The Sufficiency Principle

Maura Mezzetti

Department of Economics and Finance
Università Tor Vergata

Outline

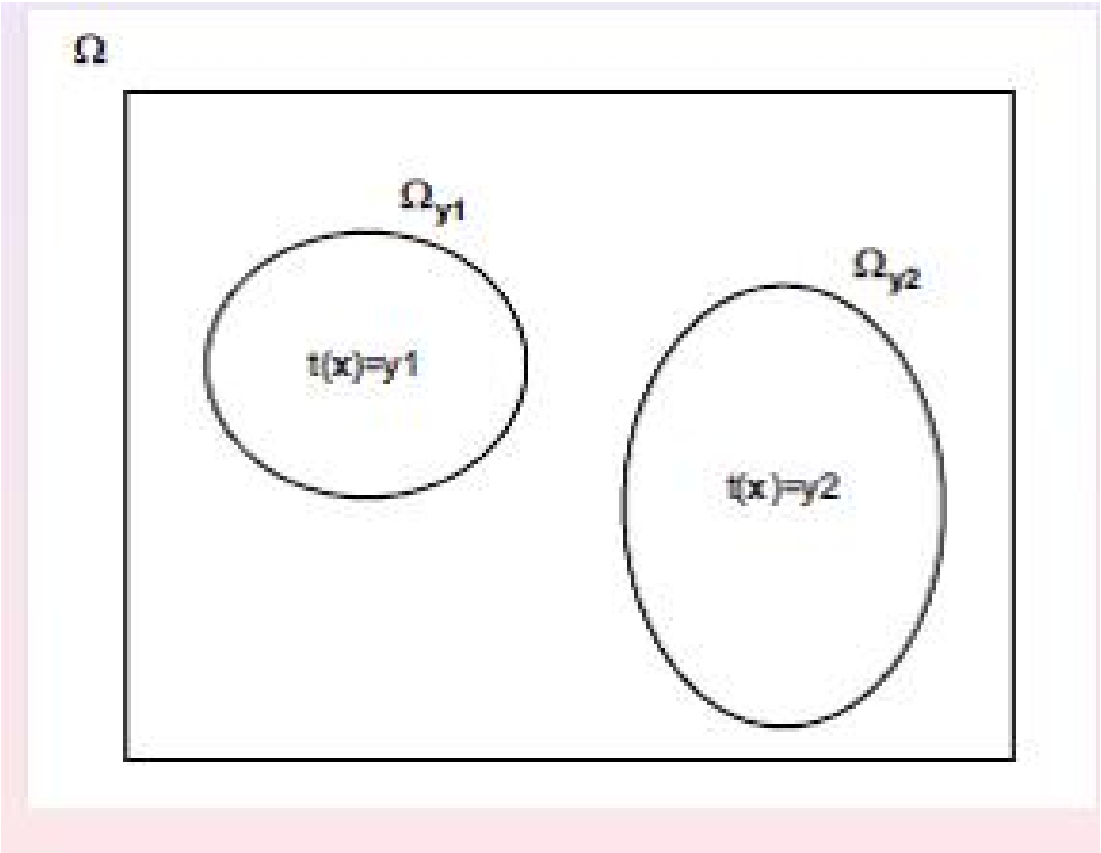
Principle of Data Reduction

"*. . . we are suffering from a plethora of surmise, conjecture and hypothesis. The difficulty is to detach the framework of fact - of absolutely undeniable fact - from the embellishments of theorists and reporters.*"

Sherlock Holmes
Silver Blaze

Sufficiency

Any sample statistic $Y = t(\mathbf{X})$ provides a partition of the sample space into the subsets Ω_y, where $\Omega_y = \{\mathbf{x} : t(\mathbf{x}) = y\}$



Sufficiency

- If two observed samples, say x_1 and x_2 , belong Ω_y, we have that $t(x_1) = t(x_2)$ and they lead to the same inferential conclusion on θ, even if the two samples are different.
- Any sample statistic defines a form of data reduction. The problem is that using a sample statistic we can loose information on θ. It is then important to use a sample statistic that captures all the information about θ contained in the sample, in the sense that it is equivalent to the sample about the information on θ.

