Problem 1

Let n items be drawn in order without replacement from a shipment of N items of which $N\theta$ are bad. Let $X_i = 1$ if the i^{th} item drawn is bad, and =0 otherwise.

Show that $\sum_{i=1}^{n} X_i$ is sufficient for θ directly and by the factorization theorem.

<u>Solution:</u>

Since $P(X_1 = x_1, \ldots, X_n = x_n \mid \sum_{i=1}^{n} X_i = t) = \frac{1}{\binom{n}{t}}$, where $X_i = \begin{cases} 1 & if \quad X_i \quad is \quad bad \\ 0 & if \quad X_i \quad is \quad good \end{cases}$;

, which is free of θ, then $T = \sum_{i=1}^{n} X_i$ is a sufficient statistics of θ

On the other hand, $\sum_{i=1}^{n} X_i$ is in a hyper-geometric distribution, so

$$P(\sum_{i=1}^{n} X_i = t \mid \theta)$$

$$= \frac{\binom{N\theta}{t}\binom{N-N\theta}{n-t}}{\binom{N}{n}}$$

Then,

$$P(X_1 = x_1, ..., X_n = x_n)$$
$$= P(X_1 = x_1, ..., X_n = x_n, \sum_{i=1}^{n} X_i = t)$$
$$= P(X_1 = x_1, ..., X_n = x_n \mid \sum_{i=1}^{n} X_i = t) P(\sum_{i=1}^{n} X_i = t)$$
$$= \frac{1}{\binom{n}{t}} \frac{\binom{N\theta}{t}\binom{N-N\theta}{n-t}}{\binom{N}{n}}$$
$$= g(T(\underline{x}), \theta) h(\underline{x})$$

, where $g(T(\underline{x}), \theta) = \frac{1}{\binom{n}{t}} \frac{\binom{N\theta}{t}\binom{N-N\theta}{n-t}}{\binom{N}{n}}$, and $h(\underline{x}) = 1$.

So, by factorization theorem, $T = \sum_{i=1}^{n} X_i$ is a sufficient statistics of θ.

Problem 2

Suppose $X_1, \ldots, X_n$ is a sample from a population with one of the following densities.

a). $p(x,\theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0$. This is beta, $\beta(\theta,1)$, density.

b). $p(x,\theta) = \theta a x^{a-1} \exp(-\theta x^a), x > 1, \theta > 0, a > 0$. This is known as the *Weibull* Density.

c). $p(x,\theta) = \theta a^{\theta} / x^{(\theta+1)}, x > a, \theta > 0, a > 0$. This is known as the *Parto* density.

In all cases, find the real-value sufficient statistic for θ and a fixed.

<u>Solution:</u>

a) Since

$$P(X_1 = x_1, ..., X_n = x_n \mid \theta) = \prod_{i=1}^{n} P(X_i = x_i \mid \theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} I(0 < x_i < 1)$$
$$= \theta^n \prod_{i=1}^{n} x_i^{\theta-1} I(0 < x_i < 1)$$

Let

$$T(\underline{X}) = \prod_{i=1}^{n} X_i$$

, then

$$P\left(X_1 = x_1, \dots, X_n = x_n \mid \theta\right) = g\left(T\left(\underline{x}\right) \mid \theta\right) h(\underline{x})$$

, where

$$g\left(T\left(\underline{x}\right) \mid \theta\right) = \theta^n T\left(\underline{x}\right)^{\theta - 1}$$

, and

$$h(\underline{x}) = \prod_{i=1}^{n} I(0 < x_i < 1)$$

Then by factorization theorem, $T(\underline{X}) = \prod_{i=1}^{n} X_i$ is sufficient for θ.

b) Since

$$\begin{aligned} & P\left(X_1 = x_1, \dots, X_n = x_n \mid \theta\right) \\ & = \prod_{i=1}^{n} \theta a x_i^{a-1} \exp\left(-\theta x_i^a\right) I\left(x_i > 0\right) \\ & = \theta^n a^n \left(\prod_{i=1}^{n} x_i\right)^{a-1} \exp\left(-\theta \sum x_i^a\right) \left(\prod_{i=1}^{n} I\left(x_i > 0\right)\right) \end{aligned}$$

Let

$$T(\underline{X}) = \sum X_i^a$$

, then

$$P\left(X_1 = x_1, \dots, X_n = x_n \mid \theta\right) = g\left(T\left(\underline{x}\right) \mid \theta\right) h(\underline{x})$$

