Lecture 9: Sufficient Statistics

11 Feb 2016

In the previous lecture, the concept and definition of sufficient statistics were covered. In this lecture, an equivalent definition for sufficient statistics, Factorization Theorem, concept of minimal sufficient statistics and relation between sufficiency and hypothesis testing are discussed.

1 Sufficient Statistics

Definition 1.1. If $p_\theta(\mathbf{x})$ is the joint pdf (or pmf) of $\mathbf{X}$ and $q_\theta(t)$ is the pdf (or pmf) of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic, if the ratio $\frac{p_\theta(\mathbf{x})}{q_\theta(T(\mathbf{x}))}$ does not depend on θ, $\forall \mathbf{x}$.

How can we find the sufficient statistics of a given population? Following theorem helps us in finding the sufficient statistics in a systematic way.

Theorem 1.2 (Factorization Theorem). *Let $f_\theta(\mathbf{x})$ denote the joint pdf of a sample $\mathbf{x}$ from population with parameter θ. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exists functions $g_\theta(t), h(\mathbf{x})$ such that for all sample points $\mathbf{x}$ and all parameter points θ,*

$$f_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))h(\mathbf{x}). \tag{1}$$

Proof. The following proof is for the discrete distribution. Suppose $T(\mathbf{X})$ is a sufficient statistic. Let $g_\theta(t) = P_\theta(T(\mathbf{x}) = t)$ and $h(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$. As $T(\mathbf{X})$ is a sufficient statistic, the conditional probability defining $h(\mathbf{x})$ does not depend on θ. Thus the above choice is valid and we have,

$$\begin{aligned} f_\theta(\mathbf{x}) &= P_\theta(\mathbf{X} = \mathbf{x}), \\ &= P_\theta((\mathbf{X} = \mathbf{x}) \text{ and } T(\mathbf{X}) = T(\mathbf{x})), \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})), \\ &= g_\theta(T(\mathbf{X}))h(\mathbf{x}). \end{aligned} \tag{2}$$

So factorization (1) is proved. It is also clear that $P_\theta(T(\mathbf{X}) = T(\mathbf{x})) = g_\theta(T(\mathbf{x}))$. Thus, $g_\theta(T(\mathbf{x}))$ is the pmf of $T(\mathbf{X})$.

Now assume, factorization (1) exists. Let $q_\theta(t)$ be the pmf of $T(\mathbf{X})$. To show $T(\mathbf{X})$ is sufficient, examine the ratio $\frac{f_\theta(\mathbf{x})}{q_\theta(T(\mathbf{x}))}$. Define, $A_{T(\mathbf{x})} = \mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})$.

$$\begin{aligned}
\frac{f_\theta(\mathbf{x})}{q_\theta(T(\mathbf{x}))} &= \frac{g_\theta(T(\mathbf{x}))h(\mathbf{x})}{q_\theta(T(\mathbf{x}))}, &(3)\\
&= \frac{g_\theta(T(\mathbf{x}))h(\mathbf{x})}{\sum_{\mathbf{y}\in A} g_\theta(T(\mathbf{x}))h(\mathbf{y})}, &(4)\\
&= \frac{g_\theta(T(\mathbf{x}))h(\mathbf{x})}{g_\theta(T(\mathbf{x}))\sum_{\mathbf{y}\in A} h(\mathbf{y})}, &(5)\\
&= \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A} h(\mathbf{y})}. &(6)
\end{aligned}$$

where, (3) follows from (1), (4) from the definition of pmf of T and (5) because T is constant on $A_{T(\mathbf{x})}$. Since the ratio does not depend on θ, $T(\mathbf{X})$ is a sufficient statistic for θ. □

Example 1.3 (Normal sufficient statistics, with variance 1). Let $X_1, ..., X_n$ be iid $n(\mu, \sigma^2)$, where $\sigma^2 = 1$. Let $\bar{X} = (X_1 + ... + X_n)/n$. The joint pdf of the sample $\mathbf{X}$ is

$$\begin{aligned}
f_\mu(\mathbf{x}) &= \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{(x_i-\mu)^2}{2}\right),\\
&= (2\pi)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2}\right),\\
&= (2\pi)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n}\frac{(x_i-\bar{x}+\bar{x}-\mu)^2}{2}\right),\\
&= (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2)}{2}\right),\\
&= (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2}\right) \exp\left(-\frac{n(\bar{x}-\mu)^2}{2}\right). \qquad (7)
\end{aligned}$$

The above expression for the pdf was already derived in the last lecture.

