STAT 408 Homework 1

Due by 11:55 pm, Sunday, 09/18/2022

50 points

Please provide detailed calculation and explanation in your solution. Points will be deducted for skimpily written answers. This homework will also require coding in R. On the coding part, the homework solutions should also include detailed description, R code, and output. Write your answers, scan them, and combine to a single pdf file. Name this file as yourname_hw1 and upload to Sakai.

1. (5 points) The pmf of the amount of memory X (GB) in a purchased flash drive is

x	1	2	4	8	16
$p(x)$	.05	.10	.35	.40	.10

Compute the following

a. E(X)

b. V(X) directly from the definition

c. The standard deviation of X

2. (5 points) Consider the following sample of observations on coating thickness for low-viscosity paint:

.83	.88	.88	1.04	1.09	1.12	1.29	1.31
1.48	1.49	1.59	1.62	1.65	1.71	1.76	1.83

a. Calculate a point estimate of the mean value of coating thickness, and state which estimator you used
b. Calculate a point estimate of the variance of coating thickness, and state which estimator you used

3. (5 points) A confidence interval is desired for the true average stray-load loss μ (watts) for a certain type of induction motor. Assume that stray-load loss is normally distributed with $\sigma = 3$.

a. Compute a 95% CI for μ when n = 25 and $\bar{x} = 58.3$
b. Compute a 95% CI for μ when n = 100 and $\bar{x} = 58.3$
c. Compute a 99% CI for μ when n = 25 and $\bar{x} = 58.3$

4. (5 points) To determine whether the pipe welds in a nuclear power plant meet specifications, a random sample of welds is selected, and tests are conducted on each weld in the sample. Suppose the specifications state that the mean strength of welds should exceed 100 lb/in^2

a. What hypotheses should be tested? Write down H_0 and H_a and explain your reason.
b. Describe type I and II errors in the context of this problem situation.

5. (10 points) Dataset births.csv contains the information for 1992 newborns and their parents.

a. Download the data set births.csv from Sakai, set your working directory, and import it into RStudio. Name the data frame as NCbirths.

b. Extract the weight variable as a vector from the data frame and name it as weights. What units do you think the weights are in?

c. Create a new vector named weights_in_pounds which are the weights of the babies in pounds. You can look up conversion factors on the internet.

d. Print the first 20 babies' weight in pounds.

e. What is the mean weight of all babies in pounds?

f. The habit variable records the smoking status for mothers of each baby. What percentage of the mothers in the sample smoke? Hint: consider table() function.

g. According to the Centers for Disease Control, approximately 14% of adult Americans are smokers. How far off is the percentage you found in (b) from the CDC's report?

6. (10 points) The dataset flint.csv records the water pollution levels in different locations at Flint, Michigan.

a. Download the flint.csv from Sakai and read it into R. When you read in the data, name your object “flint”.

b. The EPA states a water source is especially dangerous if the lead level (Pb) is 15 PPB or greater. What proportion of the locations tested were found to have dangerous lead levels?

c. Report the mean copper level for only test sites in the North region.

d. Report the mean copper level for only test sites with dangerous lead levels (at least 15 PPB).

e. Report the mean lead and copper levels for all locations.

f. Create a box plot with a good title for the lead levels. Hint: consider boxplot() function.

g. Based on what you see in part (f), does the mean seem to be a good measure of center for the data? Report a more useful statistic for this data.

7. (10 points) We will use a simulation study to show central limit theorem.

a. Set random seed to 2022. Use hist() function to plot a histogram on the weight variable in the NCbirths. Do you think weight follows a normal distribution? Why?

b. Use sample() function to randomly select 10 observations from weight. Show the mean of these 10 observations.

c. Use a for loop to repeat the (b) 1000 times. Save 1000 means in a vector. Show the histogram for 1000 means. Is this distribution close to normal?

d. Change the sample size 10 in (b) to 30 and 100, Repeat (c) for these two sample sizes. Are these two distributions close to normal? Interpret your reason.