II Sufficiency

1. Measures of Quality of Estimators

Suppose $f(x,\theta)$, $\theta \in \Theta$ is the "density" of a variable X. Consider a point estimator $Y_n = u(X_1,..,X_n)$ based on a sample $X_1,..,X_n$.

Definition 1

For a given integer n, $Y = u(X_1,..,X_n)$ is called a minimum variance unbiased estimator, (MVUE), of the parameter θ, if Y is unbiased and its variance is smaller than or equal to the variance of every other unbiased estimate of θ.

Example 1

$X_1,..,X_9$ i.i.d. $X_1 \sim N(\theta, \sigma^2)$.

X_1 - unbiased estimator of θ, $Var X_1 = \sigma^2$

$\overline{X}$ - —"— —"—, $Var \overline{X} = \frac{\sigma^2}{n}$

Problem: Is there ~~any~~ ~~another~~ an unbiased estimate of θ with variance smaller than $\frac{\sigma^2}{n}$?

2. Sufficient Statistics

Suppose that $X_1,..,X_n$ are i.i.d. random variables with the density $f(x,\theta)$, $\theta \in \Theta$, ~~and~~ Let $Y_1 = u_1(X_1,..,X_n)$ be a statistic.

Example 1

$X_1, \dots, X_n$ i.i.d.

$$X_i \sim f(x, \theta) = \theta^x (1-\theta)^{1-x}, \; x = 0, 1, \; \theta \in (0,1).$$

Then,

$$Y_1 = \sum_{i=1}^{n} X_i \sim f_{Y_1}(y_1, \theta) = \binom{n}{y_1} \theta^{y_1} (1-\theta)^{n-y_1}, \; y_1 = 0, 1, \dots, n.$$

We will find the conditional probability

$$P(X_1 = x_1, \dots, X_n = x_n \mid Y_1 = y_1) = P(A \mid B),$$

say, where $y_1 = 0, 1, \dots, n$.

(i) If $\sum_{i=1}^{n} x_i \neq y_1$, then $P(A|B) = 0$ because $A \cap B = \emptyset$.

(ii) If $\sum_{i=1}^{n} x_i = y_1$, then $A \subset B$, ~~and~~ Thus $A \cap B = A$ & $P(A|B) = \frac{P(A)}{P(B)} =$

$$\frac{\theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}}{\binom{n}{y_1} \theta^{y_1} (1-\theta)^{n-y_1}} = \frac{\theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}}{\binom{n}{\sum_i x_i} \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}} = \frac{1}{\binom{n}{\sum_i x_i}}$$ and

it does not depend on θ.

In general, let $f_{Y_1}(y_1, \theta)$ be the "density" of the statistic $Y_1 = u_1(X_1, \dots, X_n)$. The conditional probability

$$X_1 = x_1, \dots, X_n = x_n \mid Y_1 = y_1$$

equals

$$\frac{f(x_1, \theta) \cdot \dots \cdot f(x_n, \theta)}{f_{Y_1}(u_1(X_1, \dots, X_n), \theta)}.$$

Definition 1

The statistic Y_1 is called a sufficient statistic for the parameter θ if and only if

$$\frac{f(x_1,\theta)\cdot\ldots\cdot f(x_n,\theta)}{f_{Y_1}\left(u_1(x_1,\ldots,x_n),\theta\right)} = H(x_1,\ldots,x_n),$$

where $H(x_1,\ldots,x_n)$ does not depend upon θ.

Example 2

$X_1,\ldots,X_n$ i.i.d $X_1 \sim \Gamma(2,\theta)$ $\quad f(x,\theta) = \frac{x e^{-\frac{x}{\theta}}}{\Gamma(2)\theta^2} \mathbb{1}_{(0,+\infty)}(x)$

$Y_1 = \sum_{i=1}^n X_i \sim \Gamma(2n,\theta)$ $\quad f_{Y_1}(y_1,\theta) = \frac{1}{\Gamma(2n)\theta^{2n}} y_1^{2n-1} e^{-y_1/\theta} \mathbb{1}_{(0,+\infty)}(y_1)$

