes from the Research and Innovation Technology Administration at the Bureau of Transportation statistics: hflights. Make sure you have installed the packages `hflights` before suing them.

```
# Load packages
# install.packages("hflights")
library(hflights)
data(hflights)
```

(a) How many rows and columns of `hflights`? Get the names of the columns.

(b) Select the first 15 rows make it a data frame called `phflights`. Suppose we would like check three variables, DepTime, ArrTime and FlightNum. Select these three columns and call it `sflights`. Only Show the first few lines of `sflights`.

(c) Create a new column vector Called `BNum` indicating if the FlightNum is greater than 1000 and append this column to `sflights`. Show the first few lines.

(d) Compute the average arrival delay (ArrDelay) to each destination for `hflights`. (Hint: use na.rm = TRUE to remove missing values) Only show the first 10 results. Then for each carrier, calculate the percentage of flights cancelled or diverted.