, where

$$g\left(T\left(\underline{x}\right) \mid \theta\right) = \theta^n a^n \exp\left(-\theta T\left(\underline{x}\right)\right)$$

, and

$$h(\underline{x}) = \left(\prod_{i=1}^{n} x_i\right)^{a-1} \prod_{i=1}^{n} I(x_i > 1).$$

Then by factorization theorem, $T(\underline{X}) = \sum X_i^a$ is sufficient for θ.

c).

$$\begin{aligned} & P\left(X_1 = x_1, \dots, X_n = x_n \mid \theta\right) \\ & = \prod_{i=1}^{n} \theta a^{\theta} / x_i^{(\theta+1)} I\left(x_i > a\right) \\ & = \frac{\theta^n a^{n\theta}}{\prod_{i=1}^{n} x_i^{\theta+1}} \prod_{i=1}^{n} I\left(x_i > a\right) \end{aligned}$$

Let

$$T(\underline{X}) = \prod_{i=1}^{n} X_i$$

, then

$$P\left(X_1 = x_1, \dots, X_n = x_n \mid \theta\right) = g\left(T\left(\underline{x}\right) \mid \theta\right) h(\underline{x})$$

, where

$$g\left(T\left(\underline{x}\right)|\,\theta\right)=\frac{\theta^n a^{n\theta}}{T\left(\underline{x}\right)^{\theta+1}}$$

, and

$$h(\underline{x})=\prod_{i=1}^{n} I(x_i>a)\,.$$

Then by factorization theorem, $T(\underline{X})=\prod_{i=1}^{n} X_i$ is sufficient for θ.

Problem 3

Let $X_1, \ldots, X_n$ be a sample from a population with density

$$f_\theta(x) = \begin{cases} a(\theta)h(x) & \text{if } \theta_1 \le x \le \theta_2 \\ 0 & \text{otherwise} \end{cases}$$

, where $h(x) \geq 0, \theta = (\theta_1, \theta_2)$ with $-\infty < \theta_1 \leq \theta_2 < \infty$, and $a(\theta = \left[\int_{\theta_1}^{\theta_2} h(x)dx\right]^{-1}$ is assumed to exist. Find a two-dimensional sufficient statistic for this problem and apply your results to the $u[\theta_1, \theta_2]$ family of distributions.

Solutions:

$$f_\theta(x) = \prod_{i=1}^{n} a(\theta) h(x_i) I(\theta_1 \leq x_i \leq \theta_2)$$
$$= a(\theta)^n \prod_{i=1}^{n} h(x_i) I(\theta_1 \leq x_i \leq \theta_2)$$

Let

$$T(\underline{X}) = (T_1(\underline{X}), \quad T_2(\underline{X})) = \left(\min_{1\leq i\leq n} X_i, \max_{1\leq i\leq n} X_i\right) = (X_{(1)}, X_{(n)})$$

, then

$$f_\theta(x) = g(T(\underline{x}))h(\underline{x})$$

, where

$$g(T(\underline{x})) = a(\theta)^n I(\theta_1 \leq T_1(\underline{x}) \leq \theta_2) I(\theta_1 \leq T_2(\underline{x}) \leq \theta_2)$$

, and

$$h(\underline{x}) = \prod_{i=1}^{n} h(x_i).$$

Then by factorization theorem,

$$T(\underline{X}) = (T_1(\underline{X}), \quad T_2(\underline{X})) = \left(\min_{1\leq i\leq n} X_i, \max_{1\leq i\leq n} X_i\right) = (X_{(1)}, X_{(n)})$$

is sufficient statistics for $\theta = (\theta_1, \theta_2)$.

Especially for $u[\theta_1, \theta_2]$,

$$a(\theta) = \frac{1}{\theta_2 - \theta_1}$$

, and $h(x) = 1$.

So

$$T(\underline{X}) = (T_1(\underline{X}), \quad T_2(\underline{X})) = \left(\min_{1\leq i\leq n} X_i, \max_{1\leq i\leq n} X_i\right) = (X_{(1)}, X_{(n)})$$

is sufficient statistics for $\theta = (\theta_1, \theta_2)$.

