Lecture 11: Sufficiency and minimal sufficiency

Data reduction without loss of information

A statistic $T(X)$ provides a reduction of the σ-field $\sigma(X)$
Does such a reduction results in any loss of information about P?
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 $P \in \mathscr{P}$, where $\mathscr{P}$ is a family of populations.
A statistic $T(X)$ is said to be *sufficient* for $P \in \mathscr{P}$ (or for $\theta \in \Theta$ when $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric family) iff the conditional distribution of X given T is *known* (does not depend on P or θ).

- Once we compute a sufficient statistic $T(X)$, the original data X do not contain any further information about P and can be discarded.
- The concept of sufficiency depends on the given family $\mathscr{P}$.

Example 2.10

$X = (X_1, ..., X_n)$ and $X_1, ..., X_n$ are i.i.d. from the binomial distribution
Consider the statistic $T(X) = \sum_{i=1}^n X_i$, which is the number of ones in X.
For any realization x of X, x is a sequence of n ones and zeros.
T contains all information about θ, since θ 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.
Let $t = 0, 1, ..., n$ and $B_t = \{(x_1, ..., x_n) : \ x_i = 0, 1, \ \sum_{i=1}^n x_i = t\}$.
To show T is sufficient for θ, we compute, for $x \in B_t$,

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

It is 0 if $x \notin B_t$.
Then

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

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

is a known p.d.f. (does not depend on θ).
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 $P \in \mathscr{P}$ and $\mathscr{P}$ is a family of probability measures on $(\mathscr{R}^n, \mathscr{B}^n)$ dominated by a σ-finite measure ν. Then $T(X)$ is sufficient for $P \in \mathscr{P}$ iff there are nonnegative Borel functions h (which does not depend on P) on $(\mathscr{R}^n, \mathscr{B}^n)$ and $g_{_P}$ (which depends on P) on the range of T such that

$$\frac{dP}{d\nu}(x) = g_{_P}\big(T(x)\big)h(x).$$

Lemma 2.1

If a family $\mathscr{P}$ is dominated by a σ-finite measure, then $\mathscr{P}$ 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 $P_i \in \mathscr{P}$.

Proof of Theorem 2.2

(i) Suppose that T is sufficient for $P \in \mathscr{P}$.
For any $A \in \mathscr{B}^n$, $P(A|T)$ does not depend on P.
Let Q be the probability measure in Lemma 2.1.
By Fubini's theorem and Exercise 35 of §1.6, for any $B \in \sigma(T)$,

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

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.
For any $A \in \mathscr{B}^n$, with $g_P(T) = dP/dQ$ on the space $(\mathscr{R}^n, \sigma(T), Q)$,

$$P(A) = \int P(A|T) dP = \int E_Q(I_A|T) dP = \int E_Q(I_A|T) g_P(T) dQ$$

$$= \int E_Q[I_A g_P(T)|T] dQ = \int I_A g_P(T) dQ = \int_A g_P(T) \frac{dQ}{d\nu} d\nu$$

Hence,

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

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

Proof of Theorem 2.2 (continued)

(ii) Suppose that (1) holds.

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

where the second equality follows from Exercise 35 in §1.6.
Let $A \in \sigma(X)$, $P \in \mathscr{P}$, and $E_Q(I_A|T)$ be given in part (i) of the proof.
By (2), dP/dQ is a Borel function of T.
For any $B \in \sigma(T)$,

$$\int_B E_Q(I_A|T)dP = \int_B E_Q(I_A|T)\frac{dP}{dQ}dQ$$

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

This proves

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

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

Exponential famlilies

If $\mathscr{P}$ 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 2.12 (Order statistics)

Let $X_1,...,X_n$ be i.i.d. random variables having a distribution $P \in \mathscr{P}$, where $\mathscr{P}$ is the family of distributions on $\mathscr{R}$ having Lebesgue p.d.f.'s.
Let $X_{(1)},...,X_{(n)}$ 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, $T(X) = (X_{(1)},...,X_{(n)})$ is sufficient for $P \in \mathscr{P}$.

