Statistical Inference

B. Statistical Data Science 2nd Year   Indian Statistical Institute

Teacher: Soham Sarkar

Exercise Series 2 (Solutions)

Solution 1. (a)

$$p_\theta(\mathbf{x}) = \begin{cases} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} & \text{if } x_i \in \{0,1\}\,\forall i = 1,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

We can factorize $p_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))h(\mathbf{x})$, with

$$g_\theta(t) = p^t(1-p)^{n-t}, t \in \mathbb{R}, \qquad h(\mathbf{x}) = \begin{cases} 1 & \text{if } x_i \in \{0,1\}\,\forall i = 1,\ldots,n, \\ 0 & \text{otherwise,} \end{cases}$$

where $T(\mathbf{x}) = \sum_{i=1}^n x_i$. As discussed in the class, the factorization is not unique. It also holds, for example, with

$$g_\theta(t) = \begin{cases} p^t(1-p)^{n-t} & \text{if } t = 0,1,\ldots,n, \\ 0 & \text{otherwise,} \end{cases} \qquad h(\mathbf{x}) = \begin{cases} 1 & \text{if } x_i \in \{0,1\}\,\forall i = 1,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

So, a sufficient statistic for p is $T(\mathbf{X}) = \displaystyle\sum_{i=1}^n X_i$.

(b)

$$p_\theta(\mathbf{x}) = \begin{cases} e^{-n\lambda}\dfrac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} & \text{if } x_i \in \{0,1,2,\ldots\}\,\forall i = 1,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

We can factorize $p_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))h(\mathbf{x})$, with

$$g_\theta(t) = e^{-n\lambda}\lambda^t, t \in \mathbb{R}, \qquad h(\mathbf{x}) = \begin{cases} \dfrac{1}{\prod_{i=1}^n x_i!} & \text{if } x_i \in \{0,1,2,\ldots\}\,\forall i = 1,\ldots,n, \\ 0 & \text{otherwise,} \end{cases}$$

where $T(\mathbf{x}) = \sum_{i=1}^n x_i$. We can also use

$$g_\theta(t) = \begin{cases} e^{-n\lambda}\lambda^t & \text{if } t = 0,1,2,\ldots, \\ 0 & \text{otherwise,} \end{cases} \qquad h(\mathbf{x}) = \begin{cases} \dfrac{1}{\prod_{i=1}^n x_i!} & \text{if } x_i \in \{0,1,2,\ldots\}\,\forall i = 1,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

A sufficient statistic for λ is $T(\mathbf{X}) = \displaystyle\sum_{i=1}^n X_i$

(c) Depending on how you define geometric distribution:

$$p_\theta(\mathbf{x}) = p^n(1-p)^{\sum_{i=1}^n x_i - n}, x_i \in \{1, 2, \ldots\}\ \forall i = 1, \ldots, n \qquad \text{or}$$
$$p_\theta(\mathbf{x}) = p^n(1-p)^{\sum_{i=1}^n x_i}, x_i \in \{0, 1, 2, \ldots\}\ \forall i = 1, \ldots, n.$$

For the first one, factorization holds with

$$g_\theta(t) = p^n(1-p)^{t-n}, t \in \mathbb{R}, \qquad h(\mathbf{x}) = \begin{cases} 1 & \text{if } x_i \in \{1, 2, \ldots\}\ \forall i = 1, \ldots, n, \\ 0 & \text{otherwise,} \end{cases}$$

where $T(\mathbf{x}) = \sum_{i=1}^n x_i$. Therefore, a sufficient statistic for p is $T(\mathbf{X}) = \sum\limits_{i=1}^n X_i$.

(d) For $\mathsf{Uniform}(\theta, 1)$, $\theta < 1$:

$$p_\theta(\mathbf{x}) = \begin{cases} \dfrac{1}{(1-\theta)^n} & \text{if } \theta < x_i < 1\ \forall i = 1, \ldots, n, \\ 0 & \text{otherwise,} \end{cases}$$
$$= \begin{cases} \dfrac{1}{(1-\theta)^n} & \text{if } \theta < \min_{i=1,\ldots,n} x_i \text{ and } \max_{i=1,\ldots,n} x_i < 1, \\ 0 & \text{otherwise.} \end{cases}$$

$p_\theta(\cdot)$ can be factorized with

$$g_\theta(t) = \begin{cases} \dfrac{1}{(1-\theta)^n} & \text{if } t > \theta, \\ 0 & \text{otherwise,} \end{cases} \qquad h(\mathbf{x}) = \begin{cases} 1 & \text{if } \max_{i=1,\ldots,n} x_i < 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $T(\mathbf{x}) = \min_{i=1,\ldots,n} x_i$. So, $T(\mathbf{X}) = \min\limits_{i=1,\ldots,n} X_i = X_{(1)}$ is a sufficient statistic for θ.

