Lecture 23: Minimal sufficiency

Maximal reduction without loss of information

- There are many sufficient statistics for a given problem.
- In fact, $X$ (the whole data set) is sufficient.
- If $T$ is a sufficient statistic and $T = \psi(S)$, where $\psi$ is a function and $S$ is another statistic, then $S$ is sufficient.
  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 $\theta$, where $m$ is any fixed integer between 1 and $n$.
- If $T$ is sufficient and $T = \psi(S)$ with a measurable function $\psi$ that is not one-to-one, then $T$ is more useful than $S$, since $T$ provides a further reduction of the data without loss of information.
- Is there a sufficient statistics that provides “maximal" reduction of the data?

Definition 6.2.11 (minimal sufficiency)

A sufficient statistic $T(X)$ is a minimal sufficient statistic if, for any other sufficient statistic $U(X)$, $T(x)$ is a function of $U(x)$. 
\( T(x) \) is a function of \( U(x) \) iff \( U(x) = U(y) \) implies that \( T(x) = T(y) \) for any pair \((x, y)\).

In terms of the partitions of the range of the data set, if \( T(x) \) is a function of \( U(x) \), then

\[
\{ x : U(x) = t \} \subset \{ x : T(x) = t \}
\]

Thus, a minimal sufficient statistic achieves the greatest possible data reduction as a sufficient statistic.

If both \( T \) and \( S \) are minimal sufficient statistics, then by definition there is a one-to-one function \( \psi \) such that \( T = \psi(S) \); hence, the minimal sufficient statistic is unique in the sense that two statistics that are functions of each other can be treated as one statistic.

For example, if \( T \) is minimal sufficient, then so is \((T, e^T)\), but no one is going to use \((T, e^T)\).

If the range of \( X \) is \( \mathbb{R}^k \), then there exists a minimal sufficient statistic.

**Example 6.2.15.**

Let \( X_1, \ldots, X_n \) be iid from the \textit{uniform}(\( \theta, \theta + 1 \)) distribution, where \( \theta \in \mathbb{R} \) is an unknown parameter.
The joint pdf of \((X_1, \ldots, X_n)\) is
\[
\prod_{i=1}^{n} l(\{\theta < x_i < \theta + 1\}) = l(\{x_{(n)} - 1 < \theta < x_{(1)}\}), \quad (x_1, \ldots, x_n) \in \mathbb{R}^n,
\]
where \(x_{(i)}\) denotes the \(i\)th ordered value of \(x_1, \ldots, x_n\).

By the factorization theorem, \(T = (X_{(1)}, X_{(n)})\) is sufficient for \(\theta\).

Note that
\[
x_{(1)} = \sup\{\theta : f_\theta(x) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : f_\theta(x) > 0\}.
\]

If \(S(X)\) is a statistic sufficient for \(\theta\), then by the factorization theorem, there are functions \(h\) and \(g_\theta\) such that \(f_\theta(x) = g_\theta(S(x))h(x)\).

For \(x\) with \(h(x) > 0\),
\[
x_{(1)} = \sup\{\theta : g_\theta(S(x)) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : g_\theta(S(x)) > 0\}.
\]

Hence, there is a function \(\psi\) such that \(T(x) = \psi(S(x))\) when \(h(x) > 0\).

Since \(P(h(X) > 0) = 1\), we conclude that \(T\) is minimal sufficient.

In this example, the dimension of a minimal sufficient statistic is larger than the dimension of \(\theta\).
Finding a minimal sufficient statistic by definition is not convenient. The next theorem is a useful tool.

**Theorem 6.2.13.**

Let $f_\theta(x)$ be the pmf or pdf of $X$. Suppose that $T(x)$ is sufficient for $\theta$ and that, for every pair $x$ and $y$ with at least one of $f_\theta(x)$ and $f_\theta(y)$ is not 0, $f_\theta(x)/f_\theta(y)$ does not depend on $\theta$ implies $T(x) = T(y)$. (Here, we define $a/0 = \infty$ for any $a > 0$.) Then $T(X)$ is minimal sufficient for $\theta$.

**Proof.**

Let $U(X)$ be another sufficient statistic. By the factorization theorem, there are functions $h$ and $g_\theta$ such that $f_\theta(x) = g_\theta(U(x))h(x)$ for all $x$ and $\theta$.

For $x$ and $y$ such that at least one of $f_\theta(x)$ and $f_\theta(y)$ is not 0, i.e., at least one of $h(x)$ and $h(y)$ is not 0, if $U(x) = U(y)$, then

$$
\frac{f_\theta(x)}{f_\theta(y)} = \frac{g_\theta(U(x))h(x)}{g_\theta(U(y))h(y)} = \frac{h(x)}{h(y)}
$$

which does not depend on $\theta$.  

