Sample: $X_1, \cdots, X_n \overset{iid}{\sim} X$ with pdf $f(x|\theta)$

$\underset{\sim}{X} = (X_1, \cdots, X_n)^T$ n-dimensional random vector.

$\underset{\sim}{x} = (x_1, \cdots, x_n)^T$ n-dimensional real-valued vector.

Ex. $n=2$. $X_1, X_2 \sim N(\mu, \sigma^2)$

$$\underset{\sim}{X} = (X_1, X_2)^T$$

observed $(10.5, 10.9)$ $\underset{\sim}{x} = (10.5, 10.9)^T$

Statistics: $T(\underset{\sim}{X})$: a form of data reduction or data summary

Sample Space $\longrightarrow$ Image of T

$$\mathcal{X} \longrightarrow \mathcal{T} = \{t : t = T(\underset{\sim}{x}) \text{ for some } \underset{\sim}{x} \in \mathcal{X}\}$$

Example. $n=2$. $N(\mu, \sigma^2)$

$\mathcal{X} = \mathbb{R}^2$

$T(\underset{\sim}{X}) = \frac{X_1 + X_2}{2}$ $\mathbb{R}^2 \to \mathbb{R}$

Image of T : $\mathbb{R}$

Goal: find $T(\underset{\sim}{X})$ for θ

Sufficiency Principal.. If $T(\underset{\sim}{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\underset{\sim}{X}$ only through the value $T(\underset{\sim}{X})$

if we have observed two samples $\underset{\sim}{x}$ and $\underset{\sim}{y}$ such that $T(\underset{\sim}{x}) = T(\underset{\sim}{y})$, then the inference about θ should be the same whether $\underset{\sim}{x}$ or $\underset{\sim}{y}$ is observed!

Eg: $N(\theta, 1)$ point estimator: $\overline{X}_n$

confidence interval $\overline{X}_n \pm z_{\alpha/2} \frac{1}{\sqrt{n}}$

Definition of Sufficient Statistics

A statistic $T(\underset{\sim}{X})$ is a sufficient statistic for θ if the conditional distribution of the sample $\underset{\sim}{X}$ given the value of $T(\underset{\sim}{X})$ does not depend on θ.

$\underset{\sim}{X} = (X_1, \cdots X_n)^T$ $\prod_{i=1}^{n} f(x_i|\theta)$

How to find a sufficient statistic for θ?

Theorem 6.2.2 if $p(\underset{\sim}{x}|\theta)$ is the joint pdf or pmf of $\underset{\sim}{X}$

and $q(t|\theta)$ is the pdf or pmf of $T(\underset{\sim}{X})$

then $T(\underset{\sim}{X})$ is a sufficient statistic for θ

if the ratio

$$\frac{p(\underset{\sim}{x}|\theta)}{q(t(\underset{\sim}{x})|\theta)} \text{ is free of } \theta.$$

Example: (Binomial Sufficient Statistic)

Let $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli (θ) $\quad 0<\theta<1$

$$T(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \theta)$$

$$p(\underset{\sim}{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^{\Sigma x_i}(1-\theta)^{\Sigma(1-x_i)} = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

$$q(t(x)|\theta) = \binom{n}{t(x)} \theta^{t(x)}(1-\theta)^{n-t(x)} \qquad t(x) = \Sigma x_i$$

$$= \binom{n}{\Sigma x_i} \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

$$\mathcal{X} = \{ (x_1, \cdots, x_n) : x_i = 1 \text{ or } 0 \}$$

$$\frac{p(\underset{\sim}{x}|\theta)}{q(t(\underset{\sim}{x})|\theta)} = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{\sum x_i}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}} \quad \text{for all } \underset{\sim}{x} \in \mathcal{X}$$

$$= \frac{1}{\binom{n}{\sum x_i}} \quad \text{free of } \theta$$

Thus $T(\underset{\sim}{X}) = \sum X_i$ is a sufficient statistic for θ

Example: $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ where σ^2 is known

what is a sufficient statistic for θ?

