STA 312f23 Assignment Six[1]

The paper and pencil part of this assignment is not to be handed in. It is practice for Quiz 6 on October 27th. The R part may be handed in as part of the quiz. **Bring hard copy of your printout to the quiz**. Do not write anything on your printout in advance except possibly your name and student number.

The Kapan-Meier estimate of the survival function is based on discrete time. Accordingly, let the survival time T be a discrete random variable with non-zero probability on the points $t_1, t_2, \ldots$. Also, $t_0 = 0$, and $P(T = t_0) = 0$.

1. Let p_j = the probability of surviving past time t_j, given survival to time t_{j-1}. That is, $p_j = P(T > t_j | T > t_{j-1})$. Prove $p_j = \frac{S(t_j)}{S(t_{j-1})}$.

2. Prove $S(t_k) = \prod_{j=1}^{k} p_j$.

3. This question is background for the questions that follow. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p. That is, $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. You have already proved that the MLE of p is $\widehat{p} = \bar{X}$, the sample proportion. You don't have to do it again.

   (a) Write down the expected value and variance of $\widehat{p}$. Only derive them is you don't know the answer.

   (b) The asymptotic variance is just the variance; there is no need to go through the Fisher information in this case. So, what is the asymptotic distribution of $\widehat{p}$?

4. In a random sample of survival times (which can happen only at points $t_1, t_2, \ldots$), let d_j be the number of deaths at time t_j, and let n_j be the number of individuals at risk at time t_j. "At risk" means not dead yet and not censored yet at time t_j. What is a reasonable estimate of p_j? Again, p_j = the probability of surviving past time t_j, given survival to time t_{j-1}. Call the estimate $\widehat{p}_j$.

5. The estimate $\widehat{p}_j$ from the last question is clearly a sample proportion. Thinking of it as arising from a sample of Bernoullis (not quite true, but close),

   (a) What should the asymptotic distribution of $\widehat{p}_j$ be? Just write down the answer.

   (b) What should the asymptotic distribution of $\log \widehat{p}_j$ be? Show your work.

6. Based on Problem 2, the natural estimator of $S(t)$ is $\widehat{S}(t) = \prod_{t_j \le t} \widehat{p}_j$.

   (a) Write $\log \widehat{S}(t)$ as a sum.

   (b) Based on the asymptotic distribution of $\log \widehat{p}_j$, what is the (asymptotic) expected value of $\log \widehat{S}(t)$?

   (c) Based on the asymptotic distribution of $\log \widehat{p}_j$ and assuming the terms are independent (almost true), what is the (asymptotic) variance of $\log \widehat{S}(t)$?

   (d) Based on the idea that the sum of normals is normal, what should the asymptotic distribution of $\log \widehat{S}(t)$ be?

7. Assuming that your answer to the previous question is correct (you can check the lecture slides on the Kaplan-Meier estimate), derive the asymptotic distribution of $\widehat{S}(t)$. Show your work.

8. Based on your answer to the preceding question, give a reasonable standard error for $\widehat{S}(t)$. This should be something you could compute from sample data.

9. Here is a table that is stolen directly from a nice (and easy) book on survival analysis by Hosmer and Lemeshow. The similar table on p. 26 of our text is mixed up and wrong. In the table below, notice that two observations were censored between times $t = 2$ and $t = 4$; that's why there are only 83 individuals at risk at time $t = 4$, instead of 85. Fill in the empty cells.

t_j	n_j	d_j	$\widehat{p}_j$	$\widehat{S}(t_j)$
0	100	0		
2	100	15		
4	83	5		
5	73	10		0.6894

10. In the table above, how many observations were censored between times $t = 4$ and $t = 5$?

11. The file http://www.utstat.toronto.edu/brunner/data/legal/expo.data2.txt contains a data set you used last week. Read the data and use R to compute the Kaplan-Meier estimate of the survival function.

    (a) Give the Kaplan-Meier estimate of the medan, and a 95% confidence interval. The answer is 3 numbers on your printout.

    (b) Based on the output from `summary`, give $\widehat{S}(t)$ for $t = 0.062$. The answer is a number on your printout.

    (c) Give $\widehat{p}_1$, $\widehat{p}_2$, $\widehat{p}_3$ and $\widehat{p}_4$. These are numbers that you calculate from the output of `summary`. Use R as a calculator and display the numbers on your printout.

    (d) Reproduce $\widehat{S}(0.062)$ from the answer to your last question. Again, use R as a calculator and display the number on your printout.

    (e) Again using R as a calculator, reproduce the standard error of $\widehat{S}(0.062)$ and display the number on your printout. You are calculating your answer to Question 8.

    (f) Plot the Kaplan-Meier estimate of the survival function, but don't print it yet. In the next question, you are going to add another curve to the plot.

12. Last week, you did a parametric analysis assuming these data are from an exponential distribution. You may want to re-use some of your code from last week.

    (a) Compute the maximum likelihood estimate and a 95% confidence interval for the median. Compare your answer (3 numbers) to Question 11a.

    (b) Add the MLE of $S(t)$ to your Kaplan-Meier plot, and print it. Bring your printout to the quiz.

Bring your printout for Questions 11 and 12 to the quiz. All the requested numbers and the code that produced them should appear on your printout. Do not write anything on your printout in advance except possibly your name and student number.