Sufficiency

MIT 18.655

Dr. Kempthorne

Spring 2016

Sufficiency

Statistical Decision Problem

- $X \sim P_\theta, \theta \in \Theta$.
- Action space $\mathcal{A}$
- Loss function: $L(\theta, a)$
- Decision procedures: $\delta(\cdot) : \mathcal{X} \to \mathcal{A}$

Issue

- $\delta(X)$ may be inefficient ignoring important information in X that is relevant to θ.
- $\delta(X)$ may be needlessly complex using information from X that is irrelevant to θ.
- Suppose a statistic $T(X)$ summarized all the relevant information in X
- We could limit focus to decision procedures $\delta_T(t) : T(\mathcal{X}) \to R$.

Sufficiency: Examples

Example 1 Bernoulli Trials Let $X = (X_1, \dots, X_n)$ be the outcome of n i.i.d $Bernoulli(\theta)$ random variables

- The pmf function of X is:

$$\begin{aligned} p(X \mid \theta) &= P(X_1 = x_1 \mid \theta) \times P(X_2 = x_2 \mid \theta) \times \cdots \times P(X_n = x_n \mid \\ &= \theta^{x_1}(1-\theta)^{1-x_1} \times \theta^{x_2}(1-\theta)^{1-x_2} \times \cdots \theta^{x_n}(1-\theta)^{1-x_n} \\ &= \theta^{\sum x_i}(1-\theta)^{(n-\sum x_i)} \end{aligned}$$

- Consider $T(X) = \sum_{i=1}^{n} X_i$ whose distribution has pmf:

$$\binom{n}{t}\theta^t(1-\theta)^{n-t}, 0 \leq t \leq n.$$

- The distribution of X given $T(X) = t$ is uniform over the n-tuples X: $T(X) = t$.
- The unconditional distribution of X is given by generating $T \sim Binomial(n, \theta)$, and then choosing X randomly according to the uniform distribution over all tuples $\{x = (x_1, \dots, x_n) : T(x) = t\}$

- Given $T(X) = t$, the choice of tuple X does not require knowledge of θ.
- After knowing $T(X) = t$, the additional information in X is the sequence/order information which does not depend on θ.
- To make decision concerning θ, we should only need the information of $T(X) = t$, since the value of X given t reflects only the order information in X which is independent of θ.

Definition Let $X \sim P_\theta, \theta \in \Theta$ and $T(X) : \mathcal{X} \to \mathcal{T}$ is a statistic of X. The statistic T is *sufficient* for θ if the conditional distribution of X given $T = t$ is independent of θ (almost everywhere wrt $P_T(\cdot)$).

Sufficiency Examples

Example 1. Bernoulli Trials

- $X = (X_1, \ldots, X_n)$: X_i iid $Bernoulli(\theta)$
- $T(X) = \sum_1^n X_i \sim Binomial(n, \theta)$
- Prove that $T(X)$ is sufficient for X by deriving the distribution of $X \mid T(X) = t$.

Example 2. Normal Sample Let $X_1, \ldots, X_n$ be iid $N(\theta, \sigma_0^2)$ r.v.'s where σ_0^2 is known. Evaluate whether $T(X) = (\sum_1^n X_i)$ is sufficient for θ.

- Consider the transformation of

  $$X = (X_1, X_2, \ldots, X_n) \text{ to } Y = (T, Y_2, Y_3, \ldots, Y_n)$$

  where

  $$T = \sum X_i \text{ and}$$
  $$Y_2 = X_2 - X_1, Y_3 = X_3 - X_1, \ldots, Y_n = X_n - X_1$$

  (The transformation is 1-1, and the Jacobian of the transformation is 1.)

- The joint distribution of $X \mid \theta$ is $N_n(\mu \times \mathbf{1}, \sigma_0^2 I_n)$.
- The joint distribution of $Y \mid \theta$ is $N_n(\mu_Y, \Sigma_{YY})$

$$\mu_Y = (n\theta, 0, 0, \ldots, 0)^T$$

$$\Sigma_{YY} = \left[\begin{array}{c|ccccc} n\sigma_0^2 & 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & 2\sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \cdots & \sigma_0^2 \\ 0 & \sigma_0^2 & 2\sigma_0^2 & \sigma_0^2 & \cdots & \sigma_0^2 \\ 0 & \sigma_0^2 & \sigma_0^2 & 2\sigma_0^2 & \cdots & \sigma_0^2 \\ \vdots & \vdots & \vdots & & \ddots & \\ 0 & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & & 2\sigma_0^2 \end{array}\right]$$

- T and $(Y_2, \ldots, Y_n)$ are independent
  $\Longrightarrow (Y_2, \ldots, Y_n)$ given $T = t$ is
  the unconditional distribution
  $\Longrightarrow T$ is a sufficient statistic for θ.
- Note: all functions of $(Y_2, \ldots, Y_n)$ are independent of θ and T, which yields independence of $\bar{X}$ and s^2 :

$$s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n [\frac{1}{n}\sum_{j=1}^n (X_i - X_j)]^2$$

Sufficiency Examples

Example 1.5.2 Customers arrive at a service counter according to a Poisson process with arrival rate parmaeter θ.
Let X_1 and X_2 be the inter-arrival times of first two customers. (From time 0, customer 1 arrives at time X_1 and customer 2 at time $X_1 + X_2$. Prove that $T(X_1, X_2) = X_1 + X_2$ is sufficient for θ.