Principles of data reduction

Remarks

- If the sample size n is large the information in $\mathbf{x} = (x_1, \ldots, x_n)$ on θ may be hard to understand
- The information in a sample can be summarized using statistics (such as sample mean...)
- Any sample statistic $T(\mathbf{x})$ defines a form of data reduction or data summary
- The value $t = T(\mathbf{x})$ instead of the entire sample $\mathbf{x}$ is used to make inference on θ

The problem is that using a sample statistic we can loose information on θ. It is then important to use a sample statistic that captures all the information about θ contained in the sample, in the sense that it is equivalent to the sample about the information on θ.

Sufficiency

SUFFICIENCY PRINCIPLE: If $T(X)$ is a sufficient statistic for θ, then any inference about θ should depend on the sample X only through the value $T(X)$. That is, if x and y are two points such that $T(x) = T(y)$, then any inference about θ should be the same whether $X = x$ or $X = y$ is observed.

Definition A statistic $T(X)$ is a *sufficient statistic* for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

Sufficiency

Let $T(X)$ be a sufficient statistic. Consider the pair $(X, T(X))$. Obviously, $(X, T(X))$ contains the same information about θ as X alone, since $T(X)$ is a function of X. But if we know $T(X)$, then X itself has no values for us since it conditional distribution given $T(X)$ is independent of θ. Thus, by observing X (in addition to $T(X)$), we cannot say whether one particular value of parameter θ is more likely than another. Therefore, once we know $T(X)$ we can discard X completely.

Sufficiency

In other words, given the value t of a sufficient statistic T, conditionally there is no more information left in the original data regarding the unknown parameter θ. Put another way, we may think of X trying to tell us a story about θ, but once a sufficient summary T becomes available, the original story then becomes redundant. Observe that the whole data X is always sufficient for θ in this sense. But, we are aiming at a *shorter* summary statistic which has the same amount of information available in X. Thus, once we find a sufficient statistic T, we will focus only on the summary statistic T.

Example

Let $X = (X_1, \ldots, X_n)$ be a random sample from $N(\mu, \sigma^2)$. Suppose that σ^2 is known. Let consider $T(X) = \bar{X}_n$ and we know $\bar{X}_n \sim N(\mu, \sigma^2/n)$. Let's now calculate the conditional distribution of X given $T(X) = t$.

$$
\begin{aligned}
f_{X|T(X)}(x|T(X) = T(x)) &= \frac{f_X(x)}{f_T(T(x))} \\
&= \frac{\frac{exp\left(-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)\right)}{(2\pi)^{n/2}\sigma^n}}{\frac{exp(-n(\bar{x}_n-\mu)^2/(2\sigma^2))}{(2\pi)^{1/2}(\sigma/n)^{1/2}}} \\
&= \frac{exp\left(-\sum_{i=1}^n (x_i-\bar{x}_n)^2/(2\sigma^2)\right)}{(2\pi)^{(n-1)/2}\sigma^{n-1}/n^{1/2}}
\end{aligned}
$$

and this is independent on μ! $\bar{X}$ is a sufficient statistic for μ.

Example

Let X be a random sample from $Exp(\lambda)$, let's consider the statistic $T = I(X > 2)$.

$$\begin{aligned} P(X > 3 | T) &= \frac{P(X > 3 \cap T = 1)}{P(T = 1)} \\ &= \frac{P(X > 3 \cap X > 2)}{P(X > 2)} \\ &= \frac{P(X > 3)}{P(X > 2)} \\ &= \frac{exp(-3\lambda)}{exp(-2\lambda)} \\ &= exp(-\lambda) \end{aligned}$$

The statistics $T = I(X > 2)$ is NOT a sufficient statistics for λ.

Sufficiency

Theorem If $p(x|\theta)$ is the joint probability distribution function of X, and $q(t|\theta)$ is the probability distribution function of $T(X)$, then $T(X)$ is a sufficient statistic for θ if, and only if, for every x in the sample space the ratio $p(x|\theta)/q(T(x)|\theta)$ is constant as a function of θ.