Problem 4

Let $\theta = (\theta_1, \theta_2)$ be a bivariate parameter. Suppose that $T_1(X)$ is sufficient for θ_1 whenever θ_2 is fixed and known, whereas $T_2(X)$ is sufficient for θ_2 whenever θ_1 is fixed and known.

Assume that θ_1, θ_2 vary independently, $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$ and that the set $S = \{x : p(x,\theta) > 0\}$ does not depend on θ.

(a) Show that if T_1 and T_2 do not depend on θ_2 and θ_1 respectively, then $(T_1(X), T_2(X))$ is sufficient for θ.

(b) Exhibit an example in which $(T_1(X), T_2(X))$ is sufficient for θ, $T_1(X)$ is sufficient for θ_1 whenever θ_2 is fixed and known, but $T_2(X)$ is not sufficient for θ_2, when θ_1 is fixed and known.

Solution:

(a)

Since $T_1(X)$ is sufficient for θ_1 whenever θ_2 is fixed, there should exist

$$\begin{aligned} &f(X \mid \theta_1, \theta_2) \\ &= g_1(T_1(X), \theta_1, \theta_2) h_0(X, \theta_2) \end{aligned}$$

And also since $T_2(X)$ is sufficient for θ_2 whenever θ_1 is fixed, then

$$h_0(X, \theta_2) = h_1(T_2(X), \theta_2) h(X)$$

And also since $T_1(X)$ does not depend on θ_2, then

$$\begin{aligned} &g_1(T_1(X), \theta_1, \theta_2) \\ &= \tilde{g}(T_1(X), \theta_1) \tilde{h}(\theta_1, \theta_2) \end{aligned}$$

So

$$\begin{aligned} &f(X \mid \theta_1, \theta_2) \\ &= \approx{g}(T_1(X), T_2(X), \theta_1, \theta_2) h(X) \end{aligned}$$

, where

$$\begin{aligned} &\approx{g}(T_1(X), T_2(X), \theta_1, \theta_2) \\ &= \tilde{g}(T_1(X), \theta_1) \tilde{h}(\theta_1, \theta_2) h_1(T_2(X), \theta_2) \end{aligned}$$

Then, by factorization theorem, $T(X) = (T_1(X), T_2(X))$ is sufficient statistics for $\theta = (\theta_1, \theta_2)$.

(b) See bivariate normal distribution

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < \mu < \infty, \sigma > 0.$$

It is not difficult to prove that $T(X) = (T_1(X), T_2(X)) = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is sufficient statistics for (μ, σ^2).

Further, when σ^2 is fixed, $T_1(X) = \sum_{i=1}^{n} X_i$ is sufficient statistics for μ.

But when μ is fixed, $T_2(X) = \sum_{i=1}^{n} X_i^2$ is not sufficient statistics for σ^2.

Problem 5

Let $X_1, X_2, \ldots, X_m$; $Y_1, Y_2, \ldots, Y_n$, be independently distributed according to $\mathcal{N}(\mu, \sigma^2)$ and $N(\eta, \tau^2)$, respectively. Find minimal sufficient statistics for the following three cases:

(i) μ, η, σ, τ are arbitrary: $-\infty < \mu, \eta < \infty, 0 < \sigma, \tau$.

(ii) $\sigma = \tau$ and μ, η, σ are arbitrary.

(iii) $\mu = \eta$ and μ, σ, τ are arbitrary.

<u>Solution:</u>

Since $X_1, \ldots, X_m \sim iid N(\mu, \sigma^2)$, $N(\eta, \tau^2)$, and $X_1, \ldots, X_m$ & $Y_1, \ldots, Y_n$ are all independent, then

$$f(x_1,...,x_m,y_1,...,y_n)=(2\pi)^{-\frac{m+n}{2}}(\sigma^2)^{-\frac{m}{2}}(\tau^2)^{-\frac{n}{2}}\exp\{-\frac{\sum_{i=1}^m(x_i-\mu)^2}{2\sigma^2}-\frac{\sum_{i=1}^n(y_i-\mu)^2}{2\tau^2}\}$$

$$=(2\pi)^{-\frac{m+n}{2}}\exp\{-\frac{m\mu^2}{2\sigma^2}-\frac{n\mu^2}{2\tau^2}\}\cdot\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^m x_i^2+\frac{\mu}{\sigma^2}\sum_{i=1}^m x_i-\frac{1}{2\tau^2}\sum_{i=1}^n y_i^2+\frac{\mu}{\tau^2}\sum_{i=1}^n y_i\}$$

