

# Stat 709: Mathematical Statistics

## Lecture 19

Jun Shao

Department of Statistics  
University of Wisconsin  
Madison, WI 53706, USA

# Lecture 19: Sufficient statistics and factorization theorem

## Data reduction without loss of information

A statistic  $T(X)$  provides a reduction of the  $\sigma$ -field  $\sigma(X)$

Does such a reduction results in any loss of information concerning the unknown population?

If a statistic  $T(X)$  is fully as informative as the original sample  $X$ , then statistical analyses can be done using  $T(X)$  that is simpler than  $X$ .

The next concept describes what we mean by fully informative.

### Definition 2.4 (Sufficiency)

Let  $X$  be a sample from an unknown population  $P \in \mathcal{P}$ , where  $\mathcal{P}$  is a family of populations.

A statistic  $T(X)$  is said to be *sufficient* for  $P \in \mathcal{P}$  (or for  $\theta \in \Theta$  when  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  is a parametric family) iff the conditional distribution of  $X$  given  $T$  is *known* (does not depend on  $P$  or  $\theta$ ).

# Lecture 19: Sufficient statistics and factorization theorem

## Data reduction without loss of information

A statistic  $T(X)$  provides a reduction of the  $\sigma$ -field  $\sigma(X)$

Does such a reduction results in any loss of information concerning the unknown population?

If a statistic  $T(X)$  is fully as informative as the original sample  $X$ , then statistical analyses can be done using  $T(X)$  that is simpler than  $X$ .

The next concept describes what we mean by fully informative.

## Definition 2.4 (Sufficiency)

Let  $X$  be a sample from an unknown population  $P \in \mathcal{P}$ , where  $\mathcal{P}$  is a family of populations.

A statistic  $T(X)$  is said to be *sufficient* for  $P \in \mathcal{P}$  (or for  $\theta \in \Theta$  when  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  is a parametric family) iff the conditional distribution of  $X$  given  $T$  is *known* (does not depend on  $P$  or  $\theta$ ).

## Remarks

- Once we observe  $X$  and compute a sufficient statistic  $T(X)$ , the original data  $X$  do not contain any further information concerning the unknown population  $P$  (since its conditional distribution is unrelated to  $P$ ) and can be discarded.
- A sufficient statistic  $T(X)$  contains all information about  $P$  contained in  $X$  and provides a reduction of the data if  $T$  is not one-to-one.
- The concept of sufficiency depends on the given family  $\mathcal{P}$ .
- If  $T$  is sufficient for  $P \in \mathcal{P}$ , then  $T$  is also sufficient for  $P \in \mathcal{P}_0 \subset \mathcal{P}$  but not necessarily sufficient for  $P \in \mathcal{P}_1 \supset \mathcal{P}$ .

### Example 2.10

Suppose that  $X = (X_1, \dots, X_n)$  and  $X_1, \dots, X_n$  are i.i.d. from the binomial distribution with the p.d.f. (w.r.t. the counting measure)

$$f_\theta(z) = \theta^z(1 - \theta)^{1-z} I_{\{0,1\}}(z), \quad z \in \mathcal{R}, \quad \theta \in (0, 1).$$

Consider the statistic  $T(X) = \sum_{i=1}^n X_i$ , which is the number of ones in  $X$ .

## Remarks

- Once we observe  $X$  and compute a sufficient statistic  $T(X)$ , the original data  $X$  do not contain any further information concerning the unknown population  $P$  (since its conditional distribution is unrelated to  $P$ ) and can be discarded.
- A sufficient statistic  $T(X)$  contains all information about  $P$  contained in  $X$  and provides a reduction of the data if  $T$  is not one-to-one.
- The concept of sufficiency depends on the given family  $\mathcal{P}$ .
- If  $T$  is sufficient for  $P \in \mathcal{P}$ , then  $T$  is also sufficient for  $P \in \mathcal{P}_0 \subset \mathcal{P}$  but not necessarily sufficient for  $P \in \mathcal{P}_1 \supset \mathcal{P}$ .

## Example 2.10

Suppose that  $X = (X_1, \dots, X_n)$  and  $X_1, \dots, X_n$  are i.i.d. from the binomial distribution with the p.d.f. (w.r.t. the counting measure)

$$f_\theta(\mathbf{z}) = \theta^z(1 - \theta)^{1-z} I_{\{0,1\}}(\mathbf{z}), \quad \mathbf{z} \in \mathcal{R}, \quad \theta \in (0, 1).$$

Consider the statistic  $T(X) = \sum_{i=1}^n X_i$ , which is the number of ones in  $X$ .

