STA380.18
Homework on Sufficiency, SAS, and Statistics Review

Directions: *Be sure to show your work and explain your answer for each question, even if the question seems to require only a Yes or No answer. Your homework solutions are to be entirely your own effort. You may not communicate with anyone about the homework, except for the TA and/or the instructor. You may use the Canvas postings, in-class discussion, any of the recommended textbooks, and computer software, if necessary, but no other resources. In writing up your solutions for software-based questions, it is recommended to support your answers with cut-and-pasted output, provided your answers are clearly* labeled *and* circled *or* highlighted*. The grader will not search through unlabeled computer output to try to find your answers.*

Note on submission: You should assemble your solutions as a single editable file – editable in order to facilitate TA grading annotation; one file in order to ensure that the parts do not get lost – in addition, multiple files may result in overwriting some of them. I suggest embedding all in a single Word file. Scan any hand-written solutions and include the scan(s) in your master submission file.

1. **[10 points]** Amazon fulfillment centers want to ensure a uniform (and low) processing time for orders. At one center, Amazon tracked a random sample of *n* orders and compared the actual processing time of each order against Amazon's standard. The amount of time *x* that an order departed early was recorded with a negative sign ($x < 0$) or late with a positive sign ($x > 0$). For the analysis, the following statistical model was used for the *x*'s: Suppose that $X_1, X_2, ..., X_n$ are independent random variables with common density function $f(x_i;\theta) = \frac{1}{\theta\sqrt{\pi}} e^{\frac{-x_i^2}{\theta^2}}$, for $-\infty < x_i < \infty$, $i = 1, 2, \ldots, n$, where $\theta > 0$ is an unknown parameter. A small value for θ would represent uniformity of processing times. Find a one-dimensional sufficient statistic for θ.

2. **[10 points]** Computers make small "machine" errors in floating point operations that can accumulate across complex calculations. As a test, a new computer chip was given a series of *n* complex calculations for which the answers were known. For each calculation, $i = 1, 2, \ldots, n$, the machine error x_i was recorded. Interest focuses upon the distribution of machine errors (mean, variance, maximum error, etc.) The following statistical model was adopted for the machine errors: Suppose that $X_1, X_2, ..., X_n$ are independent random variables with common density function $f(x_i;\theta) = \begin{cases} \frac{1}{2\theta} & \text{for } -\theta < x_i < +\theta \\ 0 & \text{otherwise} \end{cases}$, $i = 1, 2, \ldots, n$, where $\theta > 0$ is an unknown parameter. Find a one-dimensional sufficient statistic for θ and hence for the questions of interest. *[Hint: Note the limitations on the range of X.]*

3. **[10 points]** Suppose that $X_1, X_2, ..., X_n$ are independent random variables with common density function $f(x_i;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{\sigma^2}}$, for $-\infty < x_i < \infty$, i = 1, 2, …, n, where $-\infty < \mu < \infty, \sigma > 0$ are unknown parameters. Let $Y_1, Y_2, ..., Y_n$ be the ordered values of $X_1, X_2, ..., X_n$. That is, $Y_1, Y_2, ..., Y_n$ are $X_1, X_2, ..., X_n$ rearranged in order so that $Y_1 \le Y_2 \le \cdots \le Y_n$. Specifically, $Y_1 = \min(X_1, X_2, ..., X_n), ..., Y_n = \max(X_1, X_2, ..., X_n)$. Show that $Y_1, Y_2, ..., Y_n$ are sufficient statistics for μ, σ.
[Hint: Think outside the box! You may want to approach this question through the definition of sufficiency, rather than through the factorization theorem. For example, suppose n=3 and $y_1 = 1, y_2 = 2, y_3 = 3$. *Then what is the conditional probability that* $x_1 = 3, x_2 = 1, x_3 = 2$ *given that* $y_1 = 1, y_2 = 2, y_3 = 3$*? That is, if you know that your data are the values 1, 2, 3, what is the probability that they occurred in the order 3, 1, 2?]*

For questions 4-13, I suggest you plan your SAS program for all questions, rather than doing one question at a time. That will make your work most efficient.

