STAT 5600/6600: Homework 7
(Due: 12 pm (midday) Wednesday, 12/06/2023)

Note: Show all your work for the necessary steps to receive full credit. Also, please write your name and the class you enrolled (i.e., 5600 or 6600) by your name.

Please submit your HW in pdf format as one file on Canvas. Please follow all the instructions provided; non-compliance might cause 10-20% of the total score! For computational problems, return only the relevant parts of the output with comments/annotations. The textbook "Probability and Statistics for Data Science" by Matloff is abbreviated as PS4DS.

Q3 is optional for STAT 5600 students, so they don't need to turn in the solution/answer for it.

Questions

Q1. (a) Say we have a random sampleX_i, $i = 1, 2, \ldots, n$, with population cdf $F_X(t)$ where X is a continuous random variable. Let G denote the empirical cdf. Explain your answer in each part.
(i) Is $G(t)$ a nondecreasing or nonincreasing function?
(ii) Is $G(t)$ continuous or discontinuous? If discontinuous, how many points of discontinuity does $G(t)$ have?
(iii) $G(t)$ is a random variable? Why or why not?

(b) Say we have a random sample X_i, $i = 1, 2, \ldots, n$, with population cdf $F_X(t)$. Let G denote the empirical cdf. Suppose (unknown to the one analyzing the data), X has a uniform distribution on $[40, 60]$. Find $Var[G(48.8)]$.

Q2.
(a) Suppose the number of bugs per 1,000 lines of code has a Poisson distribution with mean 5.2. Find the approximate probability of having more than 106 bugs in 20 sections of code, each 1,000 lines long. Assume the different sections act independently in terms of bugs.

(b) Using the dataset `PimaIndiansDiabetes2` in the `mlbench` package, find an approximate 99% confidence interval for the mean age in the sampled population. Would you say the mean age of the population is different from 33? How about different from 30?

Q3. Consider a certain river, and L, its level (in feet) relative to its average. There is a flood whenever $L > 8$, and it is reported that 2.5% of days have flooding. Assume that the level L is normally distributed; the above information implies that the mean is 0. Suppose the

standard deviation of L, σ, goes up by 10%. How much will the percentage of flooding days increase?

Q4. In order to win an election, Bob wants to know the proportion of population who will vote for him. The estimate is required to be within 0.1 of the true proportion with 95% confidence. Find the minimum sample size n needed for the requirement.
Hint: For any real number $u \in (0, 1)$, we have $u(1 - u) <= 1/4$.

Q5.
(a) Suppose the distribution of some random variable X is modeled as uniform distribution on (r, s) and we have a random sample $x_1, \ldots, x_n$ from this distribution. Find a closed-form expression for the MM estimator of r and s and also the MLE estimators.

(b) Consider the parametric density family ct^{c-1} for t in $(0, 1)$, 0 elsewhere. Find closed-form expressions for the MLE and the MM estimate, based on a random sample $x_1, \ldots, x_n$ from this distribution.

Q6.
(a) The hazard function for a random variable X is defined to be $h_X(t) = f_X(t)/[1 - F_X(t)]$. Write code for this for a gamma distribution with shape parameter r and rate parameter λ.

```
haz.gam <- function(t,r,lambda)
{

}

print(haz.gam(0.5,2,3.0))
```

Also plot the hazard function for $r = 2$ and $\lambda = 3$ for $(0, 50)$.

(b) Repeat part (a) for a normal distribution with mean μ and variance σ^2.

```
haz.norm <- function(t,mu,sigma)
{

}

print(haz.norm(0.5,2,3.0))
```

Also plot the hazard function for $\mu = 2$ and $\sigma = 3$ for $(-50, 50)$.