Minimal sufficiency

Maximal reduction without loss of information

- There are many sufficient statistics for a given family $\mathscr{P}$.
- In fact, X (the whole data set) is sufficient.
- If T is a sufficient statistic and $T = \psi(S)$, where ψ is measurable and S is another statistic, then S is sufficient.
- This is obvious from Theorem 2.2 if the population has a p.d.f., but it can be proved directly from Definition 2.4 (Exercise 25).
- For instance, if $X_1, \ldots, X_n$ are iid with $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$, then $(\sum_{i=1}^m X_i,\ \sum_{i=m+1}^n X_i)$ is sufficient for θ, where m is any fixed integer between 1 and n.
- If T is sufficient and $T = \psi(S)$ with a measurable function ψ that is not one-to-one, then $\sigma(T) \subset \sigma(S)$, and T is more useful than S, since T provides a further reduction of the data (or σ-field) without loss of information.
- Is there a sufficient statistics that provides “maximal" reduction of the data?

Convention

If a statement holds except for outcomes in an event A satisfying $P(A) = 0$ for all $P \in \mathscr{P}$, then we say that the statement holds a.s. $\mathscr{P}$.

Definition 2.5 Minimal sufficiency

Let T be a sufficient statistic for $P \in \mathscr{P}$.
T is called a *minimal sufficient* Statistic iff, for any other statistic S sufficient for $P \in \mathscr{P}$, there is a measurable function ψ such that $T = \psi(S)$ a.s. $\mathscr{P}$

Existence and uniqueness

Minimal sufficient statistics exist when $\mathscr{P}$ contains distributions on $\mathscr{R}^k$ dominated by a σ-finite measure (Bahadur, 1957).

If both T and S are minimal sufficient statistics, then by definition there is one-to-one measurable function ψ such that $T = \psi(S)$ a.s. $\mathscr{P}$

Hence, the minimal sufficient statistic is unique in the sense that two statistics that are one-to-one measurable functions of each other can be treated as one statistic.

Example 2.13

Let $X_1,\ldots,X_n$ be i.i.d. random variables form P_θ, the uniform distribution $U(\theta,\theta+1)$, $\theta\in R$, $n>1$.
The joint Lebesgue p.d.f. of $(X_1,\ldots,X_n)$ is

$$f_\theta(x)=\prod_{i=1}^{n} I_{(\theta,\theta+1)}(x_i)=I_{(x_{(n)}-1,x_{(1)})}(\theta),\quad x=(x_1,\ldots,x_n)\in\mathscr{R}^n,$$

where $x_{(i)}$ denotes the ith smallest value of $x_1,\ldots,x_n$.
By Theorem 2.2, $T=(X_{(1)},X_{(n)})$ is sufficient for θ.
Note that

$$x_{(1)}=\sup\{\theta:\ f_\theta(x)>0\}\ \text{ and } x_{(n)}=1+\inf\{\theta:\ f_\theta(x)>0\}.$$

If $S(X)$ is a statistic sufficient for θ, then by Theorem 2.2, there are Borel functions h and g_θ such that $f_\theta(x)=g_\theta(S(x))h(x)$.
For x with $h(x)>0$,

$$x_{(1)}=\sup\{\theta:\ g_\theta(S(x))>0\}\ \text{ and } x_{(n)}=1+\inf\{\theta:\ g_\theta(S(x))>0\}.$$

Hence, there is a measurable function ψ such that $T(x)=\psi(S(x))$ when $h(x)>0$.
Since $h>0$, a.s. $\mathscr{P}$, we conclude that T is minimal sufficient.

Finding a minimal sufficient statistic by definition is not convenient. The next theorem is a useful tool.

Theorem 2.3

Let $\mathscr{P}$ be a family of distributions on $\mathscr{R}^k$.

(i) Suppose that $\mathscr{P}_0 \subset \mathscr{P}$ and a.s. $\mathscr{P}_0$ implies a.s. $\mathscr{P}$.
If T is sufficient for $P \in \mathscr{P}$ and minimal sufficient for $P \in \mathscr{P}_0$, then T is minimal sufficient for $P \in \mathscr{P}$.

(ii) Suppose that $\mathscr{P}$ contains p.d.f.'s $f_0, f_1, f_2, \ldots$, w.r.t. a σ-finite ν.
Let $f_\infty(x) = \sum_{i=0}^\infty c_i f_i(x)$, where $c_i > 0$ for all i and $\sum_{i=0}^\infty c_i = 1$, and let $T_i(x) = f_i(x)/f_\infty(x)$ when $f_\infty(x) > 0$, $i = 0, 1, 2, \ldots$.
Then $T(X) = (T_0, T_1, T_2, \ldots)$ is minimal sufficient for $P \in \mathscr{P}$.
Furthermore, if $\{x: \ f_i(x) > 0\} \subset \{x: \ f_0(x) > 0\}$ for all i, then we may replace $f_\infty(x)$ by $f_0(x)$, in which case $T(X) = (T_1, T_2, \ldots)$ is minimal sufficient for $P \in \mathscr{P}$.

