Statistical Inference

B. Statistical Data Science 2nd Year   Indian Statistical Institute

Teacher: Soham Sarkar

Exercise Series 2

Exercise 1. Suppose that we have a random sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta$. We write $\mathbf{X}$ (or, $\mathbf{X}_n$ to emphasize on the sample size) to denote the entire sample $(X_1, \ldots, X_n)$. For the following scenarios, find the distribution $p_\theta(\mathbf{x})$ (pdf/pmf) of $\mathbf{X}$. Identify sufficient statistics for the unknown parameter θ.

(a) $f_\theta = \mathsf{Bernoulli}(p)$, $\theta = p$.

(b) $f_\theta = \mathsf{Poisson}(\lambda)$, $\theta = \lambda$.

(c) $f_\theta = \mathsf{Geometric}(p)$, $\theta = p$.

(d) $f_\theta = \mathsf{Uniform}(\theta, 1)$, $\theta < 1$. $f_\theta = \mathsf{Uniform}(\theta, \theta + 1)$, $\theta \in \mathbb{R}$.

(e) $f_\theta = \mathsf{Normal}(0, \sigma^2)$, $\theta = \sigma^2$. $f_\theta = \mathsf{Normal}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$.

(f) $f_\theta = \mathsf{Cauchy}(\theta, 1)$, $\theta \in \mathbb{R}$. $f_\theta = \mathsf{Cauchy}(0, \theta)$, $\theta > 0$. $f_\theta = \mathsf{Cauchy}(\mu, \sigma)$, $\theta = (\mu, \sigma)$.

(g) $f_\theta = \mathsf{Laplace}(\theta, 1)$, $\theta \in \mu$. $f_\theta = \mathsf{Laplace}(0, \theta)$, $\theta > 0$. $f_\theta = \mathsf{Laplace}(a, b)$, $\theta = (a, b)$.

(h) $f_\theta = \mathsf{Normal}(\theta, \theta^2)$. $f_\theta = \mathsf{Normal}(\theta, \theta)$.

Exercise 2. Let $X_1 \sim \mathsf{Uniform}(\theta - 1, \theta)$, $X_2 \sim \mathsf{Uniform}(\theta, \theta + 1)$ and $X_3 \sim \mathsf{Uniform}(\theta + 1, \theta + 2)$ be three independent random variables. Find a bivariate sufficient statistic for θ based on X_1, X_2, X_3.

Exercise 3. Let $X_1, \ldots, X_n$ be independent random variables with X_i having pdf

$$f_{\theta,i}(x) = \begin{cases} \dfrac{1}{2i\theta} & \text{if } -i(\theta - 1) < x < i(\theta + 1), \\ 0 & \text{otherwise.} \end{cases}$$

Write down the joint pdf $p_\theta(\mathbf{x})$ of $X_1, \ldots, X_n$. Find a bivariate sufficient statistic for θ.

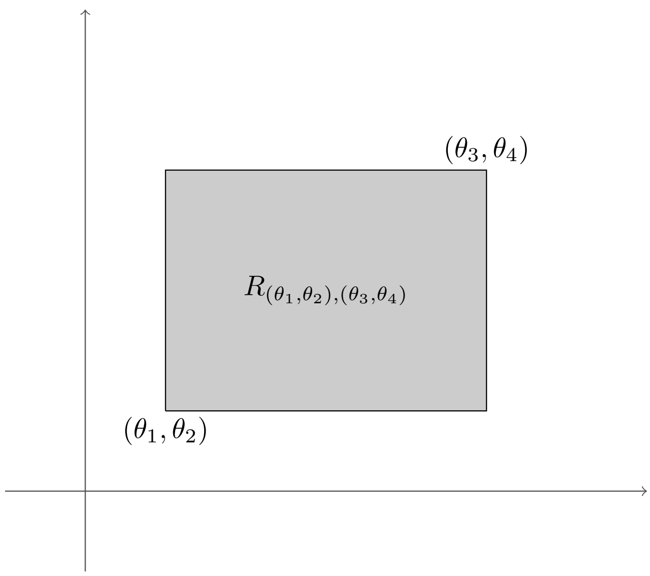
Exercise 4. Let $\theta_1, \theta_2, \theta_3, \theta_4$ be parameters satisfying $\theta_1 < \theta_3$ and $\theta_2 < \theta_4$. Denote by $R_{(\theta_1,\theta_2),(\theta_3,\theta_4)}$ the rectangular region with lower left corner (θ_1, θ_2) and upper right corner (θ_3, θ_4) (see the figure on the next page).

For $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$, let $f_{\boldsymbol{\theta}}(x, y)$ be the pdf of the uniform distribution on $R_{(\theta_1,\theta_2),(\theta_3,\theta_4)}$. That is

$$f_{\boldsymbol{\theta}}(x, y) = \begin{cases} c & \text{if } (x, y) \in R_{(\theta_1,\theta_2),(\theta_3,\theta_4)}, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Determine the value of c.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. bivariate observations with density $f_{\boldsymbol{\theta}}(x, y)$.



(b) Derive the joint pdf of the random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

(c) Find a four-dimensional sufficient statistic for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$.

Exercise 5. The blood groups of human beings are determined by the presence/absence of two different antigens, `A` and `B`. The following chart shows the determination of the four blood groups, `A`, `B`, `AB` and `O`, depending on the antigens.

	Antigen `A` present	Antigen `A` absent
Antigen `B` present	Blood group `AB`	Blood group `B`
Antigen `B` absent	Blood group `A`	Blood group `O`

Let $p_{\mathtt{A}}, p_{\mathtt{B}}, p_{\mathtt{AB}}$ and $p_{\mathtt{O}}$ denote the probabilities of a (randomly selected) person having blood group `A`, `B`, `AB`, or `O`, respectively. Suppose that we select 500 people at random and note down their blood groups.

(a) Formulate a statistical model for this data.

(b) Let $N_{\mathtt{A}}, N_{\mathtt{B}}, N_{\mathtt{AB}}$ and $N_{\mathtt{O}}$ be the number of people (out of 500) having blood groups `A`, `B`, `AB` and `O`, respectively. Show that $N_{\mathtt{A}}, N_{\mathtt{B}}, N_{\mathtt{AB}}, N_{\mathtt{O}}$ are jointly sufficient for $p_{\mathtt{A}}, p_{\mathtt{B}}, p_{\mathtt{AB}}, p_{\mathtt{O}}$.

(c) Let $q_{\mathtt{A}}$ denote the probability of the antigen `A` being present. Similarly, $q_{\mathtt{B}}$ denotes the probability of the antigen `B` being present. Moreover, we believe that the presence/absence of the two antigens do not affect each other. How would you modify your model to take this additional piece of information into account? Also, find sufficient statistics for $q_{\mathtt{A}}$ and $q_{\mathtt{B}}$ in this updated model.