Note: $p(x|\theta)$ indicates either discrete or continuous distribution (in this case $p(x|\theta) = f(x|\theta)$)

Sufficiency

Factorization Theorem Let $p(x|\theta)$ denote the joint probability distribution function of a sample X. A statistic $T(X)$ is a sufficient statistics for θ if and only if there exists a function $g(t|\theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ:

$$p(x|\theta) = g(T(x)|\theta)h(x).$$

Proof See Casella and Berger pag. 276

Factorization Theorem

Brief proof for more details see See Casella and Berger pag. 276
Let's consider only the discrete case

$$\begin{aligned} P_\theta(X = x) &= \sum_t P_\theta(X = x \cap T = t) \\ &= P_\theta(X = x \cap T = T(x)) \\ &= P_\theta(X = x | T = T(x)) P_\theta(T = T(x)) \end{aligned}$$

Since T is sufficient $P_\theta(X = x | T = T(x))$ is independent of θ and so it is equal to $h(x)$ and $P_\theta(X = x) = g(T(x)|\theta)h(x)$

Factorization Theorem

Brief proof for more details see See Casella and Berger pag. 276
On the other hand...

$$\begin{aligned} P_\theta(X = x | T = T(x)) &= \frac{P_\theta(X = x)}{P_\theta(T = T(x))} \\ &= \frac{g(T(x)|\theta)h(x)}{\sum_{T(y)=t} g(T(y)|\theta)h(y)} \end{aligned}$$

Since if $T(x) \neq t$ then $P_\theta(X = x | T = T(x)) = 0$, $g(T(y)|\theta)$ is a constant

$$P_\theta(X = x | T = T(x)) = \frac{h(x)}{\sum_{T(y)=t} h(y)}$$

does not depend on θ

Sufficiency and exponential family

- As a consequence of the factorization Theorem, we have that when a sample statistic, say $s(\mathbf{X})$, is a one-to-one function of a sufficient statistic for θ, also $s(\mathbf{X})$ is a sufficient statistic.
- When we sample from an exponential family, we have that

$$f(x|\theta) = h(x)c(\theta)exp\left(\sum_{i=1}^{k}\omega_i(\theta)t_i(x)\right)$$

then, form factorization theorem follows:
$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$ is a sufficient statistics for θ.

Examples Sufficient Statistics

Example: The distribution of the sample under the Bernoulli distribution may be expressed as:

$$f(\mathbf{x};\pi) = \pi^{\sum_i x_i}(1-\pi)^{n-\sum_i x_i} = g(t(\mathbf{x});\pi)h(\mathbf{x})$$

with $t(\mathbf{x}) = \sum_i x_i$, $g(t;\pi) = \pi^t(1-\pi)^{n-t}$, and $h(\mathbf{x}) = 1$. So, according to factorization theorem: $Y = \sum_{i=1}^n X_i$ is a sufficient statistics for π.

Examples Sufficient Statistics

Example: Let $X_1, \ldots, X_n$ be iid r.v. distributed as continuous uniform distribution on $[0, \theta]$. The probability distribution function of X_i for each i is:

$$f(x|\theta) = \begin{cases} \theta^{-1}, & 0 \leq x \leq \theta \\ 0, & \textit{otherwise} \end{cases}$$

Thus the joint probability distribution function of $X_1, \ldots, X_n$ is

$$f(\underline{x}|\theta) = \begin{cases} \theta^{-n}, & 0 \leq x_i \leq \theta, \textit{ for } i = 1, \ldots, n \\ 0, & \textit{otherwise} \end{cases}$$

Examples Sufficient Statistics

Example continued: Since for each i, $x_i \leq \theta$, we can write $\max_i x_i \leq \theta$, and let define $T(x) = max_i\, x_i$:

$$f(\underline{x}|\theta) = \theta^{-n} \prod_i I_{[0,\theta]}(x_i)$$

$$f(\underline{x}|\theta) = \theta^{-n} I_{[max(x_i),\infty]}(\theta)$$

and

$$g(t|\theta) = \begin{cases} \theta^{-n}, & t \leq \theta \\ 0, & otherwise \end{cases}$$

It can be easily verified that $f(x|\theta) = g(T(x)|\theta)h(x)$ for all x and all θ. So the statistics $T(\mathbf{X}) = max_i X_i$ is a sufficient statistics for θ,

Examples Sufficient Statistics: Pareto

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables distributed as a Pareto distribution with parameters α and x_{m}

$$f(x|\alpha, x_{\mathrm{m}}) = \frac{\alpha\, x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} \qquad \text{for } x \geq x_m.$$