(iii) Suppose that $\mathscr{P}$ contains p.d.f.'s f_P w.r.t. a σ-finite measure and that there exists a sufficient statistic $T(X)$ such that, for any possible values x and y of X, $f_P(x) = f_P(y)\phi(x, y)$ for all P implies $T(x) = T(y)$, where ϕ is a measurable function.
Then $T(X)$ is minimal sufficient for $P \in \mathscr{P}$.

Proof

(i) If S is sufficient for $P \in \mathscr{P}$, then it is also sufficient for $P \in \mathscr{P}_0$ and, therefore, $T = \psi(S)$ a.s. $\mathscr{P}_0$ holds for a measurable function ψ. The result follows from the assumption that a.s. $\mathscr{P}_0$ implies a.s. $\mathscr{P}$.

(ii) Note that $f_\infty > 0$ a.s. $\mathscr{P}$.
Let $g_i(T) = T_i,\ i = 0,1,2,\ldots.$
Then $f_i(x) = g_i(T(x))f_\infty(x)$ a.s. $\mathscr{P}$.
By Theorem 2.2, T is sufficient for $P \in \mathscr{P}$.
Suppose that $S(X)$ is another sufficient statistic.
By Theorem 2.2, there are Borel functions h and $\tilde{g}_i$ such that

$$f_i(x) = \tilde{g}_i(S(x))h(x), \quad i = 0,1,2,\ldots.$$

Then

$$T_i(x) = \tilde{g}_i(S(x)) \Big/ \sum_{j=1}^{\infty} c_j \tilde{g}_j(S(x))$$

for x's satisfying $f_\infty(x) > 0$.
By Definition 2.5, T is minimal sufficient for $P \in \mathscr{P}$.
The proof for the case where f_∞ is replaced by f_0 is the same.

Proof (continued)

(iii) From Bahadur (1957), there is a minimal sufficient statistic $S(X)$.
The result follows if we can show that $T(X) = \psi(S(X))$ a.s. $\mathscr{P}$ for a measurable function ψ.
By Theorem 2.2, there are Borel functions h and g_P such that $f_P(x) = g_P(S(x))h(x)$ for all P.
Let $A = \{x : \ h(x) = 0\}$.
Then $P(A) = 0$ for all P.
For x and y such that $S(x) = S(y), \ x \notin A$ and $y \notin A$,

$$f_P(x) = g_P(S(x))h(x) = g_P(S(y))h(x) = f_P(y)h(x)/h(y)$$

for all P.
Hence $T(x) = T(y)$.
This shows that there is a function ψ such that $T(x) = \psi(S(x))$ except for $x \in A$.
It remains to show that ψ is measurable.
Since S is minimal sufficient, $g(T(X)) = S(X)$ a.s. $\mathscr{P}$ for a measurable function g. Hence g is one-to-one and $\psi = g^{-1}$.
By Theorem 3.9 in Parthasarathy (1967), ψ is measurable.

Example 2.14

Let $\mathscr{P} = \{f_\theta : \ \theta \in \Theta\}$ be an exponential family with p.d.f.'s

$$f_\theta(x) = \exp\{[\eta(\theta)]^\tau T(x) - \xi(\theta)\} h(x).$$

Suppose that there exists $\Theta_0 = \{\theta_0, \theta_1, \ldots, \theta_p\} \subset \Theta$ such that the vectors $\eta_i = \eta(\theta_i) - \eta(\theta_0)$, $i = 1, \ldots, p$, are linearly independent in $\mathscr{R}^p$.
(This is true if the family is of full rank).
We have shown that $T(X)$ is sufficient for $\theta \in \Theta$.
We now show that T is in fact minimal sufficient for $\theta \in \Theta$.
Let $\mathscr{P}_0 = \{f_\theta : \ \theta \in \Theta_0\}$.
Note that the set $\{x : \ f_\theta(x) > 0\}$ does not depend on θ.
It follows from Theorem 2.3(ii) with $f_\infty = f_{\theta_0}$ that