For $\mathsf{Uniform}(\theta, \theta+1), \theta \in \mathbb{R}$:

$$p_\theta(\mathbf{x}) = \begin{cases} 1 & \text{if } \theta < \min_{i=1,\ldots,n} x_i \text{ and } \max_{i=1,\ldots,n} x_i < \theta + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Factorization holds with

$$g_\theta(\mathbf{t}) = g_\theta(t_1, t_2) = \begin{cases} 1 & \text{if } \theta < t_1 < t_2 < \theta + 1, \\ 0 & \text{otherwise,} \end{cases} \qquad h(\mathbf{x}) = 1,\ \forall \mathbf{x} \in \mathbb{R}^n,$$

where $\mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = \left(\min\limits_{i=1,\ldots,n} x_i, \max\limits_{i=1,\ldots,n} x_i\right)$. So, a (bivariate) sufficient statistic for θ is $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\min_{i=1,\ldots,n} X_i, \max_{i=1,\ldots,n} X_i) = \big(X_{(1)}, X_{(n)}\big)$.

(e) For $\mathsf{Normal}(0, \sigma^2), \theta = \sigma^2$:

$$p_\theta(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2}, \quad x_i \in \mathbb{R},\ \forall i = 1, \ldots, n.$$

Factorization holds with

$$g_\theta(t) = \begin{cases} \dfrac{1}{\sigma^n} e^{-\frac{t}{2\sigma^2}} & \text{if } t > 0, \\ 0 & \text{otherwise,} \end{cases} \qquad h(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}}, \quad \mathbf{x} \in \mathbb{R}^n,$$

where $T(\mathbf{x}) = \sum_{i=1}^n x_i^2$. So, a sufficient statistic for σ^2 (when $\mu = 0$) is $T(\mathbf{X}) = \sum_{i=1}^n X_i^2$.

For $\mathsf{Normal}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$:

$$\begin{aligned} p_\theta(\mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^n x_i^2 + n\mu^2 - 2\mu\sum_{i=1}^n x_i\right\}}, \quad x_i \in \mathbb{R}\,\forall i = 1, \ldots, n. \end{aligned}$$

This shows that $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is (jointly) sufficient for (μ, σ^2). Use, e.g.,

$$g_\theta(t_1, t_2) = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\left\{t_2 + n\mu^2 - 2\mu t_1\right\}}, t_1, t_2 \in \mathbb{R}, \qquad h(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}}, \mathbf{x} \in \mathbb{R}^n.$$

Equivalently, $\left(\bar{X}, S^2\right)$ is also sufficient for (μ, σ^2), where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Note: $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right) \mapsto \left(\bar{X}, S^2\right)$ is a bijection (one-to-one and onto function). So, any function of $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ can be expressed as some function of $\left(\bar{X}, S^2\right)$, and vice-verse.

(f)

$$p_\theta(\mathbf{x}) = \left(\frac{1}{\pi}\right)^n \prod_{i=1}^n \frac{\sigma}{\sigma^2 + (x_i - \mu)^2}, \quad x_i \in \mathbb{R},\ \forall i = 1, \ldots, n.$$

Apart from the entire sample $\mathbf{X}$, the order statistics $\left(X_{(1)}, \ldots, X_{(n)}\right)$ are sufficient for θ, irrespective of whether $\mu = 0$ or $\sigma = 1$. It is not possible to find any further reduced sufficient statistic.

(g)

$$p_\theta(\mathbf{x}) = \frac{1}{(2b)^n} e^{-\frac{1}{b}\sum_{i=1}^n |x_i - a|}, \quad x_i \in \mathbb{R}\,\forall i = 1, \ldots, n.$$

If a is unknown, then no further reduction than the order statistics $\left(X_{(1)}, \ldots, X_{(n)}\right)$ is possible. When $a = 0$, a sufficient statistic for b is $T(\mathbf{X}) = \sum_{i=1}^n |X_i|$.

