LECTURE - (15)

Agenda:

(1) Sufficiency
(2) Examples

Until now, we have not studied any rigorous mathematical procedure to come up with statistical estimators. All our estimators have been derived from intuitive ideas. We now will present a method which follows the principle —

" A good statistical estimator combines all the information in the sample about the target parameter".

This method is known as the "method of sufficiency".

Definition: Let $Y_1, Y_2, \ldots, Y_n$ denote a [redacted] sample from a [redacted] population with an unknown parameter θ. Then the estimator $W = g(Y_1, Y_2, \ldots, Y_n)$ is defined to be **sufficient** for θ if, the conditional distribution of $Y_1, Y_2, \ldots, Y_n$ given U, does not depend on θ.

The idea is that if the conditional distribution of the sample given U, does not depend on θ, then U in this sense contains all the information in the sample about θ.

Example: Suppose $Y_1, Y_2, \ldots, Y_n$ are the outcomes of n independent tosses of a coin with $P(\text{head}) = p$, i.e.,

$$Y_i = \begin{cases} 1 & \text{if } i^{th} \text{ trial is heads,} \\ 0 & \text{if } i^{th} \text{ trial is tails,} \end{cases}$$

and $P(Y_i = 1) = p$ for every $i = 1, 2, \ldots, n$.
Note that our standard estimator for p is the sample mean $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Task: Is $\bar{Y}$ sufficient?

To answer this question, let us find the conditional joint probability mass function of $Y_1, Y_2, \ldots, Y_n$ given $\bar{Y}$, and observe if it depends on p.

Note that

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \bar{Y} = y)$$
$$= \frac{P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n, \bar{Y} = y)}{P(\bar{Y} = y)}$$

Hence,

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \mid \bar{Y} = y)$$

$$= \frac{P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n, \sum_{i=1}^{n} Y_i = ny)}{P(\sum_{i=1}^{n} Y_i = ny)}$$

Clearly, $P(Y_1 = y_1, \dots, Y_n = y_n, \sum_{i=1}^{n} Y_i = ny) = 0$

if $ny \neq \sum_{i=1}^{n} y_i$ (How can it be that $Y_1 = y_1, \dots, Y_n = y_n$ and $\sum_{i=1}^{n} Y_i \neq \sum_{i=1}^{n} y_i$). Otherwise, if $ny = \sum_{i=1}^{n} y_i$,

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n, \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} y_i)$$

$$= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$$

$$= \prod_{i=1}^{n} P(Y_i = y_i) \qquad (\because \text{By independence})$$

$$= \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}$$

$$= p^{\sum_{i=1}^{n} y_i} (1-p)^{n - \sum_{i=1}^{n} y_i}$$

On the other hand, $\sum_{i=1}^{n} Y_i$ is a Binomial random variable with parameters n and p. Hence,

$$P(\sum_{i=1}^{n} Y_i = ny) = \binom{n}{ny} p^{ny} (1-p)^{n-ny}.$$

Since $n\bar{y} = \sum_{i=1}^{n} y_i$, it follows that,

$$P\left(\sum_{i=1}^{n} Y_i = n\bar{y}\right) = \binom{n}{\sum_{i=1}^{n} y_i} p^{\sum_{i=1}^{n} y_i} (1-p)^{n-\sum_{i=1}^{n} y_i}$$

Hence,

$$P\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \,\middle|\, \sum_{i=1}^{n} Y_i = n\bar{y}\right)$$

$$= \begin{cases} 0 & \text{if } n\bar{y} \neq \sum_{i=1}^{n} y_i, \\ \dfrac{p^{\sum_{i=1}^{n} y_i} (1-p)^{\sum_{i=1}^{n} y_i}}{\binom{n}{\sum_{i=1}^{n} y_i} p^{\sum_{i=1}^{n} y_i} (1-p)^{n-\sum_{i=1}^{n} y_i}} & \text{if } n\bar{y} = \sum_{i=1}^{n} y_i. \end{cases}$$

$$= \begin{cases} 0 & \text{if } n\bar{y} \neq \sum_{i=1}^{n} y_i, \\ \dfrac{1}{\binom{n}{\sum_{i=1}^{n} y_i}} & \text{if } n\bar{y} = \sum_{i=1}^{n} y_i. \end{cases}$$

Hence, $\bar{Y}$ is a sufficient estimator for p.

Sufficient estimators are also often known as sufficient statistics.