Statistical Inference and Multivariate Analysis (MA324)

Lecture Slides
Lecture 11

Point Estimation: Minimal Sufficiency, Information

Indian Institute of Technology Guwahati

Jan-May 2023

Minimal Sufficient Statistic

- We want to **reduce the data** by considering an appropriate summary statistic. Of course, we should take a summary which has all "**information**" that present in the original data.
- Thus, we want a **shortest summary statistic** that has all "information" regarding the parameter θ. Now, a natural question arises: How to define a **"shortest" sufficient statistic**?

Def: [**Minimal Sufficiency Statistic**] A sufficient statistic T is called minimal sufficient statistic if T is a function of any other sufficient statistic.

- A minimal sufficient statistic T **cannot be reduced any further** to another sufficient statistic. In this sense, minimal sufficient statistic is the shortest and best sufficient statistic.
- The next theorem provides us a way to find minimal sufficient statistic.

Minimal Sufficient Statistic

Theorem

Let $X_1, X_2, \ldots, X_n$ be a RS from a population with PMF/PDF $f(\cdot, \boldsymbol{\theta})$. Consider

$$h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \frac{\prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})}{\prod_{i=1}^{n} f(y_i, \boldsymbol{\theta})} \quad \text{for } \boldsymbol{x} = (x_1, \ldots, x_n), \boldsymbol{y} = (y_1, \ldots, y_n) \in \chi^n.$$

Suppose that there is a statistic $\boldsymbol{T}$ such that for any two points $\boldsymbol{x}, \boldsymbol{y} \in \chi^n$, the expression $h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$ does not involve $\boldsymbol{\theta}$ if and only if $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$. Then $\boldsymbol{T}$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.

Examples:

Example 1: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$. We have seen that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic. Is this minimal sufficient statistic?

Example 2: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Here, $\chi^n = \mathbb{R}^n$ and $\boldsymbol{\theta} = (\mu, \sigma^2)$. Is $\boldsymbol{T} = \left(\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i\right)$ a minimal sufficient statistic?

Example 3: Let us take $n = 3$ in Example 1. We have seen that $T = X_1 + X_2 + X_3$ is minimal sufficient statistic for p. Let us consider the statistic $U = X_1 X_2 + X_3$. Is U sufficient for p?

Information:

- We have mentioned that we would work with sufficient or minimal sufficient statistic, as **they provide reduction of dimension and preserve all “information”** that are present in the RS. However, we have not quantify information.

- Let X be a RV with PMF or PDF $f(\cdot, \theta)$, which depends on a real valued parameter $\theta \in \Theta$. The **variation in the PMF or PDF** $f(x, \theta)$ with respect to $\theta \in \Theta$ for fixed value of x **provides us information** about θ.

Example:

For example, suppose that X has a binomial distribution with PMF

$$f\left(X=x,\,\theta\right)=\binom{n}{x}\theta^{x}\left(1-\theta\right)^{n-x},$$

and let $n=10$ and $x=2$. Then we have $f(x,\,\theta)$ varies with θ as given in the Table below.

- Note that $P(X=2)$ at $\theta=0.8$ is given as 0.000 in the Table. However, that is not exactly zero. These probabilities are rounded off to three decimal places.
- It is this **variation that provides some information** about θ. If the variation is large, then we have more information about θ. On the other hand if the variation is less, we have less information.

Table: Variation in PMF with respect to parameter

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$f(2,\,\theta)$	0.194	0.302	0.233	0.121	0.044	0.011	0.001	0.000

Regularity conditions:

We **measure the change** in a function with respect to a variable **using derivative** of the function with respect to the variable. Following it, here we consider $\frac{\partial}{\partial\theta}\ln f\,(x,\,\theta)$. However, this partial derivative, in general, depend on x. As we are interested to measure the variation (with respect to x) in change, we can consider the **variance of the partial derivative**, $Var\left(\frac{\partial}{\partial\theta}\ln f(X,\,\theta)\right)$. Now, to define information, we need following assumptions, which are called **regularity conditions.**

1. Let $S_\theta = \{x \in \mathbb{R} : f(x,\,\theta) > 0\}$ denote the support of the PMF or PDF $f(\cdot,\,\theta)$ and $S = \cup_{\theta\in\Theta} S_\theta$. Here, we assume that S_θ does not depend on θ, *i.e.*, $S_\theta = S$ for all $\theta \in \Theta$.

2. We also assume that the PDF (or PMF) $f(\cdot,\,\theta)$ is such that differentiation (with respect to θ) and integration (or sum) (with respect to x) are interchangeable.