(h) $\mathsf{Normal}(\theta, \theta^2)$:

$$p_\theta(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\theta^n} e^{-\left\{\frac{1}{2\theta^2}\sum_{i=1}^n x_i^2 + \frac{n}{2} - \frac{1}{\theta}\sum_{i=1}^n x_i\right\}}, \quad x_i \in \mathbb{R},\ \forall i = 1, \ldots, n.$$

$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is jointly sufficient for θ.

$\mathsf{Normal}(\theta, \theta)$:

$$p_\theta(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\theta^{n/2}} e^{-\left\{\frac{1}{2\theta}\sum_{i=1}^n x_i^2 + \frac{n\theta}{2} - \sum_{i=1}^n x_i\right\}}, \quad x_i \in \mathbb{R},\ \forall i = 1, \ldots, n.$$

A sufficient statistic for θ is $\sum_{i=1}^n X_i^2$.

Solution 2. In this example, the joint pdf of X_1, X_2, X_3 is

$$\begin{aligned}
p_\theta(x_1,x_2,x_3) &= \begin{cases} 1 & \text{if } \theta-1 < x_1 < \theta, \theta < x_2 < \theta+1, \theta+1 < x_3 < \theta+2, \\ 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} 1 & \text{if } \theta < x_1+1, x_2, x_3-1 \text{ and } \theta > x_1, x_2-1, x_3-2, \\ 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} 1 & \text{if } \theta < \min\{x_1+1, x_2, x_3-1\} \text{ and } \theta > \max\{x_1, x_2-1, x_3-2\}, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

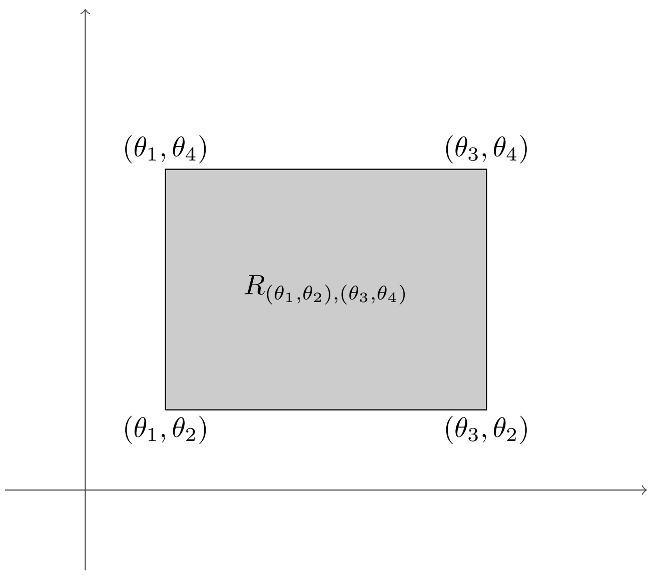
So, a bivariate sufficient statistic for θ is $(\min\{X_1+1, X_2, X_3-1\}, \max\{X_1, X_2-1, X_3-2\})$.

Solution 3. Since X_i's are independent,

$$\begin{aligned}
p_\theta(\mathbf{x}) = \prod_{i=1}^n f_{\theta,i}(x_i) &= \begin{cases} \prod_{i=1}^n \dfrac{1}{2i\theta} & \text{if } -i(\theta-1) < x_i < i(\theta+1),\ \forall i = 1,\dots,n, \\ 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} \dfrac{1}{(2\theta)^n n!} & \text{if } -\theta+1 < \dfrac{x_i}{i} < \theta+1,\ \forall i = 1,\dots,n, \\ 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} \dfrac{1}{(2\theta)^n n!} & \text{if } -\theta < \dfrac{x_i}{i} - 1 < \theta,\ \forall i = 1,\dots,n, \\ 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} \dfrac{1}{(2\theta)^n n!} & \text{if } -\theta < \min_{i=1,\dots,n}\left(\dfrac{x_i}{i} - 1\right), \max_{i=1,\dots,n}\left(\dfrac{x_i}{i} - 1\right) < \theta, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

A bivariate sufficient statistic for θ is $\left(\min\limits_{i=1,\dots,n} \dfrac{X_i}{i}, \max\limits_{i=1,\dots,n} \dfrac{X_i}{i}\right)$.

