ECON6190 Section 4

Sept. 20th, 2024

Zhuoheng Xu

Sufficient Statistics

Motivation: Reduce the data without losing any information about the parameter of interest.

Example: $X_i \sim N(\mu, \sigma^2)$ where σ^2 is known and μ is the parameter of interest.

Instead of needing the entire data set $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, the sample mean is a sufficient statistic for μ as it captures all the information about μ contained in the data.

Def: A statistic $T(\mathbf{X})$ is sufficient for θ if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on θ.

The definition motivates the following theorem

Thm (Sufficient Statistics)

If $p(\mathbf{X}|\theta)$ is the joint pdf/pmf of $\mathbf{X}$

and $q(t|\theta)$ is the joint pdf/pmf of a statistic $T(\mathbf{X})$

$\frac{p(\mathbf{X}|\theta)}{q(t|\theta)}$ does not depend on θ $\forall \mathbf{X}$ in the sample space

$\Rightarrow$ $T(\mathbf{X})$ is a sufficient statistic for θ

Remark: This is a "Guess and Verify" approach

Step 1: Write down the pdf/pmf $p(\mathbf{X}|\theta)$

Step 2: Guess a statistic $T(\mathbf{X})$ and write down the pdf/pmf $q(t|\theta)$

Step 3: Verify that the ratio $\frac{p(\mathbf{X}|\theta)}{q(t|\theta)}$ does not depend on θ

$\Rightarrow$ Drawback: This method may not be practical as it requires making a guess

A more practical approach: Factorization Theorem

Thm (Factorization Theorem)

Let $f(\mathbf{X}|\theta)$ be the joint pdf/pmf of $\mathbf{X}$

$T(\mathbf{X})$ is a sufficient statistic for θ $\Leftrightarrow$ $f(\mathbf{x}|\theta) = g(T(\mathbf{X})|\theta)h(\mathbf{X})$ $\forall \mathbf{x}$ and $\forall \theta$

Remark: Write down the pdf/pmf $f(\mathbf{x}|\theta)$ and decompose it into two parts:

one part depends on θ while the other part does not depend on θ.

Then we can find the sufficient statistic $T(\mathbf{X})$.

Minimal Sufficient Statistic

Motivation: To find the most informative sufficient statistic

Def: $T^*(\mathbf{X})$ is a minimal sufficient statistic if for any sufficient statistic $T(\mathbf{X})$

$\exists\, r(\cdot)$ s.t. $T^*(\mathbf{X}) = r(T(\mathbf{X}))$

Thm (Minimal Sufficient Statistic)

Let $f(\mathbf{x}|\theta)$ be the joint pdf/pmf of $\mathbf{X}$.

If $\exists\, T(\mathbf{X})$ s.t. for every two sample points $\mathbf{x}$ and $\mathbf{y}$

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} \text{ does not depend on } \theta \Leftrightarrow T(\mathbf{x}) = T(\mathbf{y})$$

Then $T(\mathbf{X})$ is a minimal sufficient statistic.

Remark: Write down $f(\mathbf{x}|\theta)$ and $f(\mathbf{y}|\theta)$ and find $T(\mathbf{x})=T(\mathbf{y})$ so that the ratio $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ does not depend on the parameter θ.

Practice Questions

Question 1

5. Let $\{X_1, \ldots, X_n\}$ be a random sample from the discrete uniform distribution on $\{1, 2 \ldots \theta\}$. That is, the pmf for X_i is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & x = 1, 2 \ldots \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Show that $\max_i X_i$ is a sufficient statistic for θ.

Method: Factorization Theorem

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a random sample

Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a sample point of $\mathbf{X}$

The joint pmf of $\mathbf{X}$ is

$$f(x|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1, 2, \dots, \theta\}, i=1,\dots,n \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \theta^{-n} & \max_i x_i \le \theta, i=1,\dots,n \\ 0 & \text{otherwise} \end{cases}$$

Define $h(x)=1$ and

$$g(t|\theta) = \begin{cases} \theta^{-n} & t \le \theta, i=1,\dots,n \\ 0 & \text{otherwise} \end{cases}$$

By the factorization theorem, $T(X) = \max_i x_i$.

Question 2 (ECON6190 2023 FALL Midterm)

3. **[55 pts]** If X is normal with mean μ and variance σ^2, it has the following pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad \text{for } x \in \mathbb{R}.$$

Let X and Y be jointly normal with the joint pdf

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} - 2\frac{\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right)\right), \quad \text{for } x, y \in \mathbb{R} \tag{3}$$

where $\sigma_X > 0, \sigma_Y > 0$ and $-1 \le \rho \le 1$ are some constants.

(d) Now, suppose I observe a random sample $\{(X_i, Y_i)_{i=1}^n\}$ from the population distribution (3).

i. **[10 pts]** Find a sufficient statistic for the parameters of interest $(\sigma_X^2, \sigma_Y^2, \rho)$. Clearly state your reasoning.