We have

$$\frac{\prod_{i=1}^n f(x_i,\theta)}{f_{Y_1}(y_1,\theta)} = \frac{\frac{x_1 e^{-\frac{x_1}{\theta}}}{\Gamma(2)\theta^2}\cdots\frac{x_n e^{-\frac{x_n}{\theta}}}{\Gamma(2)\theta^2}}{\frac{1}{\Gamma(2n)\theta^{2n}} y_1^{2n-1} e^{-y_1/\theta}} = \frac{\frac{1}{\theta^{2n}}\left(\prod_{i=1}^n x_i\right) e^{-\sum_{i=1}^n x_i/\theta}}{\frac{1}{\Gamma(2n)\theta^{2n}}\left(\sum_{i=1}^n x_i\right)^{2n-1} e^{-\sum_{i=1}^n x_i/\theta}} =$$

$$\frac{\Gamma(2n)\left(\prod_{i=1}^n x_i\right)}{\left(\sum_{i=1}^n x_i\right)^{2n-1}}$$ does not depend on θ.

Y_1 - sufficient statistic for θ.

Example 3

$Y_1 \leq \ldots \leq Y_n$ - the order statistics of a random sample of size n from the distribution with pdf

$$f(x,\theta) = e^{-(x-\theta)} \mathbb{1}_{(\theta,+\infty)}(x).$$

The pdf of $Y_1 = \min\{X_1, \dots, X_n\}$ is

$$f_{Y_1}(y_1, \theta) = n e^{-n(y_1 - \theta)} \mathbb{1}_{(\theta, \infty)}(y_1).$$

We have

$$\forall_i \quad \theta < x_i < +\infty$$

$$\frac{\prod_{i=1}^n f(x_i, \theta)}{f_{Y_1}(y_1, \theta)} = \frac{\prod_{i=1}^n e^{-(x_i - \theta)} \mathbb{1}_{(\theta, +\infty)}(x_i)}{n e^{-n(y_1 - \theta)} \mathbb{1}_{(\theta, \infty)}(y_1)} =$$

$$\frac{e^{-\sum_{i=1}^n x_i + n\theta} \mathbb{1}_{(\theta, +\infty)}(x_{(1)}) \mathbb{1}_{(-\infty, +\infty)}(x_{(n)})}{n e^{-n x_{(1)} + n\theta} \mathbb{1}_{(\theta, +\infty)}(x_{(1)})} =$$

$$\frac{e^{-\sum_{i=1}^n x_i}}{n e^{-n x_{(1)}}} \quad \text{does not depend on } \theta.$$

Y_1 - sufficient statistic for θ.

Theorem 1 (factorization theorem)

Let $X_1, \dots, X_n$ denote a random sample from a distribution $f(x, \theta)$, $\theta \in \Theta$. The statistic $Y_1 = u_1(X_1, \dots, X_n)$ is a sufficient statistic for θ if and only if there are two nonnegative functions k_1 and k_2, such that

$$\prod_{i=1}^n f(x_i, \theta) = k_1(u_1(x_1, \dots, x_n), \theta) \cdot k_2(x_1, \dots, x_n),$$

where $k_2(x_1, \dots, x_n)$ does not depend upon θ.

Example 4

$X_1, \dots, X_n$ i.i.d $X_1 \sim N(\theta, \sigma^2)$, σ^2 - known

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We have

$$\sum_{i=1}^{n}(x_i - \theta)^2 = \sum_{i=1}^{n}[(x_i - \bar{x}) + (\bar{x} - \theta)]^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$$

because

$$2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \theta) = 2(\bar{x} - \theta)\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

The joint density of $X_1, \dots, X_n$ has the form

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\sum_{i=1}^{n}(x_i - \theta)^2 / 2\sigma^2\right] = \cancel{\exp}$$

$$\exp\left[-n(\bar{x} - \theta)^2 / 2\sigma^2\right] \left\{ \frac{\exp\left[-\sum_{i=1}^{n}(x_i - \bar{x})^2 / 2\sigma^2\right]}{\left(\sqrt{2\pi}\,\sigma\right)^n} \right\}$$

Thus, $\bar{X}$ is ~~the~~ a sufficient statistic for the parameter θ.

Example 5

$X_1, \dots, X_n$ i.i.d $X_i \sim f(x, \theta) = \theta x^{\theta - 1} \mathbb{1}_{(0,1)}(x)$, $\theta > 0$.

The joint density of $X_1, \dots, X_n$

$$\theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta - 1} = \underbrace{\left[\theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta}\right]}_{k_1(u_1(x_1, \dots, x_n), \theta)} \underbrace{\frac{1}{\prod_{i=1}^{n} x_i}}_{k_2(x_1, \dots, x_n)}$$

$\prod_{i=1}^{n} X_i$ - sufficient statistic for θ.