## Example 2.10 (continued)

For any realization  $x$  of  $X$ ,  $x$  is a sequence of  $n$  ones and zeros.  $T$  contains all information about  $\theta$ , since  $\theta$  is the probability of an occurrence of a one in  $x$  and given  $T = t$ , what is left in the data set  $x$  is the redundant information about the positions of  $t$  ones.

To show  $T$  is sufficient for  $\theta$ , we compute  $P(X = x | T = t)$ .

Let  $t = 0, 1, \dots, n$  and  $B_t = \{(x_1, \dots, x_n) : x_i = 0, 1, \sum_{i=1}^n x_i = t\}$ .

If  $x \notin B_t$ , then  $P(X = x, T = t) = 0$ .

If  $x \in B_t$ , then

$$P(X = x, T = t) = \prod_{i=1}^n P(X_i = x_i) = \theta^t (1 - \theta)^{n-t} \prod_{i=1}^n I_{\{0,1\}}(x_i).$$

Also

$$P(T = t) = \binom{n}{t} \theta^t (1 - \theta)^{n-t} I_{\{0,1,\dots,n\}}(t).$$

Then

$$P(X = x | T = t) = \frac{P(X = x, T = t)}{P(T = t)} = \frac{1}{\binom{n}{t}} I_{B_t}(x)$$

is a known p.d.f. (does not depend on  $\theta$ ).

Hence  $T(X)$  is sufficient for  $\theta \in (0, 1)$  according to Definition 2.4.

## How to find a sufficient statistic?

Finding a sufficient statistic by means of the definition is not convenient. It involves guessing a statistic  $T$  that might be sufficient and computing the conditional distribution of  $X$  given  $T = t$ .

For families of populations having p.d.f.'s, a simple way of finding sufficient statistics is to use the factorization theorem.

### Theorem 2.2 (The factorization theorem)

Suppose that  $X$  is a sample from  $P \in \mathcal{P}$  and  $\mathcal{P}$  is a family of probability measures on  $(\mathcal{R}^n, \mathcal{B}^n)$  dominated by a  $\sigma$ -finite measure  $\nu$ . Then  $T(X)$  is sufficient for  $P \in \mathcal{P}$  iff there are nonnegative Borel functions  $h$  (which does not depend on  $P$ ) on  $(\mathcal{R}^n, \mathcal{B}^n)$  and  $g_P$  (which depends on  $P$ ) on the range of  $T$  such that

$$\frac{dP}{d\nu}(x) = g_P(T(x))h(x).$$

To prove Theorem 2.2, we need the following lemma whose proof can be found in the textbook.

## How to find a sufficient statistic?

Finding a sufficient statistic by means of the definition is not convenient. It involves guessing a statistic  $T$  that might be sufficient and computing the conditional distribution of  $X$  given  $T = t$ .

For families of populations having p.d.f.'s, a simple way of finding sufficient statistics is to use the factorization theorem.

### Theorem 2.2 (The factorization theorem)

Suppose that  $X$  is a sample from  $P \in \mathcal{P}$  and  $\mathcal{P}$  is a family of probability measures on  $(\mathcal{R}^n, \mathcal{B}^n)$  dominated by a  $\sigma$ -finite measure  $\nu$ . Then  $T(X)$  is sufficient for  $P \in \mathcal{P}$  iff there are nonnegative Borel functions  $h$  (which does not depend on  $P$ ) on  $(\mathcal{R}^n, \mathcal{B}^n)$  and  $g_P$  (which depends on  $P$ ) on the range of  $T$  such that

$$\frac{dP}{d\nu}(x) = g_P(T(x))h(x).$$

To prove Theorem 2.2, we need the following lemma whose proof can be found in the textbook.

## How to find a sufficient statistic?

Finding a sufficient statistic by means of the definition is not convenient. It involves guessing a statistic  $T$  that might be sufficient and computing the conditional distribution of  $X$  given  $T = t$ .

For families of populations having p.d.f.'s, a simple way of finding sufficient statistics is to use the factorization theorem.

### Theorem 2.2 (The factorization theorem)

Suppose that  $X$  is a sample from  $P \in \mathcal{P}$  and  $\mathcal{P}$  is a family of probability measures on  $(\mathcal{R}^n, \mathcal{B}^n)$  dominated by a  $\sigma$ -finite measure  $\nu$ . Then  $T(X)$  is sufficient for  $P \in \mathcal{P}$  iff there are nonnegative Borel functions  $h$  (which does not depend on  $P$ ) on  $(\mathcal{R}^n, \mathcal{B}^n)$  and  $g_P$  (which depends on  $P$ ) on the range of  $T$  such that

$$\frac{dP}{d\nu}(x) = g_P(T(x))h(x).$$

To prove Theorem 2.2, we need the following lemma whose proof can be found in the textbook.