Solution 4. The four corners of $R_{(\theta_1,\theta_2),(\theta_3,\theta_4)}$, starting from (θ_1,θ_2) and going in the clockwise direction, are (θ_1,θ_2), (θ_1,θ_4), (θ_3,θ_4) and (θ_3,θ_2).



For $f_{\boldsymbol{\theta}}$ to be a pdf, the value of c must be $\dfrac{1}{\text{area}(R_{(\theta_1,\theta_2),(\theta_3,\theta_4)})} = \dfrac{1}{(\theta_3-\theta_1)(\theta_4-\theta_2)}$.

$$p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Y}) = \begin{cases} \dfrac{1}{(\theta_3-\theta_1)^n(\theta_4-\theta_2)^n} & \text{if } (X_i,Y_i) \in R_{(\theta_1,\theta_2),(\theta_3,\theta_4)}\, \forall i = 1,\ldots,n, \\ 0 & \text{otherwise,} \end{cases}$$
$$= \begin{cases} \dfrac{1}{(\theta_3-\theta_1)^n(\theta_4-\theta_2)^n} & \text{if } \theta_1 < X_i < \theta_3 \text{ and } \theta_2 < Y_i < \theta_4\, \forall i = 1,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

$X_{(1)}, Y_{(1)}, X_{(n)}, Y_{(n)}$ are jointly sufficient for $\theta_1, \theta_2, \theta_3, \theta_4$.

Solution 5. We can assume that the blood groups of different persons are independent and identically distributed. In that case, we can use a Multinomial model for this problem.

Define $\mathbf{X} = (X_{\mathtt{A}}, X_{\mathtt{B}}, X_{\mathtt{AB}}, X_{\mathtt{O}})$, where

$$X_{\mathtt{A}} = \begin{cases} 1 & \text{if blood group is } \mathtt{A}, \\ 0 & \text{otherwise,} \end{cases} \qquad X_{\mathtt{B}} = \begin{cases} 1 & \text{if blood group is } \mathtt{B}, \\ 0 & \text{otherwise,} \end{cases}$$
$$X_{\mathtt{AB}} = \begin{cases} 1 & \text{if blood group is } \mathtt{AB}, \\ 0 & \text{otherwise,} \end{cases} \qquad X_{\mathtt{O}} = \begin{cases} 1 & \text{if blood group is } \mathtt{O}, \\ 0 & \text{otherwise.} \end{cases}$$

The random vector $\mathbf{X}$ can take four possible values $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, or $(0,0,0,1)$, corresponding to blood groups A, B, AB, and O, respectively. It is easy to check that

$$\mathbb{P}(X_{\mathtt{A}} = x_{\mathtt{A}}, X_{\mathtt{B}} = x_{\mathtt{B}}, X_{\mathtt{AB}} = x_{\mathtt{AB}}, X_{\mathtt{O}} = x_{\mathtt{O}}) = p_{\mathtt{A}}^{x_{\mathtt{A}}}\, p_{\mathtt{B}}^{x_{\mathtt{B}}}\, p_{\mathtt{AB}}^{x_{\mathtt{AB}}}\, p_{\mathtt{O}}^{x_{\mathtt{O}}},$$
$$x_{\mathtt{A}}, x_{\mathtt{B}}, x_{\mathtt{AB}}, x_{\mathtt{O}} \in \{0,1\}, x_{\mathtt{A}} + x_{\mathtt{B}} + x_{\mathtt{AB}} + x_{\mathtt{O}} = 1.$$

From the collected data, we have observations $\mathbf{x}_1, \ldots, \mathbf{x}_{500}$ for the 500 individuals. Moreover, $\sum_{i=1}^{500} \mathbf{x}_i = (N_{\mathtt{A}}, N_{\mathtt{B}}, N_{\mathtt{AB}}, N_{\mathtt{O}})$. The joint pmf of $\mathbf{X}_1, \ldots, \mathbf{X}_{500}$ is