By the assumption of the theorem, $T(x) = T(y)$.
This shows that there is a function $\psi$ such that $T(x) = \psi(S(x))$
Hence, $T$ is minimal sufficient for $\theta$

**Example 6.2.15.**

We re-visit this example and show how to apply Theorem 6.2.13.
If $X_1, ..., X_n$ are iid form the $\text{uniform}(\theta, \theta + 1)$ distribution, then the joint pdf is

$$f_\theta(x) = I(\{x_{(n)} - 1 < \theta < x_{(1)}\}), \quad x = (x_1, ..., x_n) \in \mathbb{R}^n,$$

Let $x$ and $y$ be two data points. Each of $f_\theta(x)$ and $f_\theta(y)$ only takes two possible values, 0 and 1.

Let $A_x = (x_{(n)} - 1, x_{(1)})$ and $A_y = (y_{(n)} - 1, y_{(1)})$.
Then

$$\frac{f_\theta(x)}{f_\theta(y)} = \begin{cases} 
0 & \theta \notin A_x, \theta \in A_y \\
1 & \theta \in A_x, \theta \in A_y \\
\infty & \theta \in A_x, \theta \notin A_y 
\end{cases}$$

This depends on $\theta$ unless $A_x = A_y$. 
Thus, if the ratio does not depend on $\theta$, we must have $x(1) = y(1)$ and $x(n) = y(n)$.

Therefore, by Theorem 6.2.13, $(X(1), X(n))$ is minimal sufficient for $\theta$.

**Example 6.2.14.**

Let $X_1, \ldots, X_n$ be iid from $N(\mu, \sigma^2)$ with unknown $\mu \in \mathbb{R}$ and $\sigma > 0$.

Earlier, we showed that $T = (\bar{X}, S^2)$ is sufficient for $\theta = (\mu, \sigma^2)$.

Is $T$ minimal sufficient?

Let $f_\theta(x)$ be the joint pdf of the sample, $x$ and $y$ be two sample points.

Then

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

$$
= \exp\left[-\frac{n(\bar{x}^2 - \bar{y}^2)}{2\sigma^2} + \frac{2n\mu(\bar{x} - \bar{y}) - (n-1)(s_x^2 - s_y^2)}{2\sigma^2}\right]
$$

where $\bar{x}$ and $s_x^2$ are the sample mean and variance based on the sample point $x$, and $\bar{y}$ and $s_y^2$ are the sample mean and variance based on the sample point $y$. 
This ratio does not depend on $\theta = (\mu, \sigma^2)$ iff $\bar{x} = \bar{y}$ and $s_x^2 = s_y^2$.

By Theorem 6.2.13, $T = (\bar{X}, S^2)$ is minimal sufficient for $\theta = (\mu, \sigma^2)$.

Like the concept of sufficiency, minimal sufficiency also depends on the parameters we are interested in.

In Example 6.2.14, let us consider two sub-families.

- Suppose that it is known that $\sigma = 1$.
  Then $T = (\bar{X}, S^2)$ is still sufficient but not minimal sufficient for $\mu$.
  In fact, using Theorem 6.2.13 we can show that $\bar{X}$ is minimal sufficient for $\mu$ when $\sigma = 1$ and, $T = (\bar{X}, S^2)$ is not a function of $\bar{X}$ (why?).

- Suppose that it is known that $\mu^2 = \sigma^2$ so that we only have one unknown parameter $\mu \in \mathbb{R}, \mu \neq 0$. (This is called a curved exponential family.)
  Is $T = (\bar{X}, S^2)$ minimal sufficient for $\theta = \mu$?
  Note that $T$ is certainly sufficient for $\theta$.
  The answer can be found as a special case of the following result.
Minimal sufficient statistics in exponential families

Let $X_1, \ldots, X_n$ be iid from an exponential family with pdf

$$f_\theta(x) = h(x)c(\theta)\exp(\eta(\theta)'T(x)), \quad \theta \in \Theta$$

where $\eta(\theta)' = (w_1(\theta), \ldots, w_k(\theta))$ and $T(x)' = (t_1(x), \ldots, t_k(x))$.

Suppose that there exists $\{\theta_0, \theta_1, \ldots, \theta_k\} \subset \Theta$ such that the vectors $\eta_i = \eta(\theta_i) - \eta(\theta_0), \ i = 1, \ldots, k$, are linearly independent in $\mathbb{R}^k$.

