MA 413 - Statistics for Data Science

Solutions to Exercise 5

1. Let $X = X_1, \ldots, X_n$ be i.i.d exponential random variables. Their joint probability density function given the parameter θ is given by :

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

Since we can separate this joint density function as $f_{X|\Theta}(x|\theta) = g(T(x), \theta)h(x)$, by the Fisher Neymann factorization theorem, the statistic is sufficient.

2. First of all, notice that the uniform distribution is not part of the exponential family. Thus we can't simply use the result of the course. Nevertheless, we could still use the Fisher Neyman factorization theorem.

Using indicator functions to express the joint probability density function given parameters $\theta = (\alpha, \beta)$, we get :

$$f_{X|\Theta}(x|\theta) = \frac{1\{\min x_i \geq \alpha\} 1\{\max x_i \leq \beta\}}{(\beta - \alpha)^n}$$

   (a) With $g(T(x), \theta) = 1\{\min x_i \geq \alpha\} 1\{\max x_i \leq \beta\}(\beta - \alpha)^{-n}$ and $h(x) = 1$, we directly conclude that the tuple $(\min x_i, \max x_i)$ is a sufficient statistics for (α, β) using Fisher Neyman factorization theorem.

   (b) In the case where α is known, we get $g(T(x), \theta) = 1\{\max x_i \leq \beta\}(\beta - \alpha)^{-n}$ and $h(x) = 1\{\min x_i \geq \alpha\}$, we directly conclude that the tuple $\max x_i$ is a sufficient statistics for β using Fisher Neyman factorization theorem.

   (c) In the case where β is known, we get $g(T(x), \theta) = 1\{\min x_i \leq \alpha\}(\beta - \alpha)^{-n}$ and $h(x) = 1\{\max x_i \geq \beta\}$, we directly conclude that the tuple $\min x_i$ is a sufficient statistics for α using Fisher Neyman factorization theorem.

3. With $\theta = (\mu, \sigma^2)$, recall that we can write

$$f_{X|\Theta}(x, \theta) = \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n}{2} \log(2\pi\sigma^2) - \frac{n\mu^2}{2\sigma^2} \right\}$$

Consequently, Fisher-Neyman factorization implies that the statistic $S(X) = (\sum X_i, \sum X_i^2)$ is sufficient for the parameter (μ, σ^2). Same for the statistic $T(X) = (n^{-1} \sum X_i, (n-1)^{-1} \sum (X_i - \bar{X})^2)$ since T and S are 1-1 functions of each other.

Using the definition, we can directly show that the statistic $T(X) = (X_1, \ldots, X_n)$ is sufficient for θ since the conditional probability distribution of the sample given the statistic does not depend on θ. Indeed, we get :

$$F_{X|T(X)}(x_1, \ldots, x_n | x'_1, \ldots, x'_n) = \begin{cases} 0 & \text{if} \quad \exists \quad i \quad \text{s.t.} \quad x_i \leq x'_i \\ 1 & \text{otherwise} \end{cases}$$

4. (a) Using a theorem from the course, it holds that $T(X) = \sum X_i$ is sufficient for θ where X_i are exponential random variables. Indeed, for two different samples y and z, we get :

$$\frac{f_{X|\Theta}(y|\theta)}{f_{X|\Theta}(z|\theta)} = \frac{\theta^n e^{-\theta \sum_{i=1}^n y_i}}{\theta^n e^{-\theta \sum_{i=1}^n z_i}} = e^{-\theta(\sum_{i=1}^n y_i - \sum_{i=1}^n z_i)}$$

Thus, this ratio does not depend on θ if and only if we have $T(Y) = T(Z)$.