Define,

$$h(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2}\right),$$

which does not depend on the unknown parameter μ. The factor in equation (7) that contains μ depend on the sample $\mathbf{x}$ only through the function, $T(\mathbf{x}) = \bar{x}$, the sample mean. So we have,

$$\begin{aligned} g_\mu(t) &= \exp\left(-\frac{n(t-\mu)^2}{2}\right); & (8) \\ f_\mu(\mathbf{x}) &= g_\mu(T(\mathbf{x}))h(\mathbf{x}). & (9) \end{aligned}$$

By Factorization Theorem, $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ.

Example 1.4 (Uniform sufficient statistics). Let $X_1, ..., X_n$ be iid observations, from the discrete uniform distribution on $1, ..., \theta$, where the unknown parameter θ is a positive integer and the pmf of X_i is

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & x = 1, 2, ..., \theta. \\ 0 & \text{otherwise.} \end{cases}$$

The analysis can be carried out using indicator function as follows: $\mathbb{I}_A(x)$ is the indicator function of set A, and is equal to 1, if $x \in A$ and equal to 0 otherwise. Let $\mathcal{N} = \{1, 2, ...\}$ be the set of positive integers and let $\mathcal{N}_\theta = \{1, 2, ..., \theta\}$. Then the joint pmf of $X_1, ..., X_n$ is

$$\begin{aligned} f_\theta(x) &= \prod_{i=1}^{n} \theta^{-1} \mathbb{1}_{\mathcal{N}_\theta}(x_i), \\ &= \theta^{-n} \prod_{i=1}^{n} \mathbb{1}_{\mathcal{N}_\theta}(x_i). \qquad (10) \end{aligned}$$

Define $T(\mathbf{x}) = \max\limits_i \ x_i$. Then,

$$\prod_{i=1}^{n} \mathbb{1}_{\mathcal{N}_\theta}(x_i) = \left(\prod_{i=1}^{n} \mathbb{1}_{\mathcal{N}}(x_i)\right) \mathbb{1}_{\mathcal{N}_\theta}(T(\mathbf{x})). \qquad (11)$$

Thus, from equations (10) and (11) we obtain

$$f_\theta(x) = \theta^{-n} \mathbb{1}_{\mathcal{N}_\theta}(T(\mathbf{x})) \prod_{i=1}^{n} \mathbb{1}_{\mathcal{N}}(x_i). \qquad (12)$$

The first factor depends on $x_1, ..., x_n$ only through the value of $T(\mathbf{x}) = \max\limits_i \ x_i$, and second factor does not depend on θ. By the Factorization Theorem, $T(\mathbf{x}) = \max\limits_i \ x_i$ is a sufficient statistic for θ.

Example 1.5 (Normal sufficient statistics, both parameters unknown). Consider a normal distribution as in example (1.3), but with unknown variance, σ^2. The parameter vector is $\theta(\mu, \sigma^2)$. As in example (1.3), we can write the joint pdf of $\mathbf{X}$ as

$$f_\theta(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2)}{2\sigma^2}\right). \tag{13}$$

Any part of the joint pdf that depends on either μ or σ^2 must be included in the g function. The pdf depends on the sample $\mathbf{x}$ only through the two values $T_1(\mathbf{x}) = \bar{x}$ and $T_2(\mathbf{x}) = s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}$. So, define:

$$\begin{aligned} h(\mathbf{x}) &= 1, \\ g_\theta(t) &= g_{(\mu,\sigma^2)}(t_1, t_2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)t_2 + n(t_1 - \mu)^2}{2\sigma^2}\right), \\ f_\theta(\mathbf{x}) &= g_{(\mu,\sigma^2)}(T(\mathbf{x}))h(\mathbf{x}). \end{aligned} \tag{14}$$

By Factorization Theorem, $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = ((\bar{\mathbf{X}}), S^2)$ is a sufficient statistic for (μ, σ^2).

