Sufficiency :- الكفاية

The concept of sufficiency of statistics implies finding out a function of the sample points $x_1, x_2, \cdots, x_n$ that summarizes all the information about the parameter θ in the sample data.

* **Def:**

Let $\{X_i\}$, $i = 1, 2, \cdots, n$ denote a random sample from a population with p.d.f. $f(x;\theta)$, where $\theta \in \Omega$. A statistic $T(x)$ is defined to be sufficient statistics for the unknown parameter θ iff the conditional distribution of $x_1, x_2, \cdots, x_n$ given the value of T does not involve θ for all other value t of T.

That is:

$$P(X_1 = x_1, \cdots, X_n = x_n \mid T = t) = \frac{P(X_1 = x_1, \cdots, X_n = x_n, T = t)}{P(T = t)}$$

does not involve θ.

* **Example:**

Let $x_1, x_2, \cdots, x_n$ be a r.s. from a distribution with p.d.f.

$$f(x;\theta) = \theta^x (1-\theta)^{1-x} \; I_{\{0,1\}}(x)$$

Note: $I_{\{0,1\}}(x)$ يعني $x = 0, 1$

Show that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for the parameter θ.

Solu:

To Show that, we need to know if the following equation does depend on θ or not?

$$P(X_1, X_2, \dots, X_n \mid T=t) = \frac{P(X_1, X_2, \dots, X_n, T=t)}{P(T=t)}$$

where

$$P(X_1, X_2, \dots, X_n, T=t) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta^{x}(1-\theta)^{1-x}$$

$$= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

$$= \theta^{t}(1-\theta)^{n-t}$$

since $X_i \sim Ber(\theta)$ or $X_i \sim bin(1,\theta)$

Then $T=\sum_{i=1}^{n} X_i \sim bin(n,\theta)$ يمكن اثباتها باستخدام (الدالة المولدة للعزوم M.g.f)

Therefore, $P(T=t) = \begin{cases} C_t^n \, \theta^t (1-\theta)^{n-t} & , t=0,1,2,\dots,n \\ 0 & \text{o.w.} \end{cases}$

Hence

$$P(X_1, X_2, \dots, X_n \mid T=t) = \frac{\theta^t (1-\theta)^{n-t}}{C_t^n \, \theta^t (1-\theta)^{n-t}}$$

$$= \frac{1}{C_t^n}$$ which does not contain θ.

Then $T=\sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

*** Example:**

Let $x_1, x_2, \ldots, x_n$ be a r.s. from a distribution with p.d.f. $f(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}^{(x)}$. Show that $\sum_{i=1}^{n} x_i$ is a sufficient statistic for θ.

Note: $I_{(0,\infty)}^{(x)}$ يعني $x>0$

*** Solu:**

The joint p.d.f of $(x_1, x_2, \ldots, x_n, t)$ is:

$$f(x_1, x_2, \ldots, x_n, t) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$= \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

since X is negative exponential distribution.

Then $T = \sum_{i=1}^{n} x_i$ is Gamma distributed (prove?)

$$T = \sum_{i=1}^{n} x_i \sim G(n, \theta)$$

(يمكن اثباتها باستخدام M.g.f لايجاد التوزيع الاحتمالي لـ $\sum_{i=1}^{n} x_i$ اذا علم التوزيع الاحتمالي لـ X)

Therefore:

$$P(x_1, x_2, \ldots, x_n \mid T = t) = \frac{P(x_1, x_2, \ldots, x_n, T = t)}{P(T = t)}$$

$$= \frac{\theta^n e^{-\theta \sum_{i=1}^{n} x_i}}{\frac{\theta^n}{\Gamma n} \left(\sum_{i=1}^{n} x_i\right)^{n-1} e^{-\theta \sum_{i=1}^{n} x_i}}$$

$$= \frac{\theta^n e^{-\theta t}}{\frac{\theta^n}{\Gamma n} t^{n-1} e^{-\theta t}} = \frac{\Gamma n}{t^{n-1}}$$

$\boxed{t = \sum_{i=1}^{n} x_i}$

Which does not contain θ,
Hence $T = \sum_{i=1}^{n} x_i$ is a sufficient statistic for θ.

* **Remark:**

If $\theta = \{\theta_1, \ldots, \theta_k\}$ consists of more than one parameter, then $T(x) = \{T_1(x), \ldots, T_k(x)\}$ will be more than one statistic too. Therefore, the following definition will be in order.

* **Def:**

Let $x_1, x_2, \ldots, x_n$ be a r.s. from a population with p.d.f. $f(x;\theta)$ where θ may be a vector. The statistics $T_1(x), \ldots, T_k(x)$ are defined to be jointly sufficient if and only if the conditional distribution of $x_1, x_2, \ldots, x_n$ given $T_1 = t, \ldots, T_k = t_k$ does not depend on θ, for all values of $t_1, t_2, \ldots, t_k$.

* **Example:**

Show that $\bar{X}$ is a sufficient statistics for mean of $N(\theta, \sigma^2)$ (where σ^2 = known) using Conditional probability.

* **Soln:**

$$P(x_1, x_2, \ldots, x_n \mid t(x) = \bar{x}) = \frac{P(x_1, x_2, \ldots, x_n, \bar{x})}{P(\bar{x})}$$

$$\because \; x \sim N(\theta, \sigma^2) \implies \bar{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

$$\implies p(\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}\,\sigma} e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2}$$

$$P(X_1, X_2, \ldots, X_n, \bar{X}) = \prod_{i=1}^{n} f(x_i; \theta, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2}$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\left[\sum(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2\right]}$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2} \cdot e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2}$$

$$P(x_1, x_2, \ldots, x_n \mid t(x) = \bar{x}) = \frac{P(x_1, x_2, \ldots, x_n, \bar{x})}{P(\bar{x})}$$

$$= \frac{\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2} \cdot e^{-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2}}{\frac{\sqrt{n}}{\sqrt{2\pi}\,\sigma} e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2}}$$

$$= \frac{\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n}{\sqrt{\frac{n}{2\pi}}\,\sigma} e^{-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2}$$

does not contain θ

$\Rightarrow$ $\bar{X}$ is a sufficient statistic for the parameter θ.