James K. Pringle
140.673 Theory
Dr. Constantine Frangakis
Assignment 4
April 10, 2014

Assignment 4

Problem 1

(a) Are the following families of distributions exponential families?

Proof. The Gamma(a, b) is in the exponential family. Its density is

$$\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx} = \frac{b^a}{\Gamma(a)}\exp\left\{-bx + (a-1)\log(x)\right\} = \frac{b^a}{\Gamma(a)}\exp\left\{\begin{bmatrix}a-1 & -b\end{bmatrix}\begin{bmatrix}\log x \\ x\end{bmatrix}\right\}$$

By chapter 4, slide 2, of the course notes, since the Gamma(a, b) can be written as

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = c(\boldsymbol{\theta})h(\mathbf{x})\exp\left\{\sum_{j=1}^{k}\gamma_j(\boldsymbol{\theta})t_j(\mathbf{x})\right\}$$

the Gamma(a, b) is a member of the exponential family of distributions.

The Weibull(c, d) is not an exponential family because there is a term in the exponential with both data and parameters that cannot be separated into a function of the data multiplied by a function of the parameters. Suppose $\theta = (c, d)$. There are no functions $w(\theta)$ and $f(t)$ such that

$$w(\theta)f(t) = -dt^c$$

□

(b) When using only the theorem discussed in class for judging completeness of the minimal sufficient statistics (MSS), can we deduce that the MSS for the parameters based on an i.i.d. sample from the families below is complete?

(i) Gamma$(a, b), a > 0, b > 0$, where the mean is a/b and the variance is a/b^2.

(ii) The Weibull(c, d) distributions, whose density at $t > 0$ is

$$f(t|c, d) = cdt^{c-1}\exp(-dt^c), \text{ for } c > 0, d > 0$$

Proof. For the Gamma(a, b), the natural parameter space is

$$\{(a-1, -b) : a, b > 0\}$$

This contains a rectangle, for example

$$\{(a, b) : 1 \leq a \leq 2, -1 \leq b \leq -2\}$$

According to the theorem, found on Lecture 4, slide 4, the MSS is also complete.

Since the Weibull is as given, with c allowed to vary, it is not in the exponential family. Therefore, the theorem discussed in class cannot be applied to it. □

Problem 2. Cramer Rao bounds with multiple parameters.

(Before you start answering the rst question, read all of the questions and the nal note carefully).

We want to study the length of time T that a special type of cells spends during DNA synthesis. Assume that we can conduct a study where we can observe the actual times T_i of an i.i.d. sample $i = 1, \ldots, n$ from the population of cells. Assume also that in the population of cells, the times (measured in hours) have approximately the Weibull distribution, given in part (ii) above, for some unknown true parameter values c_0, d_0.

(i) Find expressions for the scores for c and d defined, respectively, as:

$$S_{i,c_0} = \frac{\partial \log f(t|c,d)}{\partial c}|_{t=T_i, c=c_0, d=d_0}, \text{ and } S_{i,d_0} = \frac{\partial \log f(t|c,d)}{\partial d}|_{t=T_i, c=c_0, d=d_0}$$

Proof. Note

$$\log f(t|c,d) = \log[cdt^{c-1}\exp(-dt^c)] = \log c + \log d + (c-1)\log t - dt^c$$

Hence

$$S_{i,c_0} = \frac{\partial \log f(t|c,d)}{\partial c}|_{t=T_i, c=c_0, d=d_0} = 1/c_0 + \log T_i - d_0 T_i^{c_0} \log T_i$$

and

$$S_{i,d_0} = \frac{\partial \log f(t|c,d)}{\partial d}|_{t=T_i, c=c_0, d=d_0} = 1/d_0 - T_i^c$$

□

(ii) By noting that $\partial(t^c)/\partial(t) = ct^{c-1}$, show that the cumulative distribution function $\text{pr}(T < t|c,d) = 1 - \exp(-dt^c)$. Hence, or otherwise, describe explicitly how you can simulate a random sample of 100000 lengths of time from the Weibull if you knew c_0, d_0.