Regularity conditions:

Now, assume that X is a CRV. Then

$$E_\theta\left[\frac{\partial}{\partial\theta}\ln f(X,\,\theta)\right] = \int_S \frac{\partial \ln f(x,\,\theta)}{\partial\theta} f(x,\,\theta)dx = \int_S \frac{\partial f(x,\,\theta)}{\partial\theta}dx$$
$$= \frac{\partial}{\partial\theta}\int_S f(x,\,\theta)dx = 0,$$

as $\displaystyle\int_S f(x,\,\theta)dx = 1$. Thus,

$$Var\left(\frac{\partial \ln f(X,\,\theta)}{\partial\theta}\right) = E\left[\left(\frac{\partial \ln f(X,\,\theta)}{\partial\theta}\right)^2\right].$$

The DRV case can be handled in a similar manner by replacing the integration by a summation sign. This discussion give us the following quantification of information.

Fisher Information

Def: [**Fisher Information**] The Fisher information (or simply information) about parameter θ contained in X is defined by

$$\mathcal{I}_X(\theta) = E_\theta\left[\left(\frac{\partial \ln f(X,\, \theta)}{\partial\theta}\right)^2\right].$$

Note that $\mathcal{I}_X(\theta) = 0$ if and only if $\frac{\partial}{\partial\theta}\ln f(x,\, \theta) = 0$ with probability one, which means that the PMF or PDF of X does not involve θ. An alternative form of Fisher information can be obtained as follows.

$$\begin{aligned}
&\frac{\partial^2}{\partial\theta^2}\int_S f(x,\, \theta) = 0\\
\Longrightarrow\ &\frac{\partial}{\partial\theta}\int_S \frac{\partial \ln f(x,\, \theta)}{\partial\theta} f(x,\, \theta)dx = 0\\
\Longrightarrow\ &\int_S \frac{\partial^2 \ln f(x,\, \theta)}{\partial\theta^2} f(x,\, \theta)dx + \int_S \left[\frac{\partial \ln f(x,\, \theta)}{\partial\theta}\right]^2 f(x,\, \theta)dx = 0\\
\Longrightarrow\ &\int_S \left[\frac{\partial \ln f(x,\, \theta)}{\partial\theta}\right]^2 f(x,\, \theta)dx = -\int_S \frac{\partial^2 \ln f(x,\, \theta)}{\partial\theta^2} f(x,\, \theta)dx\\
\Longrightarrow\ &\mathcal{I}_X(\theta) = -E_\theta\left(\frac{\partial^2 \ln f(X,\, \theta)}{\partial\theta^2}\right).
\end{aligned}$$

Examples:

Example 4: Let $X \sim Poi(\lambda)$, where $\lambda > 0$. Find $\mathcal{I}_X(\lambda)$.

Example 5: Let $X \sim N(\mu, \sigma^2)$, where σ is known and $\mu \in \mathbb{R}$ is unknown parameters. Find $\mathcal{I}_X(\mu)$.

$\frac{1}{\sigma^2}$

Def: [Fisher Information] The Fisher information contained in a collection of RVs, say $\boldsymbol{X}$, is defined by

$$\mathcal{I}_{\boldsymbol{X}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_{\boldsymbol{X}}(\boldsymbol{X},\,\theta)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_{\boldsymbol{X}}(\boldsymbol{X},\,\theta)\right],$$

where $f_{\boldsymbol{X}}(\cdot,\,\theta)$ is the JPDF of $\boldsymbol{X}$ under θ.

Theorem

Let $X_1, X_2, \ldots, X_n$ be a RS from a population with PMF or PDF $f(\cdot,\,\theta)$, where $\theta \in \Theta$. Let $\mathcal{I}_{\boldsymbol{X}}(\theta)$ denote the Fisher information contained in the RS, then

$$\mathcal{I}_{\boldsymbol{X}}(\theta) = n\mathcal{I}_{X_1}(\theta) \quad \text{for all } \theta \in \Theta.$$

Theorem

Let $\boldsymbol{X}$ be a RS and $\boldsymbol{T}$ be a statistic. Then $\mathcal{I}_{\boldsymbol{X}}(\theta) \geq \mathcal{I}_{\boldsymbol{T}}(\theta)$ for all $\theta \in \Theta$. The equality holds for all $\theta \in \Theta$ if and only if $\boldsymbol{T}$ is a sufficient statistic for θ.

Def: [**Ancillary Statistic**] A statistic $\boldsymbol{T}$ is called an ancillary statistic for $\boldsymbol{\theta}$ if the distribution of $\boldsymbol{T}$ does not involve $\boldsymbol{\theta}$.

Example 6: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, where $\mu \in \mathbb{R}$ is unknown parameter. Then, $T_1 = X_1 - X_2$ is an ancillary statistic for μ as $T_1 \sim N(0, 2)$, which does not involve μ. Similarly, we can check that $T_2 = X_1 + X_2 + \ldots + X_{n-1} - (n-1)X_n$ and S^2 are ancillary statistics for μ.

Let us now consider $\boldsymbol{T} = (T_1, T_2)$. It is easy to check that $\boldsymbol{T} \sim N_2(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & n(n-1) \end{pmatrix}.$$

Thus, the distribution of $\boldsymbol{T}$ does not involve μ, and hence, $\boldsymbol{T}$ is ancillary for μ.