It should be noted that the definition of a sufficient statistic is model dependent. For another model, that is, another family of densities, the sample mean and variance may not be a sufficient statistic for the population mean and variance.

Definition 1.6 (Exponential family of distributions). An exponential family is a family of pdfs or pmfs of the following form:

$$f_{\boldsymbol{\theta}}(x) = h(x)c(\boldsymbol{\theta})exp[\sum_{i=1}^k W_i(\boldsymbol{\theta})t_i(x)] \tag{15}$$

where $\boldsymbol{\theta} = (\theta_1.......\theta_d)$, $W_i : R^d \rightarrow R$ and $t_i : R \rightarrow R$.

The pdf is an exponentiated weighted sums of functions of x where the weight is controlled by $\boldsymbol{\theta}$. $c(\boldsymbol{\theta})$ is the normalizing constant as $f_{\boldsymbol{\theta}}(x)$ is a pdf. $c(\boldsymbol{\theta})$ is commonly called as the partition function.

The general exponential family generalizes a lot of results. Some of the known distributions can be expressed in the form given in (15)

Example 1.7 (Binomial distribution in terms of exponential family). Consider a random variable X having a binomial distribution $B[n, p]$ with probability of success p. n is the number of trials.

$$\begin{aligned} f(x|\theta) &= \binom{n}{x} p^x (1-p)^{(n-x)}, \\ &= \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n, \\ &= \binom{n}{x} exp\left[xlog\left(\frac{p}{1-p}\right) + nlog(1-p)\right]. \end{aligned} \tag{16}$$

The binomial distribution is thus expressed in terms of the exponential family where $h(x) = \binom{n}{x}$, $c(\theta) = 1$, $W_1(\theta) = log\left(\frac{p}{1-p}\right)$, $t_1(x) = x$, $W_2(\theta) = log(1-p)$ and $t_2(x) = n$.

Similarly, normal distribution with unknown σ_2 and μ, gamma distribution etc. can be represented in terms of the exponential family.

Theorem 1.8. *Let X_1,X_2,...,X_n be iid observations from a pdf or pmf $f(x|\boldsymbol{\theta})$ that belongs to an exponential family, then,*

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), ..., \sum_{i=1}^{n} t_k(X_i)\right) \tag{17}$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Proof. The pdf for an exponential family distribution is represented by the following structure:

$$f_{\boldsymbol{\theta}}(x) = h(x)c(\boldsymbol{\theta})exp\left[\sum_{i=1}^{k} W_i(\boldsymbol{\theta})t_i(x)\right]. \tag{18}$$

The above pdf can be represented as product of $h(x) = h(x)$ and $g_{\boldsymbol{\theta}}(T(x)) = c(\boldsymbol{\theta})exp\left[\sum_{i=1}^{k} W_i(\boldsymbol{\theta})t_i(x)\right]$ and hence according to the Factorization Theorem, $T(\mathbf{X})$ is the sufficient statistic. $T(x)$ in this case can be identified to be all $t_i(x)$ for $i = 1, 2, .., k$. Hence the theorem follows. □

1.1 Minimal Sufficient Statistics

In the preceding section, we found one sufficient statistic for each model considered. In any problem, there may be many sufficient statistics. Out of all these, we are interested to find the 'smallest' (in the sense of data reduction) sufficient statistic.

Definition 1.9. A sufficient statistic $T(\mathbf{X})$ is called minimal sufficient statistic if for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$, i.e., if $T'(\mathbf{x}) = T'(\mathbf{y})$ then $T(\mathbf{x}) = T(\mathbf{y})$.