Proof. Notice

$$\lim_{t\to\infty} \text{pr}(T < t|c,d) = 1 - 0 = 1$$

and

$$\lim_{t\to 0} \text{pr}(T < t|c,d) = 1 - 1 = 0$$

Furthermore, $\text{pr}(T < t|c,d)$ is continuous in t (hence right-continuous). If $t_1 \leq t_2$, then

$$\begin{aligned} -dt_1^c &\geq -dt_2^c \\ 1 - \exp(-dt_1^c) &\leq 1 - \exp(-dt_2^c) \end{aligned}$$

so that $\text{pr}(T < t|c,d)$ is increasing. Therefore $\text{pr}(T < t|c,d)$ is a distribution function.

Calculating,

$$\frac{d\,\text{pr}(T < t|c,d)}{dt} = \frac{d}{dt}(1 - \exp(-dt^c)) = dct^{c-1}\exp(-dt^c)$$

The derivative of the distribution function is the Weibull density with parameters c and d. Hence, $\text{pr}(T < t|c,d)$ is the distribution function of the Weibull(c,d).

Let $F(t|c,d) = \text{pr}(T < t|c,d)$. The inverse distribution function is

$$F^{-1}(y|c,d) = \left(\frac{-\log(1-y)}{d}\right)^{1/c}$$

since

$$\begin{aligned} F(F^{-1}(y|c,d)) &= 1 - \exp\left(-d\left(\left(\frac{-\log(1-y)}{d}\right)^{1/c}\right)^c\right) \\ &= 1 - \exp(\log(1-y)) \\ &= y \end{aligned}$$

and similarly $F^{-1}(F(t|c,d)) = t$. By a well-known result, if $U \sim Uniform(0,1)$, then

$$\text{pr}(F^{-1}(U|c,d) \leq t) = \text{pr}(T < t|c,d) = F(t|c,d)$$

Hence, to similuate Weibull(c,d), it is sufficient to draw a uniform on $(0,1)$, then to apply F^{-1} to that random draw. This of course can be done 100,000 times. □

(iii) Suppose Dr. Fin, the labs researcher, knows only that $d_0 = .002$, and Dr. Fin wants to use the data T_i, $i = 1,\ldots,n$ to estimate c_0. Find the Cramer-Rao bound (CRb) for Dr. Fins estimation of c_0, if we also know that $c_0 = 3$. (Note: the only unknown in your result should be the sample size n). (Hint: use the relation between scores and the CRb, also, because the analytic (i.e. closed form) calculation of CRb here is difficult, you may want to use a method that is not analytic, if you do, describe the method.)

Proof. Since c is the only parameter, the Fisher information of a single observation is

$$I_1(c_0) = E(S_{i,c_0}^2 | c_0)$$

Since c_0 is a parameter, and since the data are i.i.d. (assuming there are n data points) the CRb for c_0 is

$$(nI_1(c_0))^{-1}$$

Since the expression for S_{i,c_0}^2 is complicated, we approximate its expectation numerically. According to the Law of Large Numbers,

$$\widehat{I_k(c_0)} = \frac{1}{k}\sum_{i=1}^{k} S_{i,c_0}^2 \to E(S_{i,c_0}^2 | c_0) = I_1(c_0)$$

where $\{S_{i,c_0}^2\}_{i\geq 1}$ represents a sequence of realized random variables. Therefore, for n large, we simulate n Weibull(c_0, d_0) where $c_0 = 3$ and $d_0 = 0.002$. □

(iv) Suppose Dr. Fin does not know either d_0 or c_0. Find the CRb for Dr. Fins estimation of c_0, if we know that $c_0 = 3$ and $d_0 = .002$.

Proof. Let $\theta = (c_0, d_0)$. Let $S_{i,\theta}$ be the 2×1 score vector. Since the data are i.i.d., according to the Law of Large Numbers,

$$\widehat{I_k(\theta)} = \frac{1}{k}\sum_{i=1}^{k} (S_{i,\theta})(S_{i,\theta})' \to E((S_{i,\theta})(S_{i,\theta})' | \theta) = I_1(\theta)$$

where as before, $\{S_{i,\theta}\}_{i\geq 1}$ represents a sequence of realized random variables.