Method: Factorization Theorem

Joint pdf: $f(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$

$$= \prod_{i=1}^{n} f(x_i, y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x_i^2}{\sigma_X^2} - 2\frac{\rho x_i y_i}{\sigma_X \sigma_Y} + \frac{y_i^2}{\sigma_Y^2}\right)\right)$$

$$= \frac{1}{(2\pi)^n \sigma_X^n \sigma_Y^n (1-\rho^2)^{n/2}} \exp\left(\sum_{i=1}^{n} -\frac{1}{2(1-\rho^2)}\left(\frac{x_i^2}{\sigma_X^2} - 2\frac{\rho x_i y_i}{\sigma_X \sigma_Y} + \frac{y_i^2}{\sigma_Y^2}\right)\right)$$

$$= \frac{1}{(2\pi)^n \sigma_X^n \sigma_Y^n (1-\rho^2)^{n/2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{\sum_{i=1}^{n} x_i^2}{\sigma_X^2} - \frac{2\rho\sum_{i=1}^{n} x_i y_i}{\sigma_X \sigma_Y} + \frac{\sum_{i=1}^{n} y_i^2}{\sigma_Y^2}\right)\right)$$

Define $h(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) = 1$. Then a sufficient statistic is $\left(\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i y_i, \sum_{i=1}^{n} y_i^2\right)$

ii. **[8 pts]** Let $\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2$. Find the mean of $\hat{\sigma}_Y^2$ and the finite-sample distribution of $\hat{\sigma}_Y^2$.

Note that $Y \sim N(0, \sigma_Y^2)$

Then $E[Y_i] = 0$ and $Var[Y_i] = E[Y_i^2] - E[Y_i]^2 = E[Y_i^2] = \sigma_Y^2$

$$E[\hat{\sigma}_Y^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n} Y_i^2\right] = \frac{1}{n-1}\sum_{i=1}^{n} E[Y_i^2] = \frac{1}{n-1}\sum_{i=1}^{n}\sigma_Y^2 = \frac{n}{n-1}\sigma_Y^2$$

Note that $\frac{(n-1)\hat{\sigma}_Y^2}{\sigma_Y^2} = \sum_{i=1}^{n}\frac{Y_i^2}{\sigma_Y^2} \sim \chi_n^2$. Thus $\hat{\sigma}_Y^2 \sim \frac{\sigma_Y^2}{n-1}\chi_n^2$

i. **[7 pts]** Let $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Find the mean of s_Y^2 and the finite-sample distribution of s_Y^2.

From the lecture notes, we know that $E[S_Y^2] = \sigma_Y^2$, and $\frac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi_{n-1}^2$.

Thus $S_Y^2 \sim \frac{\sigma_Y^2}{n-1}\chi_{n-1}^2$

Question 3 (ECON6190 2022 FALL Midterm)

5. **[25 pts]** Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$ with an unknown mean μ and **known** variance $\sigma^2 > 0$. We draw a random sample $\mathbf{X} := \{X_1, X_2, \ldots X_n\}$ of size n from X. We are interested in estimating μ based on $\mathbf{X}$.

   (a) Find a minimal sufficient statistic for μ.

Method: Theorem of minimal sufficient statistic

For any two sample points $\mathbf{x}$ and y

$$\frac{f(\mathbf{x}|\mu,\sigma^2)}{f(y|\mu,\sigma^2)} = \frac{\prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}{\prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)}$$

$$= \frac{\exp\left(\sum_{i=1}^{n}-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}{\exp\left(\sum_{i=1}^{n}-\frac{(y_i-\mu)^2}{2\sigma^2}\right)}$$

$$= \frac{\exp\left(\sum_{i=1}^{n}-\frac{(x_i-\bar{x})^2+(\bar{x}-\mu)^2}{2\sigma^2}\right)}{\exp\left(\sum_{i=1}^{n}-\frac{(y_i-\bar{y})^2+(\bar{y}-\mu)^2}{2\sigma^2}\right)}$$

$$= \exp\left(\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2-(x_i-\bar{x})^2}{2\sigma^2} + \frac{n(\bar{y}^2-\bar{x}^2)-2n(\bar{x}-\bar{y})\mu}{2\sigma^2}\right)$$

This ratio does not depend on μ if and only if $\bar{x}=\bar{y}$.

The minimal sufficient statistic is $T(\mathbf{X}) = \bar{\mathbf{X}}$.

Question 4

7. [Hong 6.5] Suppose $\mathbf{X}^n = (X_1, \ldots, X_n)$ is an iid $\mathrm{N}(\mu_1, \sigma_1^2)$ random sample, $\mathbf{Y}^m = (Y_1, \ldots, Y_m)$ is an iid $\mathrm{N}(\mu_2, \sigma_2^2)$ random sample, and the two random samples are mutually independent. Find the distribution of $\bar{X}_n - \bar{Y}_m$ where $\bar{X}_n$ and $\bar{Y}_m$ are the sample means of the first and second random samples, respectively.

We know that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, X_i \sim N(\mu_1, \sigma_1^2)$

$\bar{Y}_m = \frac{1}{m}\sum_{j=1}^{m} Y_j, Y_j \sim N(\mu_2, \sigma_2^2)$

Since X_i are iid and Y_j are iid, $\bar{X}_n \sim N(\mu_1, \frac{\sigma_1^2}{n})$, $\bar{Y}_m \sim N(\mu_2, \frac{\sigma_2^2}{m})$

Since $\bar{X}_n$ and $\bar{Y}_m$ are mutually independent, $\bar{X}_n - \bar{Y}_m \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$.