(b) Using a similar reasoning, for two different samples y and z, we get :

$$\frac{f_{X|\Theta}(y|\theta)}{f_{X|\Theta}(z|\theta)} = \frac{1\{\min y_i \geq \alpha\} 1\{\max y_i \leq \beta\}}{1\{\min z_i \geq \alpha\} 1\{\max z_i \leq \beta\}}$$

Clearly, if $T(Y) = T(Z)$, the ratio does not depend on $\theta = (\alpha, \beta)$. Now suppose that $T(Y) \neq T(Z)$. Thus depending on the interval $[\alpha, \beta]$, this ratio could be 1, 0 or ∞. That is it depends on θ. By proving the contrapositive, we complete the proof : $T(X)$ is minimally sufficient for (α, β).

(c) Using the result from exponential family, we know that the statistics $S(X) = (\sum X_i, \sum X_i^2)$ and $T(X) = (n^{-1} \sum X_i, (n-1)^{-1} \sum (X_i - \bar{X})^2)$ are minimaly sufficient for (μ, σ^2). However, $(X_1, \ldots, X_n)$ is not minimally sufficient. In particular, we cannot find a function $g(\cdot)$ with $(X_1, \ldots, X_n) = g(S(X))$.

5. Let Y be a binomial random variable. We fix n and observe that :

$$\begin{aligned} f_{Y|\Theta}(y|p) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp\left\{ \log\left(\frac{p}{1-p}\right) y + n \log(1-p) + \log\binom{n}{y} \right\} \end{aligned}$$

We get $\phi = \log \frac{p}{1-p}$, $T(y) = y$, $S(y) = \log \binom{n}{y}$, $\gamma(\phi) = n \log(1 + e^{\phi})$.

Remark. Notice that the support of an exponential family should not depend on its parameters. Because of the term $\binom{n}{y}$, the support of $f_{Y|\Theta}$ depends on n and thus, we have to fix n to consider a binomial distribution as an exponential family.

6. Let X be a Gamma random variable. For $x \geq 0$, we have :

$$\begin{aligned} f_{X|\Theta}(x|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \\ &= \exp\{\alpha \log x - \beta x + \alpha \log \beta - \log \Gamma(\alpha) - \log x\} \end{aligned}$$

Using natural parametrization, we get $\Phi = (\alpha, \beta)$, $T(x) = (\log x, -x)$, $S(x) = -\log x$ and $\gamma(\Phi) = -\phi_1 \log \phi_2 + \log \Gamma(\phi_1)$.

Note. In the two following questions, we have to compute the distribution of a rescaled random variable. Let's derive this distribution. Let X be a continuous random variable with c.d.f F_X and p.d.f f_X. We are interested on the distribution of X/n with $n > 0$. We get :

$$F_{X/n}(t) = \mathbb{P}(X/n \leq t) = \mathbb{P}(X \leq nt) = F_X(nt)$$

By definition of the p.d.f of a continuous random variable, we know we have to derive its c.d.f to find it. We obtain :

$$f_{X/n}(t) = F'_{X/n}(t) = n f_X(nt)$$

7. Let $Y = \sum_i X_i$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Using MGF, we get :

$$M_Y(t) = \prod_{i=1}^n M_X(t) = \prod_{i=1}^n \exp\left\{t\mu + \frac{t^2}{2}\sigma^2\right\} = \exp\left\{tn\mu + \frac{t^2}{2}n\sigma^2\right\}$$

We recognize the moment generating function of a normal random variable with parameters $(n\mu, n\sigma^2)$. Now using the above note, we get :

$$f_{\bar{X}}(t) = \frac{n}{\sqrt{2\pi n\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(nt - n\mu)^2}{n\sigma^2}\right\} = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{n(t-\mu)^2}{\sigma^2}\right\}$$

This is the p.d.f of a normal distribution with parameters $(\mu, \sigma^2/n)$

8. Let $Y = \sum_i X_i$ where $X_i \sim \mathcal{E}(\lambda)$. Using MGF, we get :

$$M_Y(t) = \prod_{i=1}^n M_X(t) = \prod_{i=1}^n \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^n$$

We recognize the moment generating function of a Gamma distribution with parameters (n, λ). Now using the above note, we get for $t \geq 0$:

$$f_{\bar{X}}(t) = n\frac{\lambda^n}{\Gamma(n)}(nt)^{n-1}e^{-\lambda nt}$$