Find a sufficient statistics for α,

Examples Sufficient Statistics: Pareto

$$
\begin{aligned}
f(x|\alpha, x_{\mathrm{m}}) &= \prod_{i=1}^{n} \frac{\alpha\, x_{\mathrm{m}}^{\alpha}}{x_i^{\alpha+1}} \\
&= \frac{\alpha^n\, x_{\mathrm{m}}^{n\alpha}}{\prod_{i=1}^{n} x_i^{\alpha+1}} \\
&= g(t, \alpha) h(x)
\end{aligned}
$$

where $t = \prod_{i=1}^{n} x_i$ $g(t, \alpha) = \alpha^n\, x_{\mathrm{m}}^{n\alpha}\, t^{-(\alpha+1)}$ and $h(x) = 1$. By the factorization theorem, $T(X) = \prod_{i=1}^{n} X_i$ is sufficient for α.

Examples Sufficient Statistics

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables distributed as a Beta distribution with parameters α and $\beta = 2$

$$f(x; \alpha) = \alpha(\alpha + 1)x^{\alpha+1}(1 - x) \qquad 0 < x < 1 \quad \alpha > 0$$

.

- Prove that $\sum log\, x_i$ is a sufficient statistics for α,

Examples Sufficient Statistics

$$f(x_1, \ldots, x_n) = \alpha^n (\alpha + 1)^n \prod x_i^{\alpha+1} \prod (1 - x_i)$$

Minimal Sufficient Statistics

- Many sufficient statistics may exists for the same parameter θ.
- Obviously, also the sample X is a sufficient statistic for θ as well as the order statistics $Y_1 \leq Y_2 \leq Y_3 \leq \ldots \leq Y_n$. But they do not provide any data reduction.
- A sufficient statistic $Y^\star = T^\star(\mathbf{X})$ is better of another sufficient statistic, $Y = T(\mathbf{X})$ if $Y^\star$ achieves more data reduction while still retaining all the information on θ

Minimal Sufficient Statistics

Definition The sufficient statistic that achieves the most data reduction is called **minimal sufficient statistic**.

Formally, a sufficient statistic for θ, $Y^\star = t^\star(\mathbf{X})$, is a minimal sufficient statistic if it is a function of any other sufficient statistic $Y = T(\mathbf{X})$

Minimal Sufficient Statistics

In practice, the minimal sufficient statistics gives the greatest data reduction without loss of information about parameters. Here a characterization:

Lehmann-Scheffe Theorem A statistic T is minimal sufficient if the following property holds:
For any two sample points x and y, $f(x;\theta)/f(y;\theta)$ does not depend on θ if and only if $T(x) = T(y)$.

Corollary Minimal sufficient statistic is not unique. But any two are in one-to-one correspondence, so are equivalent.

Minimal sufficiency in exponential family

Theorem: For iid observations from an exponential family

$$f(x|\theta) = h(x)c(\theta)exp\left(\sum_{i=1}^{n}\omega_i(\theta)t_i(x)\right)$$

the statistic

$$T(X) = \left(\sum_{j=1}^{n} T_1(X_j), \ldots, \sum_{j=1}^{n} T_k(X_j)\right)$$

is minimal sufficient for θ

Example

Let consider the pdf

$$f(x;\theta) = \frac{1}{\theta}\exp\left(1 - \frac{x}{\theta}\right) \qquad 0 < \theta < x$$

Find a sufficient statistic for parameter θ.
We know it is possible to find a sufficient statistic by applying the Factorization Theorem. Given the hypothesis of the random sampling we can factorize the given joint pdf as follows:

Example

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n; \theta) &= \prod_{i=1}^{n} \left(\frac{1}{\theta} \exp\left(1 - \frac{x_i}{\theta}\right) I_{(\theta,\infty)}(x_i) \right) \\
&= \frac{1}{\theta^n} \exp\left(\sum_{i=1}^{n} \left(1 - \frac{x_i}{\theta}\right) \right) \prod_{i=1}^{n} I_{(\theta,\infty)}(x_i) \\
&= \frac{1}{\theta^n} \exp\left(n - \frac{1}{\theta} \sum_{i=1}^{n} x_i \right) I_{(\theta,\infty)}(x_{(1)})
\end{aligned}
$$