## Lemma 2.1

If a family  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure, then  $\mathcal{P}$  is dominated by a probability measure  $Q = \sum_{i=1}^{\infty} c_i P_i$ , where  $c_i$ 's are nonnegative constants with  $\sum_{i=1}^{\infty} c_i = 1$  and  $P_i \in \mathcal{P}$ .

## Proof of Theorem 2.2

(i) Suppose that  $T$  is sufficient for  $P \in \mathcal{P}$ .

For any  $A \in \mathcal{B}^n$ ,  $P(A|T)$  does not depend on  $P$ .

Let  $Q$  be the probability measure in Lemma 2.1.

By Fubini's theorem and the result in Exercise 35 of §1.6,

$$\begin{aligned} Q(A \cap B) &= \sum_{j=1}^{\infty} c_j P_j(A \cap B) = \sum_{j=1}^{\infty} c_j \int_B P(A|T) dP_j \\ &= \int_B \sum_{j=1}^{\infty} c_j P(A|T) dP_j = \int_B P(A|T) dQ \end{aligned}$$

for any  $B \in \sigma(T)$ .

Hence,  $P(A|T) = E_Q(I_A|T)$  a.s.  $Q$ , where  $E_Q(I_A|T)$  denotes the conditional expectation of  $I_A$  given  $T$  w.r.t.  $Q$ .

## Lemma 2.1

If a family  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure, then  $\mathcal{P}$  is dominated by a probability measure  $Q = \sum_{i=1}^{\infty} c_i P_i$ , where  $c_i$ 's are nonnegative constants with  $\sum_{i=1}^{\infty} c_i = 1$  and  $P_i \in \mathcal{P}$ .

## Proof of Theorem 2.2

(i) Suppose that  $T$  is sufficient for  $P \in \mathcal{P}$ .

For any  $A \in \mathcal{B}^n$ ,  $P(A|T)$  does not depend on  $P$ .

Let  $Q$  be the probability measure in Lemma 2.1.

By Fubini's theorem and the result in Exercise 35 of §1.6,

$$\begin{aligned} Q(A \cap B) &= \sum_{j=1}^{\infty} c_j P_j(A \cap B) = \sum_{j=1}^{\infty} c_j \int_B P(A|T) dP_j \\ &= \int_B \sum_{j=1}^{\infty} c_j P(A|T) dP_j = \int_B P(A|T) dQ \end{aligned}$$

for any  $B \in \sigma(T)$ .

Hence,  $P(A|T) = E_Q(I_A|T)$  a.s.  $Q$ , where  $E_Q(I_A|T)$  denotes the conditional expectation of  $I_A$  given  $T$  w.r.t.  $Q$ .

## Lemma 2.1

If a family  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure, then  $\mathcal{P}$  is dominated by a probability measure  $Q = \sum_{i=1}^{\infty} c_i P_i$ , where  $c_i$ 's are nonnegative constants with  $\sum_{i=1}^{\infty} c_i = 1$  and  $P_i \in \mathcal{P}$ .

## Proof of Theorem 2.2

(i) Suppose that  $T$  is sufficient for  $P \in \mathcal{P}$ .

For any  $A \in \mathcal{B}^n$ ,  $P(A|T)$  does not depend on  $P$ .

Let  $Q$  be the probability measure in Lemma 2.1.

By Fubini's theorem and the result in Exercise 35 of §1.6,

$$\begin{aligned} Q(A \cap B) &= \sum_{j=1}^{\infty} c_j P_j(A \cap B) = \sum_{j=1}^{\infty} c_j \int_B P(A|T) dP_j \\ &= \int_B \sum_{j=1}^{\infty} c_j P(A|T) dP_j = \int_B P(A|T) dQ \end{aligned}$$

for any  $B \in \sigma(T)$ .

Hence,  $P(A|T) = E_Q(I_A|T)$  a.s.  $Q$ , where  $E_Q(I_A|T)$  denotes the conditional expectation of  $I_A$  given  $T$  w.r.t.  $Q$ .

## Proof of Theorem 2.2 (continued)

Let  $g_p(T)$  be the Radon-Nikodym derivative  $dP/dQ$  on the space  $(\mathcal{B}^n, \sigma(T), Q)$ .