Guess: try $T(\underset{\sim}{X}) = \overline{X_n} \sim N(\theta, \frac{\sigma^2}{n})$

$$p(\underset{\sim}{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\} \quad \underset{\sim}{x} \in \mathbb{R}^n$$

$$q(t(\underset{\sim}{x})|\theta) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} \exp\left\{-\frac{(\frac{1}{n}\sum_{i=1}^{n} x_i - \theta)^2}{2\sigma^2/n}\right\} \quad \underset{\sim}{x} \in \mathbb{R}^n$$

proportional to

$$\frac{p(\underset{\sim}{x}|\theta)}{q(t(\underset{\sim}{x})|\theta)} \propto \frac{\exp\left(-\frac{\sum(x_i-\theta)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(\frac{1}{n}\sum x_i - \theta)^2}{2\sigma^2/n}\right)}$$

$$(x_i-\theta)^2 = (\underbrace{x_i - \tfrac{1}{n}\textstyle\sum x_i}_{a} + \underbrace{\tfrac{1}{n}\textstyle\sum x_i - \theta}_{b})^2$$

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

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

free of θ ! ! !

Ex (Sufficient order statistic)

$$X_1 \cdots X_n \sim f(x|\theta)$$

$$p(\underset{\sim}{x}|\theta) = \cancel{\prod_{i=1}^{n} f(x_i|\theta)}$$

$$T(\underset{\sim}{X}) = (X_{(1)}, \cdots, X_{(n)})^T \quad \text{order statistic}$$

$$q(t(\underset{\sim}{x})|\theta) = n! \prod_{i=1}^{n} f(x_{(i)}|\theta) = n! \cancel{\prod_{i=1}^{n} f(x_i|\theta)}$$

$$\xrightarrow{\text{ratio}} \frac{1}{n!} \quad \text{free of } \theta \text{ !}$$

Theorem 6.2.6 (Factorization Theorem)

Let $p(\underset{\sim}{x}|\theta)$ be the joint pdf/pmf of $\underset{\sim}{X}$

a statistic $T(\underset{\sim}{X})$ is a sufficient statistic for θ

if and only if

there exist function $g(t(\underset{\sim}{x})|\theta)$ and $h(\underset{\sim}{x})$

such that $p(\underset{\sim}{x}|\theta) = g(t(\underset{\sim}{x})|\theta)\, h(\underset{\sim}{x})$ for all $\underset{\sim}{x} \in \mathcal{X}$

Example (Uniform Sufficient statistic)

$$X_1, \dots, X_n \overset{iid}{\sim} Unif(1, \theta)$$

$$p(\underset{\sim}{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\theta - 1} I(1 < x_i < \theta)$$

$$= \left(\frac{1}{\theta-1}\right)^n \prod_{i=1}^{n} I(1 < X_i < \theta)$$

$$= \left(\frac{1}{\theta-1}\right)^n I(1 < X_{(1)} < X_{(n)} < \theta)$$

$$= \left(\frac{1}{\theta-1}\right)^n I(1 < X_{(1)}) \times I(X_{(n)} < \theta)$$

$$= \underbrace{\left(\frac{1}{\theta-1}\right)^n I(X_{(n)} < \theta)}_{g(t(\underset{\sim}{x})|\theta)} \times \underbrace{I(1 < X_{(1)})}_{h(\underset{\sim}{x})}$$

$$t(\underset{\sim}{x}) = X_{(n)}$$

Eg $\quad X_1 \cdots X_n \sim Unif(\theta_1, \theta_2) \quad \underset{\sim}{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$

$$p(\underset{\sim}{x}|\theta) = \underbrace{\left(\frac{1}{\theta_2 - \theta_1}\right)^n I(\theta_1 < X_{(1)} < X_{(n)} < \theta_2)}_{g(t(\underset{\sim}{x})|\theta)} \times \underbrace{1}_{h(\underset{\sim}{x})}$$

$$t(\underset{\sim}{x}) = \begin{pmatrix} X_{(1)} \\ X_{(n)} \end{pmatrix}$$