Statistics 147 Assignment #2

Summer 2020; 60 pts

DUE DATE: Wednesday, July 1, 2020, by 6:00 pm

GENERAL INSTRUCTIONS:

- Your write-up should be neat, well-organized and concise but complete. Your write-up should include a cover page which includes your name, the last 4 digits of your student ID. Your write-up must be typed. Use a word processor such as Word or LaTeX, etc. No hand-written papers will be accepted for credit!
- Be sure to include copies of all charts and graphs.
- You will need the following data files, that is available for download on **Ilearn** under **Data Files**

  `agility.dat`

- For your SAS session, name your program file *hwk2_su20.sas*.

  (i) **(2 pts)** Include the following titles:

```
title1 'Statistics 147 Assignment #2';
title2 'Summer 2020';
title3 'Your Name';
title4 'SAS Question XX'    where XX = question number
title5 'SubPart (xx)';     /* where xx is something like ii */
```

  (ii) **(3 pts)** Be sure to include documentation in your SAS program.

  (iii) **(2 pts)** Be sure to a full copy of your SAS program at the end!

- For your R session,

  ♠ **(2 pts)** include the following titles (as comments)

```
# Statistics 147 Assignment #2
# Summer 2020
# Your Name
# R Question ZZ (where ZZ = question number)
# Subpart XX (where XX = subpart number, something like ii)
```

  ♠ **(2 pts)** Be sure to include a full copy of your R code at the end.

  ♠ **(2 pts)** Be sure to include descriptive documentation/comments in your R code.

The Questions

R

1. (Use **R** for this problem!) **(13 pts total)** A production process that creates specialized dog collars operates with 7% nonconforming (defective) output. A sample of 12 collars is selected at random and the number of nonconforming collars counted.

Let Y = number of nonconforming collars. Then $Y \sim Bin(n = 12, p = 0.07)$. Use this information to complete the following.

(i) **(1 pt)** Generate the sequence 0 through 12 and store the sequence in a variable named **y**. (Be sure to print the values of y.)

(ii) **(1 pt)** Generate the *probability distribution function* of Y and store the values the variable **pdf1**. (Be sure to print the values of pdf1.)

(iii) **(1 pt)** Generate the *cumulative distribution function* of Y and store the values in the variable **cdf1**. (Be sure to print the values of cdf1.)

(iv) **(1 pt)** Generate the values of $P(Y > y)$ and store the values in the variable **sdf1**. (Be sure to print the values of sdf1.)

(v) **(1 pt)** Use the **cbind** function to generate a data frame with y1, pdf1, cdf1 and sdf1 as the columns. (Be sure to print the data set.)

(vi) **(1 pt)** Find the probability that *exactly 3* of the 12 collars are nonconforming (defective).

(vii) **(1 pt)** Find the probability that *at most 3* of the 12 collars are nonconforming (defective).

(viii) **2 pts)** Find the probability that *between, and including*, 2 and 4 of the 12 collars are nonconforming (defective).

(ix) **(2 pts)** Generate a plot of y vs pdf1. Be sure to give your plot an appropriate title.

(x) **(2 pts)** Generate a plot of y vs cdf1. Be sure to give your plot an appropriate title.

2. **(6 pts total)** (Use **R** for this problem!) Suppose the filling process of a particular type of dry dog food bag is normally distributed with a mean of 40 pounds and a standard deviation of 0.50 pounds. The filling process is considered to be functioning at an appropriate level (functioning "in control") if the amount of fill in the dog food bags is between 38.75 pounds and 41.25 pounds. Ruihan selects a dog food bag at random from the assembly line.

   (i) **(1 pt)** Find the probability that the bag is under-filled.

   (ii) **(1 pt)** Find the probability that the the bag is not under-filled.

   (iii) **(1 pt)** Find the probability that the bag is over-filled or under-filled. (1 pt)

   (iv) **(2 pts)** Find the probability that the bag has a fill amount that meets specifications.

   (v) **(1 pt)** Find the 97^{th} percentile.

3. **(3 pts total)** (Use **R** for this problem!) Linda has three dogs (Cody, Dusty and Shadow) that she is training for a national agility championship. Linda records their times to finish the course (in seconds) for 30 runs. Luke does not believe there is a significant difference in mean finishing times between the three dogs. To test this claim, Luke takes three independent random samples of 8 times for each of the three dogs, yielding the following data:

```
File name: agility.dat
Cody Dusty Shadow
75      69      83
85      79      93
70      77      87
79      51      72
73      53      82
81      69      72
84      59      62
70      64      77
```

NOTE:

- ♠ The data is located in a datafile named **agility.dat**.
- ♠ The headings are included in the data file. The actual data begins on line 3.
- ♠ **Assume** the *Cody* is dog 1, *Dusty* is dog 2, and *Shadow* is dog 3.
- ♠ μ_i = true average finishing time for dog i and σ_i^2 = variance of the finishing time for dog i, $\quad i = 1, 2, 3$

(i) **(2 pts)** Write the R code to read the data into a data frame and then print it out.

