IE605: Engineering Statistics

Lecture 10

Manjesh K. Hanawal

Previous Lecture:

- Data Reduction
- Sufficiency principle
- Sufficient Statistics
- Factorization theorem

This Lecture:

- Minimum statistics
- Ancillary statistics
- Complete statistics

Factorization theorem

How to come up with a sufficient statistic for a parameter

- guess a statistics (required good intuition)
- find its pdf/pmf (expression can be tedious)
- find the ration to acertain

Factorization Theorem: For a random sample X with pdf/pmf $f(x|\theta)$, let $T(X)$ is a statistic for θ. Then $T(X)$ is sufficient statistic **if and only if** if there exists functions $g(t|\theta)$ and $h(x)$ such that, for all x and parameters points θ

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

Example 1:

Population distribution $\sim \mathcal{N}(\mu, \sigma^2)$, with μ unknown and σ^2 known

$$f(x|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\sum_{i=1}^{2}(x_i - \mu)^2/2\sigma^2\right\}$$

$$f(x|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2 - n(\bar{x} - \mu)^2\right)/2\sigma^2\right\}$$

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

$$= h(x)g(\bar{x}|\mu)$$

Hence $T(x) = \bar{x}$ is a sufficient statistics.

Example 2:

Population distribution $\sim \mathcal{N}(\mu, \sigma^2)$, both μ and σ^2 unknown

$$f(x|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\sum_{i=1}^{2}(x_i - \mu)^2/2\sigma^2\right\}$$

$$\begin{aligned}
f(x|\mu) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2 - n(\bar{x} - \mu)^2\right)/2\sigma^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\sum_{i=1}^{n}(x_i - \bar{x})^2/2\sigma^2\right\} \exp\left\{-n(\bar{x} - \mu)^2/2\sigma^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{(n-1)s^2/2\sigma^2\right\} \exp\left\{-n(\bar{x} - \mu)^2/2\sigma^2\right\} \\
&= h(x)g(\bar{x}, s^2|(\mu, \sigma^2))
\end{aligned}$$

Set $T_1(x) = \bar{x}$ and $T_2(x) = s^2$. Then $T(x) = (T_1(x), T_2(x))$ is a sufficient statistics for $\theta = (\mu, \sigma^2)$

Sufficient Statistics for Exponential Family

Recall that $f(x|\theta)$ is belongs to exponential family with parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ if

$$f(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\}.$$

$$T(x) = \left(\sum_{j=1}^{n} t_1(x_j), \ldots, \sum_{j=1}^{n} t_k(x_j)\right)$$

is a sufficient statistics for θ $(d \leq k)$.

Functions of Sufficient Statistics

- Can there be one sufficient statistic or multiple?
- The entire sample is always sufficient statistic. Set $T(x) = (x_1, x_2, \ldots, x_n)$ and $h(x) = 1$, then

$$f(x|\theta) = f(T(x)|\theta)h(x)$$

- Any one-to-one function of a sufficient statistics is also a sufficient statistics.
- Let $T^\star(X) = r(T(X))$ for some invertible r and $T(X)$ is a sufficient statistics

$$\begin{aligned} f(x|\theta) &= g(T(x)|\theta)h(x) \\ &= g(r^{-1}(T^\star(x)|\theta))h(x) \\ &= g^\star(T^\star(x)|\theta)h(x) \end{aligned}$$

Hence $T^\star(x)$ is a sufficient statistics

Minimal Sufficient Statistics

A sufficient Statistics $T(X)$ is called a minimal sufficient statistics, if for any sufficient statistics $T'(X)$,

$$T'(x) = T'(y) \implies T(x) = T(y)$$

- $\mathcal{T} = \{t : t = T(x), x \in \mathcal{X}\}$ and $A_t = \{x : T(x) = t\}$
- $\mathcal{T}' = \{t' : t' = T'(x), x \in \mathcal{X}\}$ and $B_{t'} = \{x : T'(x) = t'\}$
- for any $t' \in \mathcal{T}'$, there exists $t \in \mathcal{T}$ such that $B_{t'} \subset A_t$.

Example: Minimal Sufficient Statistics

Samples are drawn from $\mathcal{N}(\mu, \sigma^2)$ with unknown μ and known σ^2.