(This is true if the family is of full rank).

We have shown that $T(X)$ is sufficient for $\theta$.

We now show that $T$ is in fact minimal sufficient for $\theta$.

Note that we can focus on the points $x$ at which $h(x) > 0$.

For any $\theta$,

$$\frac{f_\theta(x)}{f_\theta(y)} = \exp(\eta(\theta)'[T(x) - T(y)]) \frac{h(x)}{h(y)}$$

If this ratio equals $\phi(x, y)$ not depending on $\theta$, then

$$\eta(\theta)'[T(x) - T(y)] = \log(\phi(x, y)) + \log(h(y)/h(x))$$

which implies

$$[\eta(\theta_i) - \eta(\theta_0)]'[T(x) - T(y)] = 0 \quad i = 1, \ldots, k.$$
Since $\eta(\theta_i) - \eta(\theta_0)$, $i = 1, \ldots, k$, are linearly independent, we must have $T(x) = T(y)$.

By Theorem 6.2.13, $T$ is minimal sufficient for $\theta$.

**Application to curved normal family**

We can now show that $T = (\bar{X}, S^2)$ minimal sufficient for $\theta = \mu$ when the sample is a random sample from $N(\mu, \mu^2)$, $\mu \in \mathbb{R}$, $\mu \neq 0$.

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

$$\eta(\theta_1) - \eta(\theta_0) = \left( \begin{array}{c} -n \\ -(n-1)/2 \end{array} \right) - \left( \begin{array}{c} n \\ -(n-1)/2 \end{array} \right) = \left( \begin{array}{c} -2n \\ 0 \end{array} \right)$$

$$\eta(\theta_2) - \eta(\theta_0) = \left( \begin{array}{c} n/2 \\ -2(n-1) \end{array} \right) - \left( \begin{array}{c} n \\ -(n-1)/2 \end{array} \right) = \left( \begin{array}{c} -n/2 \\ -3(n-1)/2 \end{array} \right)$$
Are these two vectors linearly independent? If there are $c$ and $d$ such that
\[ c[\eta(\theta_1) - \eta(\theta_0)] + d[\eta(\theta_2) - \eta(\theta_0)] = 0 \]
i.e.,
\[ c \begin{pmatrix} -2n \\ 0 \end{pmatrix} + d \begin{pmatrix} -n/2 \\ -3(n-1)/2 \end{pmatrix} = 0 \]
Then, we must have $d = 0$ and then $c = 0$. That means vectors $\eta(\theta_1) - \eta(\theta_0)$ and $\eta(\theta_2) - \eta(\theta_0)$ are linearly independent and the condition for the result of exponential family is satisfied. Therefore, $T = (\bar{X}, S^2)$ is minimal sufficient for $\theta = \mu$.

**Example.**

Let $(X_1, ..., X_n)$ be a random sample from $Cauchy(\mu, \sigma)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are unknown parameters. The joint pdf of $(X_1, ..., X_n)$ is
\[
 f_{\mu, \sigma}(x) = \frac{\sigma^n}{\pi^n} \prod_{i=1}^{n} \frac{1}{\sigma^2 + (x_i - \mu)^2}, \quad x = (x_1, ..., x_n) \in \mathbb{R}^n.
\]
For any \( x = (x_1, \ldots, x_n) \) and \( y = (y_1, \ldots, y_n) \), if
\[
f_{\mu, \sigma}(x) = \psi(x, y)f_{\mu, \sigma}(y)
\]
holds for any \( \mu \) and \( \sigma \), where \( \psi \) does not depend on \((\mu, \sigma)\), then
\[
\prod_{i=1}^{n} \left[ 1 + (y_i - \mu)^2 \right] = \psi(x, y) \prod_{i=1}^{n} \left[ 1 + (x_i - \mu)^2 \right] \quad \mu \in \mathbb{R}
\]
Both sides of the above identity are polynomials of degree \( 2n \) in \( \mu \). Comparison of the coefficients to the highest terms gives \( \psi(x, y) = 1 \) and hence
\[
\prod_{i=1}^{n} \left[ 1 + (y_i - \mu)^2 \right] = \prod_{i=1}^{n} \left[ 1 + (x_i - \mu)^2 \right] \quad \mu \in \mathbb{R}
\]
As a polynomial of \( \mu \), the left-hand side of the above identity has \( 2n \) complex roots \( x_i \pm i, \ i = 1, \ldots, n \), while the right-hand side of the above identity has \( 2n \) complex roots \( y_i \pm i, \ i = 1, \ldots, n \).
Since the two sets of roots must agree, the ordered values of \( x_i \)’s are the same as the ordered values of \( y_i \)’s.
By Theorem 6.2.13, the order statistics of \( X_1, \ldots, X_n \) is minimal sufficient for \((\mu, \sigma)\).
The following result may be useful in establishing minimal sufficiency outside of the exponential family.