(i)

Let $\tilde{\eta}=(-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2},-\frac{1}{2\tau^2},\frac{\eta}{\tau^2})'$, and $T(\underline{x})=(\sum_{i=1}^m x_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i^2,\sum_{i=1}^n y_i)'$, then by the property of exponential family, $T(\underline{x})=(\sum_{i=1}^m x_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i^2,\sum_{i=1}^n y_i)'$ is the natural sufficient statistics for $\tilde{\eta}=(-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2},-\frac{1}{2\tau^2},\frac{\eta}{\tau^2})'$.

Since $-\infty<\mu,\eta<\infty$ and $\sigma,\tau>0$, then $0<\frac{1}{\sigma^2},\frac{1}{\tau^2}<\infty$, and $-\infty<\frac{\mu}{\sigma^2},\frac{\eta}{\tau^2}<\infty$, and further $-\infty<-\frac{1}{2\sigma^2},-\frac{1}{2\tau^2}<0$. We can see that the natural parameter space Σ of η contains open rectangles, which implies that $T(\underline{x})=(\sum_{i=1}^m x_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i^2,\sum_{i=1}^n y_i)'$ is also minimal sufficient.

(ii)

If $\sigma=\tau$, then $\tilde{\eta}=(-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2},\frac{\eta}{\sigma^2})'$. Let $T(\underline{x})=(\sum_{i=1}^m x_i^2+\sum_{i=1}^n y_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i)'$, then by the property of exponential family, $T(\underline{x})=(\sum_{i=1}^m x_i^2+\sum_{i=1}^n y_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i)'$ is the natural sufficient statistics for $\tilde{\eta}=(-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2},\frac{\eta}{\sigma^2})'$.

Since $-\infty<\mu,\eta<\infty$ and $\sigma>0$, then $0<\frac{1}{\sigma^2}<\infty$, and $-\infty<\frac{\mu}{\sigma^2},\frac{\eta}{\sigma^2}<\infty$, and further $-\infty<-\frac{1}{2\sigma^2}<0$. In this case the natural parameter space Σ of η contains open rectangles, which implies that $T(\underline{x})=(\sum_{i=1}^m x_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i^2,\sum_{i=1}^n y_i)'$ is also minimal sufficient.

(iii)

If $\mu=\eta$, then $\tilde{\eta}=(-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2},-\frac{1}{2\tau^2},\frac{\mu}{\tau^2})'$. Let $T(\underline{x})=(\sum_{i=1}^m x_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i^2,\sum_{i=1}^n y_i)'$, then by the property of exponential family, $T(\underline{x})=(\sum_{i=1}^m x_i^2,\sum_{i=1}^m x_i,\sum_{i=1}^n y_i^2,\sum_{i=1}^n y_i)'$ is the natural sufficient statistics for $\tilde{\eta}=(-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2},-\frac{1}{2\tau^2},\frac{\mu}{\tau^2})'$.

Without loss of generality, for $\underline{x_1}=(x_{11},...,x_{1m},y_{11},...,y_{1n})'$ and $\underline{x_2}=(x_{21},...,x_{2m},y_{21},...,y_{2n})'$, assume $\sum_{i=1}^m x_{1i}^2\geq\sum_{i=1}^m x_{2i}^2$, $\sum_{i=1}^n y_{1i}^2\geq\sum_{i=1}^n y_{2i}^2$, $\sum_{i=1}^m x_{1i}\geq\sum_{i=1}^m x_{2i}$ and $\sum_{i=1}^n y_{1i}\geq\sum_{i=1}^n y_{2i}$.

Now suppose $\exists\ \ k(x_1,x_2)>0$ s.t.

$$f\left(x_1 \mid \tilde{\eta}\right) = f\left(x_2 \mid \tilde{\eta}\right) k\left(x_1, x_2\right)$$

hold for $\forall \quad \tilde{\eta} = \left(\frac{1}{\sigma^2}, \frac{1}{\tau^2}, \frac{\mu}{\sigma^2}, \frac{\mu}{\tau^2}\right)$