Example

Therefore we can set: $h(x) = 1$ and

$$g(T_1(x), T_2(x); \theta) = \frac{1}{\theta^n} \exp\left(n - \frac{1}{\theta}\sum_{i=1}^{n} x_i\right) I_{(\theta,\infty)}(x_{(1)})$$

where the two sufficient statistics are the sample sum and sample minimum $T_1(x) = \sum_{i=1}^{n} x_i$ and $T_2(x) = x_{(1)}$. $T_1(x)$ and $T_2(x)$ are two jointly sufficient statistics for the parameter θ. Thus, the dimension of a minimal sufficient statistic (two) is greater than the dimension of the parameter space (one).

Example: Geometric distribution

Let consider the pdf

$$p(x;\pi) = \pi(1-\pi)^{x-1}$$

We know it is possible to find a sufficient statistic by applying the Factorization Theorem:

$$\begin{aligned} p(x_1, x_2, \ldots, x_n; \pi) &= \prod_{i=1}^{n} \frac{\pi}{1-\pi} (1-\pi)^{x_i} \\ &= \left(\frac{\pi}{1-\pi}\right)^n \exp\left(\sum_i x_i log(1-\pi)\right) \end{aligned}$$

So that sample sum, $\sum_i x_i$, is sufficient and complete statistics.

Example: Uniform distribution

Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from a uniform distribution on the interval $[\theta - 1, \theta + 1]$, with $-\infty < \theta < \infty$.

$$\begin{aligned} f(x|\theta) &= 1 \qquad \theta - 1 \leq x \leq \theta + 1 \\ f(x_i|\theta) &= 1 \qquad \theta - 1 \leq x_i \leq \theta + 1 \\ f(x_1, \ldots, x_n|\theta) &= 1 \qquad \theta \leq min(x_i) + 1 \, and \, \theta \geq max(x_i) - 1 \end{aligned}$$

$f(x_1, \ldots, x_n|\theta) = f(y_1, \ldots, y_n|\theta)$ is independent of θ if and only if $min(x_i) = min(y_i)$ and $max(x_i) = max(y_i)$, therefore $T(X) = (min(X_i), max(X_i))$ is minimal sufficient.

Exercises:

Assume that the random variables $X_1, X_2, \ldots, X_n$ form a random sample of size n form the distribution specified, and show that the statistic T specified is a sufficient statistic for the parameter:

Exercise 1: A normal distribution for which the mean μ is known and the variance σ^2 is unknown; $T = \sum_{i=1}^{n}(X_i - \mu)^2$.

Exercise 2: A gamma distribution with parameters α and β, where the value of β is known and the value of α is unknown (α); $T = \prod_{i=1}^{n} X_i$

Exercise 3: A uniform distribution on the interval $[a, b]$, where the value of a is known and the value of b is unknown ($b > a$); $T = max(X_1, \ldots, X_n)$.

Exercise 4: A uniform distribution on the interval $[a, b]$, where the value of b is known and the value of a is unknown ($b > a$); $T = min(X_1, \ldots, X_n)$.

Exercises:

Exercise 5: Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, and the value of β is known. Show that the statistic $T = \sum_{i=1}^{n} logX_i$ is a sufficient statistic for the parameter α.

Exercise 6: The Pareto distribution has density function:

$$f(x|x_0;\theta) = \theta x_0^{\theta} x^{-\theta-1}, \qquad x \geq x_0, \quad \theta > 1$$

Assume that $x_0 > 0$ is given and that $X_1, X_2, \ldots, X_n$ is an i.i.d. sample. Find a sufficient statistic for *theta* by

(a) using factorization theorem,
(b) using the property of exponential family. Are they the same? If not, why are both of them sufficient?

Example of in-sufficiency

Suppose that X_1 and X_2 are two independent Bernoulli random variables with parameter p, $\quad 0 < p < 1$. Show that the statistics $T = X_1 - X_2$ is not a sufficient statistics for p.

Example of in-sufficiency

- $(X_1, X_2, \ldots, X_n)$ iid $Poisson(\lambda)$ $T = X_1 - X_2$ is not sufficient
- $(X_1, X_2, \ldots, X_n)$ iid pmf $f(x; \theta)$ $T = (X_1, X_2, \ldots, X_{n-1})$ is not sufficient.