III Data reduction

Idea: Suppose $\underline{X}$ is a random sample with $X_j \sim f(\cdot|\theta)$ (unknown parameter; certain distribution). e.g. $n(\mu, \sigma^2)$ (θ, θ) only interested in one parameter not in the others

Can we find a statistic $T(\underline{X})$ which keeps all information on θ?

(with the aim of reducing the necessary information)

III.1 Sufficient statistics

Sufficient principle: $T(X)$ is a sufficient statistic for θ if two sample points $\underline{x}$ and $\underline{y}$ with $T(\underline{x}) = T(\underline{y})$ One gets same influence on θ.

Def. $T(\underline{X})$ is a sufficient statistic for θ if the conditional pmf or pdf of the sample $\underline{X}$ given $T(\underline{X})$ does not depend on θ.

Example: Consider $X_i \sim \text{Bernoulli}(\theta)$ with $\theta \in (0,1)$ and $T(\underline{X}) := \sum_{j=1}^{N} X_j$ with N fixed.

e.g. $\underline{x} = (0,1,0,0,0,1,1,0,\cdots,1)$ e.g. $T(\underline{x}) = 25$ (number of 1 in $\underline{x}$)

Observe that $T(\underline{X}) \sim \text{binormial}(N, \theta)$.

Set $t := \sum_{i=1}^{N} x_i$. Then the conditional pmf of $\underline{X}$ given $T(\underline{X})$ is

$$\frac{\prod_{j=1}^{N} \text{Bern}(x_j|\theta)}{\text{binomial}(t|N,\theta)} = \frac{\prod_{j=1}^{N} \theta^{x_j}(1-\theta)^{1-x_j}}{\binom{N}{t}\theta^t(1-\theta)^{N-t}} = \frac{\theta^{\sum_{j=1}^{N} x_j}(1-\theta)^{\sum_{j=1}^{N}(1-x_j)}}{\binom{N}{t}\theta^t(1-\theta)^{N-t}} = \frac{\theta^t(1-\theta)^{N-t}}{\binom{N}{t}\theta^t(1-\theta)^{N-t}}$$

$$= \frac{1}{\binom{N}{t}} \text{ indep of } \theta.$$

$\Rightarrow T$ is a sufficient statistics for θ $\Rightarrow$ any $\underline{x}, \underline{y}$ with $\sum x_j, \sum y_j$ provides the same information on θ.

Thm. (Factorization Thm)

A statistic $T(\underline{X})$ is sufficient for θ if and only if

$$f_{\underline{X}}(\underline{x}|\theta) = g(T(\underline{x})|\theta)\, h(\underline{x}) \quad \forall \underline{x} \in \mathbb{R}^N$$

($h(\underline{x})$: indep of θ)

Example: $X_j \sim (\mu, \sigma^2)$ and consider $\theta = \mu$. σ^2 is known.

$$f_{\underline{X}}(\underline{x}|\mu,\sigma^2) = \prod_{j=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_j - \bar{x} + \bar{x} - \mu)^2}{2\sigma^2}\right) \quad \bar{x} := \frac{1}{N}\sum_{j=1}^{N} x_j$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{\sum_{j=1}^{N}(x_j-\bar{x})^2 + N(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

$$= \underbrace{e^{-N(\bar{x}-\mu)^2/2\sigma^2}}_{g(\bar{x}|\mu)} \underbrace{(2\pi\sigma^2)^{-\frac{N}{2}} e^{-\sum_j (x_j-\bar{x})^2/2\sigma^2}}_{\text{indep of } \mu}$$

$\Rightarrow T(\underline{X}) := \bar{X}$ is a sufficient statistic for μ.

Remarks:

1) Sufficient statistics are θ-dependent and model dependent(X_j).
2) Vector valued $T(\underline{X})$ and vector parameters $\boldsymbol{\theta}$ are possible.
3) We can look at minimal sufficient statistics. (maybe not unique)

(any other sufficient statistics should be a function of a minimal one)