(ii) **(1 pt)** Add the appropriate lines of code to make the columns of the data frame accessible individually and obtain the column headers/names.

SAS

1. **(5 pts total)** (Use **SAS** for this problem!) According to a recent survey, 65% of dog owners prefer a male dog. A random sample of 25 dog owners is selected.

   Let X = # of dog owners that prefer a male dog. Use this information to complete the following.

   (i) **(1 pt**) Find the probability that exactly 15 of the dog owners prefer a male dog.

   (ii) **(1 pt**) Find the probability that no more than 15 of the dog owners prefer a male dog.

   (iii) **(2 pts)** Find the probability that between 16 and 19, inclusively, of the dog owners prefer a male dog.

   (iv) **(1 pt)** Find the average number of dog owners that prefer a male dog.

2. **(6 pts total)** (Use **SAS** for this problem!) Suppose the filling process of a particular type of canned dog food is normally distributed with a mean of 13.25 ounces and a standard deviation of 0.10 ounces. The filling process is considered to be functioning at an appropriate level (functioning "in control") if the amount of fill in the dog food cans is between 13.15 ounces and 13.35 ounces. Ruihan selects a can of dog food at random from the assembly line.

   (i) **(1 pt)** Find the probability that the can is underfilled.

   (ii) **(1 pt)** Find the probability that the can is *not* underfilled.

   (iii) **(2 pts)** Find the probability that the can has a fill amount that meets specifications.

   (iv) **(1 pt)** Find the probability that the can is overfilled or underfilled (does not meet specifications).

   (v) **(1 pt)** Find the 97^{th} percentile.

3. **(5 pts)** (Use **SAS** for this problem!) Let Let $y = 3n + \sqrt{4.5m^{-0.5n}}$ for n = 1,3,5 and m = 2,5,8. Use nested DO loops to calculate the values of y. (Be careful of P.E.M.D.A.S. / order-of-operations when you code!)

4. **(13 pts total)** (Use **SAS** for this problem!) Linda has three dogs (Cody, Dusty and Shadow) that she is training for a national agility championship. Linda records their times to finish the course (in seconds) for 30 runs. Ruihan does not believe there is a significant difference in mean finishing times between the three dogs. To test this claim, Ruihan takes three independent random samples of 8 times for each of the three dogs, yielding the following data:

```
File name: agility.dat
Cody Dusty Shadow
75      69      83
85      79      93
70      77      87
79      51      72
73      53      82
81      69      72
84      59      62
70      64      77
```

NOTE:

- ♠ The data is located in a datafile named **agility.dat**.
- ♠ The headings are included in the data file. The actual data begins on line 3.
- ♠ **Assume** the *Cody* is dog 1, *Dusty* is dog 2, and *Shadow* is dog 3.
- ♠ μ_i = true average finishing time for dog i and σ_i^2 = variance of the finishing time for dog i, $\quad i = 1, 2, 3$

(i) **(4 pts)** Write the SAS code to read in and print out the data. Use **agility1** as your temporary SAS dataset name. Use nested do loops! **(NOTE: DO NOT COPY AND PASTE THE DATA INTO YOUR SAS PROGRAM. READ THE DATA IN FROM THE EXTERNAL DATA FILE!)** (5 pts)

(ii) **(2 pts)** Using if-then-else structures, name the dogs as follows:

dog	1	2	3
name	**Cody**	**Dusty**	**Shadow**

(iii) **(1 pt)**Add the appropriate lines of code to sort the data by the **name** of the dog.

(iv) **(2 pts)** Add the appropriate lines of code to your program to generate the mean, standard deviation and median for each of the dogs and complete the following table. You may need to create a table of your own in your writeup.

name	**Mean**	**Standard Deviation**	**Median**
Cody			
Dusty			
Shadow			

(v) **(3 pts)** Create a new temporary SAS data set, named **onlyC**, in which the data is restricted to the times of *Cody* only. (Be sure to print the data as a check.)

(vi) **(3 pts)** Create a new temporary SAS data set, named **bothDS**, in which the data is restricted to the times of **Dusty** and **Shadow** only. (Be sure to print the data as a check.)