, then

$$\frac{f\left(\underline{x}_1 \mid \tilde{\eta}\right)}{f\left(\underline{x}_2 \mid \tilde{\eta}\right)} = \exp\left(-\frac{\sum_{i=1}^{m} x_{1i}^2 - \sum_{i=1}^{m} x_{2i}^2}{2\sigma^2} - \frac{\sum_{i=1}^{n} y_{1i}^2 - \sum_{i=1}^{n} y_{2i}^2}{2\tau^2} + \frac{\mu\left(\sum_{i=1}^{m} x_{1i} - \sum_{i=1}^{m} x_{2i}\right)}{\sigma^2} + \frac{\mu\left(\sum_{i=1}^{n} y_{1i} - \sum_{i=1}^{n} y_{2i}\right)}{\tau^2}\right)$$

Let μ=0, $\sigma^2 = \frac{\sum_{i=1}^{n} y_{1i}^2 - \sum_{i=1}^{n} y_{2i}^2}{\sum_{i=1}^{m} x_{1i}^2 - \sum_{i=1}^{m} x_{2i}^2}$, then

$$\frac{f\left(\underline{x}_1 \mid \tilde{\eta}\right)}{f\left(\underline{x}_2 \mid \tilde{\eta}\right)} = \exp\left(-\left(\sum_{i=1}^{n} y_{1i}^2 - \sum_{i=1}^{n} y_{2i}^2\right)\left(1 + \frac{1}{\tau^2}\right)\right) = k(x, y)$$

holds for $\forall \tau^2 > 0$. This requires $\sum_{i=1}^{n} y_{1i}^2 = \sum_{i=1}^{n} y_{2i}^2$.

Similarly, let μ=0, $\tau^2 = \frac{\sum_{i=1}^{m} x_{1i}^2 - \sum_{i=1}^{m} x_{2i}^2}{\sum_{i=1}^{n} y_{1i}^2 - \sum_{i=1}^{n} y_{2i}^2}$, then

$$\frac{f\left(\underline{x}_1 \mid \tilde{\eta}\right)}{f\left(\underline{x}_2 \mid \tilde{\eta}\right)} = \exp\left(-\left(\sum_{i=1}^{m} x_{1i}^2 - \sum_{i=1}^{m} x_{2i}^2\right)\left(1 + \frac{1}{\sigma^2}\right)\right) = k(x, y)$$

holds for $\forall \sigma^2 > 0$. This requires $\sum_{i=1}^{m} x_{1i}^2 = \sum_{i=1}^{m} x_{2i}^2$.

Further, known that $\sum_{i=1}^{m} x_{1i}^2 = \sum_{i=1}^{m} x_{2i}^2$ and $\sum_{i=1}^{n} y_{1i}^2 = \sum_{i=1}^{n} y_{2i}^2$

, let μ=1, $\sigma^2 = \frac{\sum_{i=1}^{n} y_{1i} - \sum_{i=1}^{n} y_{2i}}{\sum_{i=1}^{m} x_{1i} - \sum_{i=1}^{m} x_{2i}}$, then

$$\frac{f\left(\underline{x}_1 \mid \tilde{\eta}\right)}{f\left(\underline{x}_2 \mid \tilde{\eta}\right)} = \exp\left(-\left(\sum_{i=1}^{n} y_{1i} - \sum_{i=1}^{n} y_{2i}\right)\left(1 + \frac{1}{\tau^2}\right)\right) = k(x, y)$$

holds for $\forall \tau^2 > 0$. This requires $\sum_{i=1}^{n} y_{1i} = \sum_{i=1}^{n} y_{2i}$.

Similarly, let μ=1, $\tau^2 = \frac{\sum_{i=1}^{m} x_{1i} - \sum_{i=1}^{m} x_{2i}}{\sum_{i=1}^{n} y_{1i} - \sum_{i=1}^{n} y_{2i}}$, then

$$\frac{f(\underline{x}_1 \mid \tilde{\eta})}{f(\underline{x}_2 \mid \tilde{\eta})} = \exp\left(-\left(\sum_{i=1}^{m} x_{1i} - \sum_{i=1}^{m} x_{2i}\right)\left(1+\frac{1}{\sigma^2}\right)\right) = k(x,y)$$

holds for $\forall \sigma^2 > 0$. This requires $\sum_{i=1}^{m} x_{1i} = \sum_{i=1}^{m} x_{2i}$.

From above, we see that $T(\underline{x}_1) = T(\underline{x}_2)$, and thus $T = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i)\right)$ is minimal sufficient.