AUBURN UNIVERSITY

STAT 5600/6600 - Probability and Statistics for Data Science

Final Exam December 7, 2023

Instructions:

1. Write your name and the class you enrolled (i.e., 5600 or 6600) by your name.
2. Highest possible score in this exam is 100. You must always **explain your answers** and **show your work** to receive **full credit** or proper **partial credit**. If necessary, you can use both sides of the sheets, but make sure you have indicated doing so.
3. There are 5 questions. Please make sure you have the file with the 5 questions.
4. You are permitted both sides of two 8.5-11 sheets containing formulas, notes, or anything from the course as deemed helpful. You will turn in your exam. You may keep your note/formula sheet. No other aids of any kind are permitted.
5. Calculators are permitted but must be supplied by you. Cell phones are not permitted and should be turned off and put away before the exam begins.
6. You must always **explain your answers** and **show your work** to receive **full credit** or proper **partial credit**. If necessary, you can use the back of these pages, but make sure you have indicated doing so.
7. Q1(d) and Q4(c) are optional for STAT 5600 students, so they don't need to turn in the solution/answer for it.

Name: ____________________________________

Signature: ____

Question	1	2	3	4	5	Total (out of 100)
Score						

Questions

Q1. (20 points)
An item is produced in 3 different factories, F_1, F_2 and F_3. The proportions produced in the 3 factories, and the proportions defective in each factory, are as follows:

Factory	% produced	% defective
F_1	50	2
F_2	30	3
F_3	20	4

(a) Among the three factories, which one produces the highest proportion of items? How about the highest proportion of defective items?

(b) Suppose I have a very large number of these items which are randomly collected and I know which factory produced each item. If you had to choose an item from this collection and you could ask me from which factory you would want the item, which factory would you choose to minimize the chance of getting a defective item, and why?

(c) Suppose you randomly select an item without knowing which factory it came from. What is the probability that the item is defective?

(d) If you randomly select an item that is **not** defective, what is the probability that it was produced in Factory F_3? (round your answer to 3 decimals)

Q2. (20 points)
(a) A random variable X has density $f_X(x) = cx^2$ and support $[0, 3]$, where c is the normalizing constant. Find the cdf $F_X(t)$.

(b) Suppose a random variable Y has density $f_Y(y) = \frac{1}{4}y^3$ and support $[0, 2]$. If the random variables X in part (a) and Y in this part are independent. Find $E(XY)$.

(c) Suppose the random variables X and Y in parts (a) and (b) above are independent. Compute $P(X < 2, Y > 1)$ and also $P(X < 2|Y > 1)$.

Q3. (20 points)
(a) What was the reason that the early pioneers of statistics defined sample variance in the "divide by $n-1$" way?

(b) Say a random variable X has MGF $M_X(t) = e^{2t}$. Find $Var(X)$. What would this imply for the random variable X. Explain your answer.

(c) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ population and let $\overline{X}_n$ be the sample mean. Using the MGF of normal distribution and properties of MGF, show that the standardized variable $Z = \dfrac{X-\mu}{\sigma}$ is normally distributed as $N(0,1)$. Using this, deduce that $Z = \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ has also the standard normal distribution $N(0,1)$.
Hints: For $X \sim N(\mu, \sigma^2)$, the MGF is $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$ and you may use the fact that $\overline{X}_n \sim N(\mu, \sigma^2/n)$ without deriving it.

Q4. (20 points) Suppose the distribution of some random variable X is modeled as uniform distribution on $(0, c)$ and we have a random sample $x_1, \ldots, x_n$ from this distribution.
(a) Find a closed-form expression for the MM (method of moments) estimator of c.

(b) Show that the MLE (maximum likelihood estimator) of c is $\hat{c} = \max(X_i)$ (i.e., the largest of the X_i values).

(c) Find the MLE of $\mathrm{Var}(X)$ for X from uniform distribution on $(0, c)$.

(d) For $c = 1$ and $n = 100$ and consider $W = X_1 + \ldots + X_{100}$. Find the (approximate) probability $P(W < 50)$.

Q5. (20 points) Recall that if $Z \sim N(0,1)$, $P(Z > 1.96) = .025$ and $P(Z > 1.645) = .05$.
(a) The dean of the College of Administrative Sciences and Economics wanted to know whether the graduates of his college used statistical inference techniques during their first year of employment after graduation. He surveyed 304 graduates and found that 204 used statistical inference in the first year of their employment.
Estimate with 95% confidence the proportion of all business school graduates who use their statistical education within a year of graduation. Do the data given in the question indicate that the proportion is different from 0.80? Explain.

(b) A study of college students nationwide found that the mean hours of sleep students get the night before an exam is 6 hours, with standard deviation 1.9 hours. The distribution is bell-shaped (i.e., you can assume it has a normal distribution). At Auburn University (AU), a random sample of 40 students is taken. The mean hours of sleep these students get the night before an exam is 5.5 hours. The standard deviation is found to be same as the nationwide college student population. What is the 95% CI on the mean sleep hours for AU students? Do AU students sleep less than the nationwide average? Explain.