$$S(X) = \Big(\exp\{\eta_1^\tau T(x) - \xi_1\}, \ldots, \exp\{\eta_p^\tau T(x) - \xi_p\} \Big)$$

is minimal sufficient for $\theta \in \Theta_0$.
Since η_i's are linearly independent, there is a one-to-one measurable function ψ such that $T(X) = \psi(S(X))$ a.s. $\mathscr{P}_0$.
Hence, T is minimal sufficient for $\theta \in \Theta_0$.
It is easy to see that a.s. $\mathscr{P}_0$ implies a.s. $\mathscr{P}$.
Thus, by Theorem 2.3(i), T is minimal sufficient for $\theta \in \Theta$.

Example 2.14 (continued)

We now apply Theorem 2.3(iii) to obtain the same result.

For any θ,

$$f_\theta(x)/f_\theta(y) = \exp\{\eta^\tau(\theta)(T(x) - T(y))\}h(x)/h(y)$$

If, for any θ,

$$f_\theta(x)/f_\theta(y) = \phi(x,y)$$

then

$$\eta^\tau(\theta)(T(x) - T(y)) = \log(\phi(x,y)) + \log(h(y)/h(x))$$

Then

$$[\eta(\theta_i) - \eta(\theta_0)]^\tau(T(x) - T(y)) = 0 \qquad i = 1,...,p.$$

Since $\eta(\theta_i) - \eta(\theta_0)$, $i = 1,...,p$, are linearly independent, we have $T(x) = T(y)$.
By Theorem 2.3(iii), T is minimal sufficient

The result in Example 2.13 can be proved by using Theorem 2.3(iii).

Application to curved normal family

Let $X_1,...,X_n$ be i.i.d. from $N(\mu,\mu^2)$, $\mu \in \mathscr{R}$, $\mu \neq 0$.
It can be shown that

$$\eta(\theta) = \left(\frac{n\mu}{\sigma^2}, -\frac{n-1}{2\sigma^2}\right) = \left(\frac{n}{\mu}, -\frac{n-1}{2\mu^2}\right)$$

Points $\theta_0 = (1,1)$, $\theta_1 = (-1,1)$, and $\theta_2 = (1/2,1/2)$ are in the parameter space, and

$$\eta(\theta_1)-\eta(\theta_0) = \begin{pmatrix} -n \\ -(n-1)/2 \end{pmatrix} - \begin{pmatrix} n \\ -(n-1)/2 \end{pmatrix} = \begin{pmatrix} -2n \\ 0 \end{pmatrix}$$

$$\eta(\theta_2)-\eta(\theta_0) = \begin{pmatrix} n/2 \\ -2(n-1) \end{pmatrix} - \begin{pmatrix} n \\ -(n-1)/2 \end{pmatrix} = \begin{pmatrix} -n/2 \\ -3(n-1)/2 \end{pmatrix}$$

If

$$c\begin{pmatrix} -2n \\ 0 \end{pmatrix} + d\begin{pmatrix} -n/2 \\ -3(n-1)/2 \end{pmatrix} = 0$$

we must have $d = 0$ and then $c = 0$.
The vectors $\eta(\theta_1)-\eta(\theta_0)$ and $\eta(\theta_2)-\eta(\theta_0)$ are linearly independent.
Therefore, $T = (\bar{X}, S^2)$ is minimal sufficient for $\theta = \mu$.

Remarks

- The sufficiency (and minimal sufficiency) depends on the postulated family $\mathscr{P}$ of populations (statistical models).
- It may not be a useful concept if the proposed statistical model is wrong or at least one has some doubts about the correctness of the proposed model.
- From the examples in this section and some exercises in §2.6, one can find that for a wide variety of models, statistics such as the sample mean $\bar{X}$, the sample variance S^2, $(X_{(1)}, X_{(n)})$ in Example 2.11, and the order statistics in Example 2.9 are sufficient.
- Thus, using these statistics for data reduction and summarization does not lose any information when the true model is one of those models but we do not know exactly which model is correct.
- A minimal sufficient statistic is not always the “simplest sufficient statistic".
- For example, if $\bar{X}$ is minimal sufficient, then so is $(\bar{X}, \exp\{\bar{X}\})$.