A minimal sufficient statistic is a sufficient statistic which can be compared with every other sufficient statistic and hence can be viewed to be the dominant one. It is not unique though. For a minimal sufficient statistic, a one to one function is also a minimal sufficient statistic.

Theorem 1.10. *Let $f(\mathbf{x}|\theta)$ be the pdf of a sample. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ doesn't depend on θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$, then $T(\mathbf{X})$ is a minimal sufficient statistic.*

Proof. Let $f(\mathbf{x}|\theta)$ be the pdf of a sample, for all $\mathbf{x} \in \boldsymbol{\chi}$ and θ; $f(\mathbf{x}|\theta) \geq 0$. Now, let us first show that the function $T(\mathbf{X})$ is a sufficient statistic. Let $\tau = \{t : t = T(\mathbf{x})$ for some $\mathbf{x} \in \boldsymbol{\chi}\}$ be the image of $\boldsymbol{\chi}$ under $T(\mathbf{x})$. Now, lets define partition sets induced by $T(\mathbf{x})$ as

$$A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}. \tag{19}$$

For each A_t, choose and fix one element $\mathbf{x}_t \in A_t$. For any $\mathbf{x} \in \boldsymbol{\chi}$, $\mathbf{x}_{T(\mathbf{x})}$ is the fixed element that is in the same set A_t as $\mathbf{x}$. Since $\mathbf{x}$ and $\mathbf{x}_{T(\mathbf{x})}$ are in the same set A_t, $T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$ and hence $f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ is constant as a function of θ. Thus, we can define a function on $\boldsymbol{\chi}$ by $h(\mathbf{x}) = f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ and h does not depend on θ. Define a function on τ by $g(t|\theta) = f(\mathbf{x}_t|\theta)$. Then,

$$\begin{aligned} f(\mathbf{x}|\theta) &= \frac{f(\mathbf{x}_{T(\mathbf{x})}|\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)}, & (20) \\ &= g(T(\mathbf{x})|\theta) h(\mathbf{x}). & (21) \end{aligned}$$

Now, we need to show that $T(\mathbf{X})$ is the minimal sufficient statistic. Let $T'(\mathbf{X})$ be any other sufficient statistic. By the Factorization Theorem, there exist functions g' and h' such that $f(\mathbf{x}|\theta) = g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})$. At any two sample points $\mathbf{x}$ and $\mathbf{y}$, let $T'(\mathbf{x}) = T'(\mathbf{y})$, then,

$$\begin{aligned} \frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} &= \frac{g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})}{g'(T'(\mathbf{y})|\theta) h'(\mathbf{y})}, & (22) \\ &= \frac{h'(\mathbf{x})}{h'(\mathbf{y})}. & (23) \end{aligned}$$

Hence, it does not depend on θ when $T'(\mathbf{x}) = T'(\mathbf{y})$. Thus, $T(\mathbf{X})$ is a function of $T'(\mathbf{X})$ and is the minimal sufficient statistic. □

Example 1.11 (Normal minimal sufficient statistic). Let X_1,X_2,...,X_n be iid with pdf $n(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Let $\mathbf{x} = (x_1....x_n)$ and $\mathbf{y} = (y_1....y_n)$ denote two sample points and let $\bar{x}$ and $\bar{y}$ denote the sample means and ${S_{\mathbf{x}}}^2$ and ${S_{\mathbf{y}}}^2$ denote the sample variance of the samples $\mathbf{x}$ and $\mathbf{y}$ respectively. Then,

$$\begin{aligned}
\frac{f(\mathbf{x}|(\mu,\sigma^2))}{f(\mathbf{y}|(\mu,\sigma^2))} &= \frac{(2\pi\sigma^2)^{-\frac{n}{2}}\exp(-[n(\bar{x}-\mu)^2+(n-1){S_{\mathbf{x}}}^2])/(2\sigma^2)}{(2\pi\sigma^2)^{-\frac{n}{2}}\exp(-[n(\bar{y}-\mu)^2+(n-1){S_{\mathbf{y}}}^2])/(2\sigma^2)}, && (24)\\
&= \exp\left(\frac{n((\bar{y}-\mu)^2-(\bar{x}-\mu)^2)+(n-1)({S_{\mathbf{x}}}^2-{S_{\mathbf{y}}}^2)}{2\sigma^2}\right), && (25)\\
&= \exp\left(\frac{-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)({S_{\mathbf{x}}}^2-{S_{\mathbf{y}}}^2))}{2\sigma^2}\right). && (26)
\end{aligned}$$