Then

$$\begin{aligned} P(A) &= \int P(A|T) dP = \int E_Q(I_A|T) dP = \int E_Q(I_A|T) g_p(T) dQ \\ &= \int E_Q[I_A g_p(T)|T] dQ = \int I_A g_p(T) dQ = \int_A g_p(T) \frac{dQ}{dv} dv \end{aligned}$$

for any  $A \in \mathcal{B}^n$ .

Hence,

$$\frac{dP}{dv}(x) = g_p(T(x)) h(x) \quad (1)$$

holds with  $h = dQ/dv$ .

(ii) Suppose that (1) holds.

Then

$$\frac{dP}{dQ} = \frac{dP}{dv} \bigg/ \sum_{i=1}^{\infty} c_i \frac{dP_i}{dv} = g_p(T) \bigg/ \sum_{i=1}^{\infty} g_{P_i}(T) \quad \text{a.s. } Q, \quad (2)$$

where the second equality follows from Exercise 35 in §1.6.

## Proof of Theorem 2.2 (continued)

Let  $g_P(T)$  be the Radon-Nikodym derivative  $dP/dQ$  on the space  $(\mathcal{R}^n, \sigma(T), Q)$ .

Then

$$\begin{aligned} P(A) &= \int P(A|T) dP = \int E_Q(I_A|T) dP = \int E_Q(I_A|T) g_P(T) dQ \\ &= \int E_Q[I_A g_P(T)|T] dQ = \int I_A g_P(T) dQ = \int_A g_P(T) \frac{dQ}{d\nu} d\nu \end{aligned}$$

for any  $A \in \mathcal{B}^n$ .

Hence,

$$\frac{dP}{d\nu}(x) = g_P(T(x)) h(x) \quad (1)$$

holds with  $h = dQ/d\nu$ .

(ii) Suppose that (1) holds.

Then

$$\frac{dP}{dQ} = \frac{dP}{d\nu} \bigg/ \sum_{i=1}^{\infty} c_i \frac{dP_i}{d\nu} = g_P(T) \bigg/ \sum_{i=1}^{\infty} g_{P_i}(T) \quad \text{a.s. } Q, \quad (2)$$

where the second equality follows from Exercise 35 in §1.6.

## Proof of Theorem 2.2 (continued)

Let  $A \in \sigma(X)$  and  $P \in \mathcal{P}$ .

The sufficiency of  $T$  follows from

$$P(A|T) = E_Q(I_A|T) \quad \text{a.s. } P, \quad (3)$$

where  $E_Q(I_A|T)$  is given in part (i) of the proof.

This is because  $E_Q(I_A|T)$  does not vary with  $P \in \mathcal{P}$ , and result (3) and Theorem 1.7 imply that the conditional distribution of  $X$  given  $T$  is determined by  $E_Q(I_A|T)$ ,  $A \in \sigma(X)$ .

By (2),  $dP/dQ$  is a Borel function of  $T$ .

For any  $B \in \sigma(T)$ ,

$$\begin{aligned} \int_B E_Q(I_A|T) dP &= \int_B E_Q(I_A|T) \frac{dP}{dQ} dQ \\ &= \int_B E_Q \left( I_A \frac{dP}{dQ} \middle| T \right) dQ = \int_B I_A \frac{dP}{dQ} dQ = \int_B I_A dP. \end{aligned}$$

This proves (3) and completes the proof.

## Proof of Theorem 2.2 (continued)

Let  $A \in \sigma(X)$  and  $P \in \mathcal{P}$ .

The sufficiency of  $T$  follows from

$$P(A|T) = E_Q(I_A|T) \quad \text{a.s. } P, \quad (3)$$

where  $E_Q(I_A|T)$  is given in part (i) of the proof.

This is because  $E_Q(I_A|T)$  does not vary with  $P \in \mathcal{P}$ , and result (3) and Theorem 1.7 imply that the conditional distribution of  $X$  given  $T$  is determined by  $E_Q(I_A|T)$ ,  $A \in \sigma(X)$ .

By (2),  $dP/dQ$  is a Borel function of  $T$ .

For any  $B \in \sigma(T)$ ,

$$\begin{aligned} \int_B E_Q(I_A|T) dP &= \int_B E_Q(I_A|T) \frac{dP}{dQ} dQ \\ &= \int_B E_Q \left( I_A \frac{dP}{dQ} \middle| T \right) dQ = \int_B I_A \frac{dP}{dQ} dQ = \int_B I_A dP. \end{aligned}$$

This proves (3) and completes the proof.