Context for questions 4-13:

A random sample of 15 pizza outlets were selected from the national population of outlets of a large chain of retail pizza makers. Data on monthly sales (number of pizzas), price per pizza, and advertising expenses per outlet are in the file "pizza data.xls". Use SAS to answer the remaining homework questions. Assume the validity of the regression model (i.e., the linearity of means, homoscedasticity of variances, independence of observations, and normality of residual errors) as needed for questions 4-12 (Q13 asks you to test one model specification).

4. **[7 points]** For credit for question 4, include a printout of your SAS program with your homework submission.

   You should also include SAS output to support your answers to the remaining questions. But all output must be clearly labeled by question number and circled or highlighted. You may write your answers directly on the SAS output if you can do so legibly beside the appropriate part of the labeled output, or download and edit the output.

Hints: For importing the data into SAS:

- By default, SAS expects the first row of an input Excel file to be the variable names and row 2 to start the data. With appropriate coding, SAS can handle alternatives, but for simplicity, you may wish to modify the Excel file to make variable names come in row 1 and data start in row 2.

- With appropriate coding, SAS can also handle formatted input data (like dollar signs and commas). An import may work with formatted data. If it does not, you may wish to remove the formatting to facilitate the import.
- Copy and paste of data from Excel directly into the SAS program for use with CARDS type input often does not work well. The reason: Excel contains imbedded tab marks as column separators that SAS is not expecting. By default, SAS expects blanks as column separators. You can replace the tabs with blanks in Word. If you take this approach to getting the data into SAS, you would copy and paste from Excel into Word, then search and replace "^t" in Word (Word's code for identifying tabs) with blanks, then copy and paste from Word into the SAS program.
- Note that copy and paste directly into a *SAS dataset* (rather than into a SAS program) does not work at all. The reason: Because SAS can read a vast number of formats flexibly, SAS needs guidance from you on how you want the data read.
- You may want to add an observation to the dataset to help with some of the questions. This may be more straightforward to do in Excel than in SAS. (It is not hard in SAS – see the solutions for the Basic Statistics Review for an example.)

Hints: For downloading and printing Results:

- In SAS OnDemand, the Results window has three icons for downloading the Results window as an HTML file, as a PDF file, or as a Word file.
- Once the Results file has been downloaded to your computer, you can edit, cut and paste, print it out, etc.

5. **[7 points]** A new outlet is being planned with an allocated budget of $50,000 per month for advertising and planned average price of $10.00 per pizza. Estimate the monthly sales (number of pizzas) of this outlet. *[Hint: You may want that extra observation in the data set to answer this.]*

6. **[7 points]** By how much, plus or minus, do you expect your estimate in the preceding question will miss actual monthly sales for the outlet? *[Hint: You should be as precise as you can be for this amount.]*

7. **[7 points]** Can you give an interval in which you are approximately 95% confident that actual sales for the new outlet will lie? *[Hint: You should be as precise as you can be for this amount.]*

8. **[7 points]** Suppose the chain adopts a policy that all outlets will spend $50,000 per month on advertising and maintain an average price of $10.00. Estimate the mean monthly sales among all outlets in the chain.

9. **[7 points]** By how much, plus or minus, do you expect your estimate in the preceding question will miss actual mean monthly sales for the chain? *[Hint: You should be as precise as you can be for this amount.]*

10. **[7 points]** Can you give an interval in which you are approximately 95% confident that actual mean monthly sales will lie? *[Hint: You should be as precise as you can be for this amount.]*

11. **[7 points]** In a departmental review meeting, the marketing manager stated her opinion that every additional monthly dollar spent on advertising results in an increase in mean sales of about two additional pizzas per month if the average price is not changed. How many standard errors are between her opinion and the corresponding data estimate? Do you think her opinion is consistent with these data, if being within +/- 2 standard errors is considered consistent?

12. **[7 points]** In a departmental review meeting, the marketing manager stated her opinion that every additional monthly dollar spent on advertising results in an increase in mean sales of about two additional pizzas per month. How many standard errors are between her opinion and the corresponding data estimate? Do you think her opinion is consistent with these data, if being within +/- 2 standard errors is considered consistent? *[Hint: This is **not** a repeat of the preceding question.]*

13. **[7 points]** Test normality of the residuals in the regression that you ran to answer the preceding question. Use significance level of 0.05.