**Theorem 6.6.5.**

Let \( f_0, f_1, f_2, \ldots \) be a sequence of pdf’s or pmf’s on the range of \( X \), all have the same support.

a. The statistic

\[
T(X) = \left( \frac{f_1(X)}{f_0(X)}, \frac{f_2(X)}{f_0(X)}, \ldots \right)
\]

is minimal sufficient for the family \( \{f_0, f_1, f_2, \ldots\} \) of populations for \( X \).

b. If \( \mathcal{P} \) is a family of pdf’s or pmf’s for \( X \) with common support, and

(i) \( f_i \in \mathcal{P}, \ i = 0, 1, 2, \ldots \)

(ii) \( T(X) \) is sufficient for \( \mathcal{P} \),

then \( T \) is minimal sufficient for \( \mathcal{P} \).

**Proof of part a.**

Let \( g_i(T) = T_i \), the \( i \)th component of \( T \), \( i = 0, 1, 2, \ldots \), and let \( g_0(T) = 1 \).

Then \( f_i(x) = g_i(T(x))f_0(x), \ i = 0, 1, 2, \ldots \)

By the factorization theorem, \( T \) is sufficient for \( \mathcal{P}_0 = \{f_0, f_1, f_2, \ldots\} \).
Suppose that $S(X)$ is another sufficient statistic.

By the factorization theorem, there are Borel functions $h$ and $\tilde{g}_i$ such that

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

Then

$$T_i(x) = \frac{f_i(x)}{f_0(x)} = \frac{\tilde{g}_i(S(x))}{\tilde{g}_0(S(x))}, \quad i = 0, 1, 2, ...$$

Thus, $T$ is a function of $S$.

By definition, $T$ is minimal sufficient for $\mathcal{P}_0$.

Proof of part b.

The two additional conditions are $\mathcal{P}_0 \subset \mathcal{P}$ and $T$ is sufficient for $\mathcal{P}$.

Let $S$ be any sufficient statistic for $\mathcal{P}$.

Then $S$ is also sufficient for $\mathcal{P}_0$.

From part a, $T$ is minimal sufficient for $\mathcal{P}_0$ so there is a function $\psi$ such that $T = \psi(S)$.

By definition, $T$ is minimal sufficient for $\mathcal{P}$ since $T$ is sufficient for $\mathcal{P}$. 
In fact, the result in Theorem 6.6.5 still holds if the common support condition is replaced by that the support of \( f_0 \) contains the support of any pdf or pmf in \( \mathcal{P} \).

In most applications, we can choose \( \mathcal{P}_0 \) containing finitely many elements.

**Example**

Let \( X_1, \ldots, X_n \) be a random sample from double-exponential(\( \mu, 1 \)), \( \mu \in \mathbb{R} \), i.e., the joint pdf of \( X \) is

\[
f_\mu(x) = \frac{1}{2^n} \exp \left( - \sum_{i=1}^{n} |x_i - \mu| \right)
\]

Consider \( \mathcal{P}_0 = \{ f_0, f_1, \ldots, f_n \} \), where each \( f_j \) is the pdf with \( \mu = j \), \( j = 0, 1, \ldots, n \).

By Theorem 6.6.5a, a minimal sufficient statistic for \( \mathcal{P}_0 \) is

\[
T = \left( \exp \left( \sum_{i=1}^{n} |X_i| - \sum_{i=1}^{n} |X_i - j| \right), \ j = 1, \ldots, n \right)
\]

which is equivalent to
\[ U = \left( \sum_{i=1}^{n} |X_i| - \sum_{i=1}^{n} |X_i - j|, \ j = 1, \ldots, n \right) \]

We can further show that \( U \) is equivalent to \( S \), the set of order statistics.
Since \( S \) is sufficient, by Theorem 6.6.5b, \( S, U, \) and \( T \) are all minimal sufficient.

**Definition 6.6.3 (Necessity).**
A statistic is said to be necessary if it can be written as a function of every sufficient statistic.

\( T \) is necessary if, for every sufficient \( S \), there is a function \( \psi \) such that \( T = \psi(S) \).

From the definitions of the necessity and minimal sufficiency, we can reach the following conclusion.

**Theorem 6.6.4.**
A statistic is minimal sufficient if and only if it is necessary and sufficient.