- $T_1(x) = \bar{x}$ is a sufficient statistics for μ
- $T_2(x) = (\bar{x}, s^2)$ is also a sufficinet statistics for μ (verify!)
- $T_1(x) = \bar{x} = r(\bar{x}, s^2) = r(T_2(x))$
- As both T_1 and T_2 are sufficient statistic they contain same knowledge about μ
- Additional knowledge of s^2 does not add any information about μ.
- T_1 gives better data reductions!

Test for Minimal Sufficient Statistics

> Let $f(x|\theta)$ be the pmf/pdf of a sample. Suppose there exists a function $T(X)$ such that, for every sample pair (x, y), the ratio $f(x|\theta)/f(y|\theta)$ is a constant iff $T(x) = T(y)$. Then $T(X)$ is a minimal sufficient statistics.

- Image of $\mathcal{X}$ under T:
- Define partion set $\mathcal{X}$
- Select one element in each partition
- Pair each $x \in \mathcal{X}$ with partitions
- Argue T is sufficient statistics using factorization theorem

Test for Minimal Sufficient Statistic contd..

- T' is another sufficient statistics
- Apply factorization theorem

- Consider points (x, y) such that $T'(x) = T'(y)$. Apply ratio

- Apply the assumption to T is minimal.

Example: Normal Minimal Statistics

Samples are drawn from $\mathcal{N}(\mu, \sigma^2)$. (μ, σ^2) unknown.

- Consider two sample points (x, y) with statistics $(\bar{x}, s_x^2)$ and $(\bar{x}, s_x^2)$

$$\frac{f(x|\theta)}{f(y|\theta)} = \frac{(2\pi\sigma^2)^{n/2}}{(2\pi\sigma^2)^{n/2}} \frac{\exp\left\{-\left(n(\bar{x}-\mu)^2 + (n-1)s_x^2\right)/2\sigma^2\right\}}{\exp\left\{-\left(n(\bar{y}-\mu)^2 + (n-1)s_y^2\right)/2\sigma^2\right\}}$$

- The ratio is constant iff $\bar{x} = \bar{y}$ and $s_x^2 = s_y^2$, i.e., statistics are same

Example: Uniform Minimal Statistics

Samples are drawn from $Unif(\theta, \theta+1)$, θ is unknown.
PDF of sample x is

$$f(x|\theta) = \begin{cases} 1 & \text{if} \quad \theta < x_i < \theta + 1 \quad i = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

- It can be reorganized as

$$f(x|\theta) = \begin{cases} 1 & \text{if} \quad \max_i x_i - 1 < \theta < \min_i x_i \quad i = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

- $f(x|\theta)/f(y|\theta)$ is constant provided

$$\max_i x_i = \max y_i \quad \& \quad \min x_i = \min y_i$$

- $T(X) = (\min_i X_i, \max_i X_i) = (X_{(1)}, X_{(n)})$ is MSS
- $\left(\left(X_{(n)} - X_{(1)}\right), \left(X_{(n)} + X_{(1)}\right)/2\right)$ is also a MSS.

Ancillary Statistics

- Sufficient Statistics contains all the information about parameter θ available from sample
- Ancillary Statistics, has a complementary purpose.

> A statistics $S(X)$ whose distribution does not depend on the parameter θ is called an ancillary statistics

- Alone ancillary statistics contains no information about θ
- When used in conjuection with other statistics it may reveal information about θ

Example: Uniform Ancillary Statistics

$X = (X_1, X_2, \ldots, X_n)$ are iid and $\sim Unif(\theta, \theta + 1)$.

$$f(x|\theta) = \begin{cases} 1 & \text{if} \quad \theta < x < \theta + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Joint pdf of $(X_{(1)}, X_{(n)})$ can be derived

$$g(x_{(1)}, x_{(n)}) = \begin{cases} n(n-1)(x_{(n)} - x_{(1)})^{n-2} & \theta < x_{(1)} < x_{(n)} < \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

Define $R = (X_{(n)} - X_{(1)})$ and $R = (X_{(n)} + X_{(1)})/2$. Find joint distributions of (R, M)

$$h(r|\theta) = n(n-1)r^{n-2}(1-r) \quad 0 < r < 1$$

(Complete!)