STA280 Unsupervised Learning
Homework on Sufficiency

Directions: *Be sure to show your work and explain your answer for each question. 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, and any of the recommended textbooks (if appropriate).*

Note on submission: *Please submit your solutions as a single Word file. Word is editable and will facilitate TA grading comments. You do not need to use an equation editor for math text. You may hand-write solutions. Scan all hand-written solutions (be sure they are legible) and embed them in a single Word 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 represents 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 $X_1, X_2, \ldots, X_n$ 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 $Y_1, Y_2, \ldots, Y_n$ be the ordered values of $X_1, X_2, \ldots, X_n$. That is, $Y_1, Y_2, \ldots, Y_n$ are $X_1, X_2, \ldots, X_n$ rearranged in order so that $Y_1 \le Y_2 \le \cdots \le Y_n$. Specifically, $Y_1 = \min(X_1, X_2, \ldots, X_n), \ldots, Y_n = \max(X_1, X_2, \ldots, X_n)$. Show that $Y_1, Y_2, \ldots, Y_n$ are sufficient statistics for μ, σ.

[Hint: This problem can be solved easily by using either the definition of sufficiency or the Factorization Theorem when thought about in the right way. To use the definition, 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 sequence 3, 1, 2?]*