Homework 5: Due Thursday May 2, 2019

Problem 1. Suppose $X_1, \ldots, X_n$ are i.i.d. with a shifted rate-one exponential distribution. That is, their common pdf is given by

$$f(x) = e^{-(x-\theta)}, \quad \text{for } x \geq \theta, \text{ and 0 elsewhere.}$$

i) Compute the MLE of θ. Call this MLE T_n.

ii) Compute $E[T_n]$. Is T_n unbiased or is it biased?

iii) Find a formula for $T_{n-1,i}$, $i = 1, \ldots, n$, and deduce $\overline{T}_n$ and b_{Jack}.

iv) Compute $E[T_{Jack}]$. Is the bias of T_{Jack} smaller than that of T_n? Is it of the same order in n or is it of a smaller order?

v) Compute the variance of T_n.

vi) So far, the computations were all analytical (i.e. no data was generated to do the computations). Now, generate $n = 10$ i.i.d. random variables with $\theta = 1$. Compute $\widetilde{T}_n$ and V_{Jack}. Compare V_{jack} to your answer in v).

vii) Repeat vi) with $n = 100$ and $n = 10,000$. What do you notice?

Problem 2. Say $X_1, \ldots, X_n$ are i.i.d. random variables that are exponentially distributed with rate 1. We want to figure out how to break $[0, \infty)$ into intervals to use for a chi-square test. We want the intervals to have the same probability and we want to use $m_n = \lfloor 2n^{2/5} \rfloor$ intervals (as per Mann and Wald's recommendation). Say the intervals are of the form $I_k = [a_k, a_{k+1})$, $k = 1, \ldots, m_n - 1$ and $I_{m_n} = [a_{m_n}, \infty)$.

i) Show that $P(X_1 \geq a_k) = (m_n - k + 1)/m_n$ for all $k = 1, \ldots, m_n$.

ii) Use the above to compute a_k for $k = 1, \ldots, m_n$.

iii) Generate $n = 10$ i.i.d. random variables, exponentially distributed with rate 1 then use a chi-square test with the above intervals to test the hypothesis that the data is exponentially distributed with rate 1. Give a p-value.

iv) Generate $n = 10$ i.i.d. random variables, exponentially distributed with rate 2 then use a chi-square test with the above intervals to test the hypothesis that the data is exponentially distributed with rate 1. Give a p-value.

v) Repeat iii) and iv) with $n = 100$ and $n = 10,000$. (You will need to write a code for this, as the number of intervals will be too large to be done by hand.)

Problem 3. Say $X_1, \dots, X_n$ are i.i.d. random variables that are exponentially distributed with rate θ (i.e. mean $1/\theta$).

i) Compute the MLE of θ.

ii) Break $[0, \infty)$ into intervals $I_0 = [0, 1)$, $I_1 = [1, 2), \dots, I_{m-1} = [m-1, m)$, and $I_m = [m, \infty)$. Write a formula for

$$p_k = P(X_1 \in I_k),\ k = 0, \dots, m,$$

in terms of θ. Check that $p_0 + \cdots + p_m = 1$.

iii) Assuming $\theta = 1$, give a formula for the largest m such that $np_m \geq 1$. This is how many intervals we will use this time, even when $\theta \neq 1$.

iv) Say now that all we have are the frequencies with which X's fell into the different intervals, i.e. we only know the random variables

$$N_k = \sum_{i=1}^{n} \mathbf{1}\{X_i \in I_k\}, \quad 0 \leq k \leq m.$$

Compute the MLE of θ (but now in terms of $N_1, \dots, N_m$).

Note: We could use the intervals we computed in Problem 3, but the formula for the MLE would be much more complicated.

Problem 4. Generate $n = 10$ i.i.d. random variables, exponentially distributed with rate 1. This will be the dataset we will work on in this part of the problem.

i) Using the chi-square goodness-of-fit and the intervals from Problem 3.ii, test the hypothesis that the data is exponentially distributed with rate 1 with m as computed from Problem 3.iii. Give a p-value.

ii) Repeat the above but testing for an exponential distribution with an unknown rate. Use the MLE from Problem 3.iv. Give a p-value.

iii) Recompute the statistic you used in ii) but now using the MLE from Problem 3.i. Is this a chi-square distributed random variable? How does it compare to the statistics you computed in i) and ii)?

Note: You can in fact generate the data set a large number of times and each time compute the statistic in iii), then use the Kolmogorov-Smirnov test to check whether or not this is a chi-square. But here the question is purely theoretical, based on what we have seen in class.

iv) Repeat the above steps with $n = 100$ and $n = 10,000$. (Do not forget to recalculate m.) What do you notice?