## Exponential families

If  $\mathcal{P}$  is an exponential family, then Theorem 2.2 can be applied with

$$g_{\theta}(t) = \exp\{[\eta(\theta)]^{\tau}t - \xi(\theta)\},$$

i.e.,  $T$  is a sufficient statistic for  $\theta \in \Theta$ .

In Example 2.10 the joint distribution of  $X$  is in an exponential family with  $T(X) = \sum_{i=1}^n X_i$ .

Hence, we can conclude that  $T$  is sufficient for  $\theta \in (0, 1)$  without computing the conditional distribution of  $X$  given  $T$ .

### Example 2.11 (Truncation families)

Let  $\phi(x)$  be a positive Borel function on  $(\mathcal{R}, \mathcal{B})$  such that

$$\int_a^b \phi(x) dx < \infty \text{ for any } a \text{ and } b, \quad -\infty < a < b < \infty.$$

Let  $\theta = (a, b)$ ,  $\Theta = \{(a, b) \in \mathcal{R}^2 : a < b\}$ , and

$$f_{\theta}(x) = c(\theta)\phi(x)I_{(a,b)}(x), \quad c(\theta) = \left[ \int_a^b \phi(x) dx \right]^{-1}$$

## Exponential families

If  $\mathcal{P}$  is an exponential family, then Theorem 2.2 can be applied with

$$g_{\theta}(t) = \exp\{[\eta(\theta)]^{\tau} t - \xi(\theta)\},$$

i.e.,  $T$  is a sufficient statistic for  $\theta \in \Theta$ .

In Example 2.10 the joint distribution of  $X$  is in an exponential family with  $T(X) = \sum_{i=1}^n X_i$ .

Hence, we can conclude that  $T$  is sufficient for  $\theta \in (0, 1)$  without computing the conditional distribution of  $X$  given  $T$ .

## Example 2.11 (Truncation families)

Let  $\phi(x)$  be a positive Borel function on  $(\mathcal{R}, \mathcal{B})$  such that

$$\int_a^b \phi(x) dx < \infty \text{ for any } a \text{ and } b, \quad -\infty < a < b < \infty.$$

Let  $\theta = (a, b)$ ,  $\Theta = \{(a, b) \in \mathcal{R}^2 : a < b\}$ , and

$$f_{\theta}(x) = c(\theta)\phi(x)I_{(a,b)}(x), \quad c(\theta) = \left[ \int_a^b \phi(x) dx \right]^{-1}$$

## Example 2.11 (continued)

Then  $\{f_\theta : \theta \in \Theta\}$ , called a truncation family, is a parametric family dominated by the Lebesgue measure on  $\mathcal{R}$ .

Let  $X_1, \dots, X_n$  be i.i.d. random variables having the p.d.f.  $f_\theta$ .

Then the joint p.d.f. of  $X = (X_1, \dots, X_n)$  is

$$\prod_{i=1}^n f_\theta(x_i) = [c(\theta)]^n I_{(a, \infty)}(x_{(1)}) I_{(-\infty, b)}(x_{(n)}) \prod_{i=1}^n \phi(x_i), \quad (4)$$

where  $x_{(i)}$  is the  $i$ th ordered value of  $x_1, \dots, x_n$ .

Let  $T(X) = (X_{(1)}, X_{(n)})$ ,  $g_\theta(t_1, t_2) = [c(\theta)]^n I_{(a, \infty)}(t_1) I_{(-\infty, b)}(t_2)$ , and  $h(x) = \prod_{i=1}^n \phi(x_i)$ .

By (4) and Theorem 2.2,  $T(X)$  is sufficient for  $\theta \in \Theta$ .

## Example 2.12 (Order statistics)

Let  $X = (X_1, \dots, X_n)$  and  $X_1, \dots, X_n$  be i.i.d. random variables having a distribution  $P \in \mathcal{P}$ , where  $\mathcal{P}$  is the family of distributions on  $\mathcal{R}$  having Lebesgue p.d.f.'s.

Let  $X_{(1)}, \dots, X_{(n)}$  be the order statistics given in Example 2.9.

Note that the joint p.d.f. of  $X$  is

$$f(x_1) \cdots f(x_n) = f(x_{(1)}) \cdots f(x_{(n)}).$$

Hence,  $T(X) = (X_{(1)}, \dots, X_{(n)})$  is sufficient for  $P \in \mathcal{P}$ .

The order statistics can be shown to be sufficient even when  $\mathcal{P}$  is not dominated by any  $\sigma$ -finite measure, but Theorem 2.2 is not applicable (see Exercise 31 in §2.6).