Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

Lecture 2 - Sufficiency, Factorization Theorem, Minimal sufficiency

Summarizing the data without loosing information :

- is it possible? Yes sometimes : sufficiency
- How can we do that ? Using sufficient statistics
- Is it important ? Yes : often - leads to simplified and better estimates

Sufficient statistics

Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$.

Definition (Sufficiency)

A statistic $T(X_1, \ldots, X_n)$ is a function of the data that does not depend on unknown parameters.

A statistic $T(X_1, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$, given T, does not depend on θ. That is,

$$f(x \mid t, \theta) = f(x \mid t)$$

Comment The definition says that a sufficient statistic T contains all the information there is in the sample about θ.

Definition (Sufficiency)

A statistic $T(X_1, \ldots, X_n)$ is a function of the data that does not depend on unknown parameters.

A statistic $T(X_1, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$, given T, does not depend on θ. That is,

$$f(x \mid t, \theta) = f(x \mid t)$$

What does this even mean?
It means that for any function g the map

$$\theta \mapsto \mathbb{E}_\theta[g(X)|T = t]$$

is constant.

Example 7

n independent trials where the probability of success is p.
Let $X_1, \ldots, X_n$ be indicator variables which are 1 or 0 depending if the trial is a success or failure.
Let $T = \sum_{i=1}^n X_i$. The conditional distribution of $X_1, \ldots, X_n$ given $T = t$ is

$$
\begin{aligned}
g(x_1, \ldots, x_n \mid t, p) = \frac{f(x_1, \ldots, x_n, t \mid p)}{h(t \mid p)} &= \frac{\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} \\
&= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \\
&= \binom{n}{t}^{-1},
\end{aligned}
$$

not depending on p, so T is sufficient for p.
Comment Makes sense, since no information in the order.

Theorem 1 : Factorization Criterion

Theorem

$T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if there exist two non-negative functions f_1, h such that the likelihood function $L(\theta; x)$ can be written

$$L(\theta; x) = f_1[t(x_1, \ldots, x_n); \theta] h[x_1, \ldots, x_n] = f_1[t; \theta] h[x],$$

where f_1 depends only on the sample through T, and h does not depend on θ.

Proof - for discrete random variables

1. Assume that T is sufficient, then the distribution of the sample is

$$L(\theta; x) = f(x|\theta) = f(x, t|\theta)\mathbb{1}_{T(x)=t} = f(x \mid t, \theta) f_T(t \mid \theta)$$

T is sufficient which implies

$$f(x \mid t, \theta) = f(x \mid t) := h(x)$$

$f_1(t \mid \theta)$ depends on x through $t(x)$ only so

$$L(\theta; x) = f(x \mid t) f_T(t \mid \theta)$$

We set $L(\theta; x) = h(x) f_1(t; \theta)$, where $f_1 \equiv f_T(t|\theta)$, $h \equiv f(x \mid t)$.

2. Suppose $L(\theta; x) = f(x \mid \theta) = f_1[t; \theta]h[x]$. Then

$$f_T(t \mid \theta) = \sum_{\{x:T(x)=t\}} f(x, t \mid \theta) = \sum_{\{x:T(x)=t\}} L(\theta; x) = f_1[t; \theta] \sum_{\{x:T(x)=t\}} h(x).$$

Thus

$$f(x \mid t, \theta) = \frac{f(x, t \mid \theta)}{f_T(t \mid \theta)} = \frac{L(\theta; x)}{f_T(t \mid \theta)} = \frac{h[x]}{\sum_{\{x':T(x')=t\}} h(x')},$$

not depending on θ. (f_1 cancels out in numerator and denominator.)

Minimal sufficiency

How much can we reduce the data without loosing information? Is there a minimal sufficient statistic?

Example 7 (cont.) Consider $n = 3$ Bernoulli trials

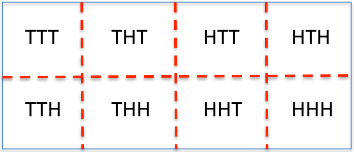
1. $T_1(X) = (X_1, X_2, X_3)$ (the individual sequences of trials)
2. $T_2(X) = (X_1, \sum_{i=1}^{3} X_i)$ (the 1st random variable and the total sum).
3. $T_4(X) = I(T_3(X) = 0)$ (I is indicator function; Exercise Prove T_4 not sufficient)

Definition (Minimality)

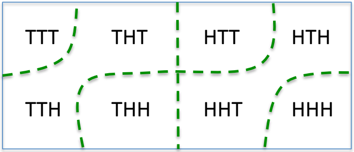
A statistic is minimal sufficient if it can be expressed as a function of every other sufficient statistic.

Minimal sufficiency and partitions of the sample space

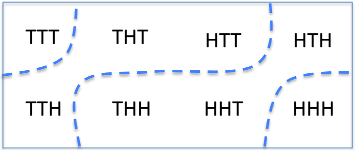
- Intuitively, a minimal sufficient statistic most efficiently captures all possible information about the parameter θ.
- Any statistic $T(X)$ partitions the sample space into subsets and in each subset $T(X)$ has constant value.
- Minimal sufficient statistics correspond to the coarsest possible partition of the sample space.
- In the example of $n = 3$ Bernoulli trials consider the following 4 statistics and the partitions they induce.



$$T_1(X) = (X_1, X_2, X_3)$$



$$T_2(X) = \left(X_1, \sum_{i=1}^{3} X_i \right)$$



$$T_3(X) = \sum_{i=1}^{3} X_i$$

$$T_4(X) = I\left(T_3(X) = 0\right)$$

Lemma 1 : Lehmann-Scheffé partitions

Theorem

If a statistic $T(X)$ *satisfies*

$$T(x) = T(y) \quad \Leftrightarrow \frac{L(\theta; y)}{L(\theta; x)} = \frac{f(y \mid \theta)}{f(x \mid \theta)} = m(x, y),$$

then it is minimal sufficient .