- X_1 and X_2 are iid $Exponential(\theta)$ r.v.'s (by A.16.4).
- The $Exponential(\theta)$ r.v. is the special case of the $Gamma(p, \theta)$ distribution with density with $p = 1$
$$f(x \mid \theta, p) = \frac{\theta^p x^{p-1} e^{-\theta x}}{\Gamma(p)}, 0 < x < \infty$$
- Theorem B.2.3: If X_1 and X_2 are independent random variables with $\Gamma(p, \lambda)$ and $\Gamma(q, \lambda)$ distributions,
  - $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$ are independent and
$$Y_1 \sim Gamma(p + r, \theta) \text{ and } Y_2 \sim Beta(p, q).$$
- So, with $p = q = 1$, $Y_1 \sim Gamma(2, \theta)$ and $Y_2 \sim Uniform(0, 1)$, independently.
- $[(X_1, X_2) \mid T = t] \sim (X, Y)$ with $X \sim Uniform(0, t)$; $Y = t - X$

Sufficiency: Factorization Theorem

Theorem 1.5.1 (Factorization Theorem Due to Fisher and Neyman). In a regular model, a statistic $T(X)$ with range $\mathcal{T}$ is sufficient for $\theta \in \Theta$, iff there exists functions

$$g(t, \theta) : \mathcal{T} \times \Theta \rightarrow R \text{ and } h : \mathcal{X} \rightarrow R,$$

such that

$$p(x \mid \theta) = g(T(x), \theta)h(x), \text{ for all } x \in \mathcal{X} \text{ and } \theta \in \Theta.$$

Proof: Consider the discrete case where $p(x \mid \theta) = P_\theta(X = x)$. First, suppose T is sufficient for θ. Then, the conditional distribution of X given T is independent of θ and we can write

$$\begin{aligned} P_\theta(x) &= P_\theta(X = x, T = t(x)) \\ &= [P_\theta(T = t(x))] \times [P(X = x \mid T = t(x))] \\ &= [g(T(x), \theta)] \times [h(x)] \end{aligned}$$

where $g(t, \theta) = P_\theta(T = t)$

and $h(x) = \begin{cases} 0, & \text{if} \quad P_\theta(x) = 0, \text{ for all } \theta \\ P_\theta(X = x \mid T = t(x)), & \text{if} \quad P_\theta(X = x) > 0 \text{ for some } \theta \end{cases}$

Sufficiency: Factorization Theorem

Proof (continued). Second, suppose that $P_\theta(x)$ satisfies the factorization:

$$P_\theta(x) = g(t(x), \theta)h(x).$$

Fix $t_0 : P_\theta(T = t_0) > 0$, for some $\theta \in \Theta$. Then

$$P_\theta(X = x \mid T = t_0) = \frac{P_\theta(X=x, T=t_0)}{P_\theta(T=t_0)}.$$

- The numerator is $P_\theta(X = x)$ when $t(X) = t_0$ and 0 when $t(X) \neq t_0$
- The denominator is $P_\theta(T = t_0) = \sum_{\{x:t(x)=t_0\}} P_\theta(X = x) = \sum_{\{x:t(x)=t_0\}} g(t(x), \theta)h(x)$

$$P_\theta(X = x \mid T = t_0) = \begin{cases} 0 & if \quad t(x) \neq t_0 \\ \dfrac{g(t_0, \theta)h(x)}{g(t_0, \theta) \sum_{\{x':t(x)=t_0\}} h(x')}, & if \quad t(x) = t_0 \end{cases}$$

(This is independent of θ as g-factors cancel)

Sufficiency: Factorization Theorem

More advanced proofs:

- Ferguson (1967) details proof for absolutely continuous X under regularity conditions of Neyman (1935).
- Lehmann (1959) *Testing Statistical Hypotheses* (Theorem 8 and corollary 1, Chapter 2) details general measure-theoretic proof.

Example 1.5.2 (continued) Let $X_1, X_2, \ldots, X_n$ be inter-arrival times for n customers which are iid *Exponential*(θ) r.v.'s

$$p(x_1, \ldots, x_n \mid \theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i}, \text{ where } 0 < x_i, i = 1, \ldots, n$$

- $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is sufficient by factorizaton theorem.
- $g(t, \theta) = \theta^n exp(-\theta \sum_1^n x_i)$ and $h(x_1, \ldots, x_n) = 1$.