In this example, $\theta = (\mu, \sigma^2)$. Hence, the above ratio doesn't depend on θ if and only if $(\bar{x})^2 - (\bar{y})^2 = 0$, $\bar{x} - \bar{y} = 0$ and ${S_{\mathbf{x}}}^2 - {S_{\mathbf{y}}}^2 = 0$. Thus, $\bar{x} = \bar{y}$ and ${S_{\mathbf{x}}}^2 = {S_{\mathbf{y}}}^2$. According to theorem 1.10, $T(\mathbf{X}) = (\bar{X}, S^2)$ is a minimal sufficient statistic for $n(\mu, \sigma^2)$.

Example 1.12 (Uniform minimal sufficient statistic). Suppose X_1,....,X_n are iid uniform observations on interval $(\theta, \theta + 1)$, $\theta \in (-\infty, \infty)$, then the joint pdf of $\mathbf{X}$ is

$$f(\mathbf{x}|\theta) = \begin{cases} 1, & \text{if } \theta < x_i < \theta + 1, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

As $\theta < x_i$ for all $i = 1, 2, ..., n$, then, $\theta < \min\limits_i x_i$. Similarly, as $x_i < \theta + 1$ for all $i = 1, 2...., n$, hence $\max\limits_i x_i - 1 < \theta$. Hence, $f(\mathbf{x}|\theta)$ can be written as,

$$f(\mathbf{x}|\theta) = \begin{cases} 1, & \text{if } \max\limits_i\ x_i - 1 < \theta < \min\limits_i\ x_i, \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Now, for any two samples $\mathbf{x}$ and $\mathbf{y}$, $\dfrac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ is positive if and only if $f(\mathbf{x}|\theta) = f(\mathbf{y}|\theta) = 1$. Hence, $\max\limits_i x_i - 1 < \theta < \min\limits_i x_i$ and $\max\limits_i y_i - 1 < \theta < \min\limits_i y_i$. Hence, $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimum sufficient statistic.

2 Sufficient Statistics in Hypothesis Testing

In the first few lectures, we learnt about detection using hypothesis testing. In this section, we will learn how sufficient statistics are related to hypothesis testing.

All the hypothesis testing were likelihood ratio tests. Likelihood ratio tests using entire sample $\mathbf{X}$ is equivalent to performing the likelihood ratio test using sufficient statistics $T(\mathbf{X})$. Hence, instead of considering the whole sample, we can just use the sufficient statistics in these tests.

Theorem 2.1. *Consider a composite hypothesis test with* $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Lambda}_0$*,* $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Lambda}_1$*. Let* $T(\mathbf{X})$ *be a sufficient statistic for* $\boldsymbol{\theta}$*. Let the generalized likelihood ratio test (GLRT) be*

$$\lambda(\mathbf{x}) = \frac{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Lambda}_1} P_{\boldsymbol{\theta}}(x)}{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Lambda}_0} P_{\boldsymbol{\theta}}(x)}, \tag{29}$$

and the GLRT statistic based on T *be*

$$\lambda^*(t) = \frac{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Lambda}_1} P_{\boldsymbol{\theta}}(T(\mathbf{x}) = t)}{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Lambda}_0} P_{\boldsymbol{\theta}}(T(\mathbf{x}) = t)}, \tag{30}$$

then,

$$\lambda(\mathbf{x}) = \lambda^*(T(\mathbf{x})), \tag{31}$$

$\forall$ $\mathbf{x}$ *in the sample space.*