$$\prod_{i=1}^{500} \mathbb{P}(X_{\mathtt{A},i} = x_{\mathtt{A},i}, X_{\mathtt{B},i} = x_{\mathtt{B},i}, X_{\mathtt{AB},i} = x_{\mathtt{AB},i}, X_{\mathtt{O},i} = x_{\mathtt{O},i}) = \prod_{i=1}^{500} p_{\mathtt{A}}^{x_{\mathtt{A},i}}\, p_{\mathtt{B}}^{x_{\mathtt{B},i}}\, p_{\mathtt{AB}}^{x_{\mathtt{AB},i}}\, p_{\mathtt{O}}^{x_{\mathtt{O},i}} = p_{\mathtt{A}}^{N_{\mathtt{A}}}\, p_{\mathtt{B}}^{N_{\mathtt{B}}}\, p_{\mathtt{AB}}^{N_{\mathtt{AB}}}\, p_{\mathtt{O}}^{N_{\mathtt{O}}}.$$

In the above, we should keep in mind that the pmf makes sense if $x_{\mathtt{A},i}, x_{\mathtt{B},i}, x_{\mathtt{AB},i}, x_{\mathtt{O},i} \in \{0,1\}, x_{\mathtt{A},i} + x_{\mathtt{B},i} + x_{\mathtt{AB},i} + x_{\mathtt{O},i} = 1$ for all $i = 1, \ldots, n$; otherwise the pmf is zero. Use factorization theorem to conclude that $N_{\mathtt{A}}, N_{\mathtt{B}}, N_{\mathtt{AB}}, N_{\mathtt{O}}$ are jointly sufficient for $p_{\mathtt{A}}, p_{\mathtt{B}}, p_{\mathtt{AB}}, p_{\mathtt{O}}$.

Note: Notice that $N_{\mathtt{A}} + N_{\mathtt{B}} + N_{\mathtt{AB}} + N_{\mathtt{O}} = 500$. So, a further reduction can be made by taking any three of these four statistics, since the remaining one can be determined from the three.

With the additional information/modeling assumption in part (c), we have

$$p_{\mathtt{A}} = q_{\mathtt{A}}(1-q_{\mathtt{B}}), \qquad p_{\mathtt{B}} = (1-q_{\mathtt{A}})q_{\mathtt{B}}, \qquad p_{\mathtt{AB}} = q_{\mathtt{A}}q_{\mathtt{B}}, \qquad p_{\mathtt{O}} = (1-q_{\mathtt{A}})(1-q_{\mathtt{B}}).$$

With this, the updated joint pmf becomes

$$\begin{aligned} &\{q_{\mathtt{A}}(1-q_{\mathtt{B}})\}^{N_{\mathtt{A}}}\, \{(1-q_{\mathtt{A}})q_{\mathtt{B}}\}^{N_{\mathtt{B}}}\, \{q_{\mathtt{A}}q_{\mathtt{B}}\}^{N_{\mathtt{AB}}}\, \{(1-q_{\mathtt{A}})(1-q_{\mathtt{B}})\}^{N_{\mathtt{O}}} \\ &= q_{\mathtt{A}}^{N_{\mathtt{A}}+N_{\mathtt{AB}}}\, q_{\mathtt{B}}^{N_{\mathtt{B}}+N_{\mathtt{AB}}}\, (1-q_{\mathtt{A}})^{N_{\mathtt{B}}+N_{\mathtt{O}}}\, (1-q_{\mathtt{B}})^{N_{\mathtt{A}}+N_{\mathtt{O}}} \\ &= q_{\mathtt{A}}^{N_{\mathtt{A}}+N_{\mathtt{AB}}}\, q_{\mathtt{B}}^{N_{\mathtt{B}}+N_{\mathtt{AB}}}\, (1-q_{\mathtt{A}})^{500-N_{\mathtt{A}}-N_{\mathtt{AB}}}\, (1-q_{\mathtt{B}})^{500-N_{\mathtt{B}}-N_{\mathtt{AB}}}. \end{aligned}$$

Therefore, in this case, we get a bivariate sufficient statistic $(N_{\mathtt{A}} + N_{\mathtt{AB}}, N_{\mathtt{B}} + N_{\mathtt{AB}})$.

Note: Notice that $N_{\mathtt{A}} + N_{\mathtt{AB}}$ is the total number of individuals having antigen `A` (regardless of the presence/absence of antigen `B`) and $N_{\mathtt{B}} + N_{\mathtt{AB}}$ is the total number of individuals having antigen `B`. With the additional modeling assumption, we only need to consider these total numbers, which is similar to the Bernoulli case.