Comment This Lemma tells us how to define partitions that correspond to minimal sufficient statistics. It says that ratios of likelihoods of two values x and y in the same partition (and hence same statistic value) should not depend on θ.

Proof (for discrete RVs)

1. Sufficiency.

Suppose T is such a statistic

$$g(x|t,\theta) = \frac{f(x \mid \theta)}{f(t \mid \theta)} = \frac{f(x \mid \theta)}{\sum_{y\in\tau} f(y \mid \theta)}, \quad \tau = \{y : T(y) = t\}$$

$$= \frac{f(x \mid \theta)}{\sum_{y\in\tau} f(x \mid \theta) m(x,y)}$$

$$= \Big[\sum_{y\in\tau} m(x,y)\Big]^{-1}$$

which does not depend on θ. Hence T is sufficient.

Proof (for discrete RVs)

1. Sufficiency.
Suppose T is such a statistic

$$
\begin{aligned}
g(x|t,\theta) = \frac{f(x \mid \theta)}{f(t \mid \theta)} &= \frac{f(x \mid \theta)}{\sum_{y\in\tau} f(y \mid \theta)}, \quad \tau = \{y : T(y) = t\} \\
&= \frac{f(x \mid \theta)}{\sum_{y\in\tau} f(x \mid \theta) m(x,y)} \\
&= \Big[\sum_{y\in\tau} m(x,y)\Big]^{-1}
\end{aligned}
$$

which does not depend on θ. Hence T is sufficient.

2. Minimal sufficiency.

Now suppose U is any other sufficient statistic and that $U(x) = U(y)$ for some pair of values (x, y). Since U is sufficient we have for all pair (x, y) s.t. $U(x) = U(y)$

$$\frac{L(\theta; y)}{L(\theta; x)} = \frac{f_1[u(y); \theta]h[y]}{f_1[u(x); \theta]h[x]} = \frac{h[y]}{h[x]}$$

which does not depend on θ. Hence $T(x) = T(y)$ and T is a function of U.

Example 7 (cont.) : Minimal sufficiency

n Bernoulli trials with $T = \sum_{i=1}^{n} X_i$.

$$\frac{p^{T(x)}(1-p)^{n-T(x)}}{p^{T(y)}(1-p)^{n-T(y)}} = \left(\frac{p}{(1-p)}\right)^{T(x)-T(y)} = m(x,y) \Leftrightarrow T(x) = T(y)$$

Sufficiency in an exponential family

$$L(\theta; x) = e^{A(\theta)^t B(x) - D(\theta) + C(x)}$$

Then $B(X)$ is a sufficient statistics for $\boldsymbol{\phi} = A(\theta)$ and for θ since

$$f_1(B(x); \theta) = e^{A(\theta)^t B(x) - D(\theta)}, \quad h(x) = e^{C(x)}$$

For a sample $X_1, \ldots, X_n$ i.i.d. from the exponential family $T(X^n) = \sum_{i=1}^n B(X_i)$ is a sufficient statistics.

$$L(\boldsymbol{\phi}; x) \;=\; \prod_{i=1}^n f(x_i; \boldsymbol{\phi}) = \exp\left\{ \sum_{j=1}^k \phi_j \left(\sum_{i=1}^n B_j(x_i) \right) - nD(\boldsymbol{\phi}) + \sum_{i=1}^n C(x_i$$

Exponential family again

What do we need to have a minimal sufficient statistics?

Minimal sufficiency in an exponential family : canonical parametrisation

Theorem

For a sample $X_1, \ldots, X_n$ i.i.d. from a regular exponential family (Φ open), then

- *The distribution of $T(x)$ belongs to a k-parameter exponential family.*
- *The statistic $T(x); = (\sum_{i=1}^n B_1(x_i), \ldots, \sum_{i=1}^n B_k(x_i))$ is minimal sufficient.*

Proof

Set $t_j = \sum_{i=1}^n B_j(x_i)$ and $C(x) = \sum_{i=1}^n C(x_i)$, then

$$L(\boldsymbol{\phi}; x) = \exp\left\{ \sum_{j=1}^k \phi_j t_j - nD(\boldsymbol{\phi}) + C(x) \right\}.$$

and

$$\frac{L(\boldsymbol{\phi}; x)}{L(\boldsymbol{\phi}; y)} = \underbrace{\exp\left(\sum_{j=1}^k \phi_j [t_j(x) - t_j(y)] \right)}_{= cst \Leftrightarrow t(x) = t(y)} \exp(C(x) - C(y))$$

because Φ open.

Sufficiency in an exponential family : general case

If $\theta \in \Theta \subset \mathbb{R}^k$ and $\Phi(\Theta)$ contains an open rectangle (full rank family) then $T(x) = (\sum_{i=1}^n B_1(x_i), \ldots, \sum_{i=1}^n B_k(x_i))$ is minimal sufficient for θ.

$$\frac{L(\theta; x)}{L(\theta; y)} = \underbrace{\exp\left(\sum_{j=1}^k \phi_j(\theta)[t_j(x) - t_j(y)]\right)}_{=cst \Leftrightarrow t(x)=t(y)} \exp(C(x) - C(y))$$

same story

Sufficiency as a rare feature

Nice property for reducing the data in a low dimensional transform without loosing information

How frequent is it within the parametric families ?

Very rare: Only exponential families or some families whose support depend on the parameter [Pitman-Koopman-Darmois theorem]