Since the data are i.i.d., the CRb in this case is given by the formula

$$\begin{bmatrix}1 & 0\end{bmatrix} (nI_1(\theta))^{-1} \begin{bmatrix}1\\0\end{bmatrix}$$

□

which we approximate by

$$\begin{bmatrix}1 & 0\end{bmatrix} (n\widehat{I_k(\theta)})^{-1} \begin{bmatrix}1\\0\end{bmatrix} =$$

(v) Interpret the CRbs in (iii) and (iv) in terms of estimators, and explain why the larger one is larger (in your explanation, do not use the word bound).

Proof. The CRBs in (iii) and (iv) are numbers for two separate scenarios (in (iii), only c_0 is unknown, and in (iv), both c_0 and d_0 are unknown). The variance of any unbiased estimator (or statistic) of c_0, under the regularity conditions discussed in class, cannot be less than those numbers, in those scenarios. The CRb in (iv) is larger because

the same data are being used to estimate two parameters, and so less than the full information in the data can be used to estimate c_0. In (iii) all the information in the data can be used to predict c_0, since that is the only parameter. □

(vi) Find an expression for the median, m_0, of T in the cell population in terms of arbitrary c_0, d_0. Hence, find the CRb for Dr. Fins estimation of m_0, if only we know that $c_0 = 3$ and $d_0 = .002$. (Note: to answer this part, you may find useful to write a small program for finding numerically the partial derivatives of functions of two variables.)

Proof. The median is the value of t such that

$$F(t|c,d) = 1 - \exp(-dt^c) = 0.5$$

Hence,

$$\begin{aligned} 1 - \exp(-dt^c) &= 0.5 \\ 0.5 &= \exp(-dt^c) \\ \log 0.5 &= -dt^c \\ \log 2 &= dt^c \\ \left(\frac{\log 2}{d}\right)^{1/c} &= t \end{aligned}$$

Therefore, the median m_0 is

$$m_0(c,d) = \left(\frac{\log 2}{d}\right)^{1/c}$$

which is a function only of the parameters of the distribution. Let $\theta = (c,d)'$. Then

$$\frac{dm_0}{d\theta} = \left[-\left(\tfrac{1}{c^2}\right)\log\left(\tfrac{\log 2}{d}\right)\left(\tfrac{\log 2}{d}\right)^{1/c} \quad -\left(\tfrac{1}{c}\right)(\log 2)^{1/c}\left(\tfrac{1}{d}\right)^{1/c+1}\right]'$$

Because the data are i.i.d., the CRb then is

$$\frac{dm_0}{d\theta}'\,(nI_1(\theta))^{-1}\frac{dm_0}{d\theta}$$

which we approximate by

$$\frac{dm_0}{d\theta}'\,(n\widehat{I_k(\theta)})^{-1}\frac{dm_0}{d\theta}$$

□

(vii) How many cells should be studied so that the latter $\sqrt{\text{CRb}}$ for estimating the median be smaller than 0.1 hours? Give one reason for why we may care about the size of the latter $\sqrt{\text{CRb}}$.

Proof. Since $\sqrt{\text{CRb}}$ should be smaller than 0.1, then CRb should be smaller than 0.01. Then

$$0.01 \geq \frac{dm_0}{d\theta}' (nI_1(\theta))^{-1} \frac{dm_0}{d\theta}$$
$$n \geq 100 \frac{dm_0}{d\theta}' (I_1(\theta))^{-1} \frac{dm_0}{d\theta}$$

Therefore, using the simulation data from before, n should be greater than or equal to

$$\lceil 100 \frac{dm_0}{d\theta}' (I_1(\theta))^{-1} \frac{dm_0}{d\theta} \rceil \approx \lceil 100 \frac{dm_0}{d\theta}' (\widehat{I_k(\theta)})^{-1} \frac{dm_0}{d\theta} \rceil$$

□

IN PRACTICE: It is we who are the researcher (i.e., Dr. Fin and we are the same person), and it is also we who do not know either c_0 or d_0 and wish to estimate them, but it is also we who need to have an idea of the CRbs before the study, in order to design it well (for example, to determine approximately how many cells we will need to examine). In that more realistic case, we should get a preliminary idea of the true values of the parameters either from theory specific to the scientific problem, or from earlier studies, or from a smaller (pilot) study that we will conduct before conducting the larger one, or from all of these sources combined. We can then use that preliminary information to elicit a range of possible values of c_0, d_0, and calculate the CRb bounds that will be relevant for designing the larger study. Another approach could be to express the preliminary information explicitly as a prior distribution.