Sufficiency: Applying Factorization Theorem

Example: Sample from Uniform Distribution Let $X_1, \ldots, X_n$ be a sample from the *Uniform*(α, β) distribution:

$$p(x_1, \ldots, x_n \mid \alpha, \beta) = \frac{1}{(\beta-\alpha)^n} \prod_{i=1}^n I_{(\alpha,\beta)}(x)$$

- The statistic
  $$T(x_1, \ldots, x_n) = (min\ x_i, max\ x_i)$$
  is sufficient for $\theta = (\alpha, \beta)$
  $$\prod_{i=1}^n I_{(\alpha,\beta)}(x_i) = I_{(\alpha,\beta)}(min\ x_i) I_{(\alpha,\beta)}(max\ x_i)$$
- If α is known, then $T = max\ \ x_i$ is sufficient for β
- If β is known, then $T = min\ \ x_i$ is sufficient for α

Sufficiency: Applying Factorization Theorem

Example 1.5.4 Normal Sample. Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, with unknown $\theta = (\mu, \sigma^2) \in R \times R_+$

The joint density is

$$
\begin{aligned}
p(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x_i - \mu)^2) \\
&= (2\pi\sigma^2)^{-n/2} exp(-\frac{n\mu^2}{2\sigma^2}) \times \\
&\quad exp\left\{-\frac{1}{2\sigma^2}(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i)\right\} \\
&= g(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i; \theta)
\end{aligned}
$$

- $T(X_1, \ldots, X_n) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient.
- $T^*(X_1, \ldots, X_n) = (\bar{X}, s^2)$ is sufficient, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $s^2 = \frac{1}{(n-1)}\sum_{i=1}^n (X_i - \bar{X})^2)$ are sufficient.

Note: Sufficient statistics are not unique (their level sets are!!).

Sufficiency: Applying Factorization Theorem

Example 1.5.5 Normal linear regression model. Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim N(\mu_i, \sigma^2)$, where

$$\mu_i = \beta_1 + \beta_2 z_i,\ i = 1, 2, \ldots, n$$

and z_i are constants.

- Under what conditions is $\theta = (\beta_1, \beta_2, \sigma^2)$ identifiable?
- Under those conditions, the joint density for $(Y_1, \ldots, Y_n)$ is

$$\begin{aligned} p(y_1, \ldots, y_n \mid \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2) \\ &= (2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1 - \beta_2 z_i)^2 \\ &= (2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (\beta_1 + \beta_2 z_i)^2) \\ &\quad \times exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i^2 - 2(\beta_1 + \beta_2 z_i)y_i)) \end{aligned}$$

which equals

$$(2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (\beta_1 + \beta_2 z_i)^2)$$
$$\times exp(-\frac{1}{2\sigma^2}[(\sum_{i=1}^n y_i^2) - 2\beta_1(\sum_{i=1}^n y_i) - 2\beta_2(\sum_{i=1}^n z_i y_i))$$

- $T = (\sum_{i=1}^n Y_i^2, \sum_{i=1}^n Y_i, \sum_{i=1}^n z_i Y_i)$ is sufficient for θ

Sufficiency and Decision Theory

Theorem: Consider a statistical decision problem with:

- $X \sim P_\theta, \theta \in \Theta$ with sample space $\mathcal{X}$ and parameter space Θ
- $\mathcal{A} = \{actions\ a\}$
- $L(\theta, a) : \Theta \times \mathcal{A} \to R$, loss function
- $\delta(X) : \mathcal{X} \to \mathcal{A}$, a decision procedure
- $R(\theta, \delta(X)) = E[L(\theta, \delta(X)) \mid \theta]$, risk function

If $T(X)$ is sufficient for θ, where $X \sim P_\theta, \theta \in \Theta$, then we we can find a decision rule $\delta^*(T(X))$ depending only on $T(X)$ that does as well as $\delta(X)$

Proof 1: Consider randomized decision rule based on $(T(X), X^*)$, where X^* is the random variable with conditional distribution:

$$X^* \sim [X \mid T(X) = t_0]$$

Note:

- δ^* will typically be randomized (due to X^*)
- δ^* specified by value $T(X) = t$ and conditionally random X^*

Proof 2:

- By sufficiency of $T(X)$, the distribution of $\delta(X)$ given $T(X) = t$ does not depend on θ.
- Draw δ^* randomly from this conditional distribution.
- The risk of δ^* satisfies:

$$\begin{aligned} R(\theta, \delta^*) &= E_T\{E_{X|T}[L(\theta, \delta^*(T)) \mid T]\} \\ &= E_T\{E_{X|T}[L(\theta, \delta(X)) \mid T]\} = R(\theta, \delta(X)) \end{aligned}$$

Example 1.5.6 Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ consists of iid $N(\theta, 1)$ r.v.'s. By the factorization theorem $T(\mathbf{X}) = \; {}_1^n X_i$ is sufficient.
Let $\delta(X) = X_1$.
Define $\delta^*(T(X))$ as follows

$$\delta^*(T(X)) = T(X) + \sqrt{\tfrac{N-1}{N}} Z, \text{ where } Z \sim N(0, 1),$$

independent of X.

- Given $T(X) = t_0$, $\delta^*(T(X)) \sim N(t_0, \frac{(n-1)}{n})$
- Unconditionally $\delta^*(T(X)) \sim N(\theta, 1)$ (identical to X_1)

Sufficiency and Bayes Models

Definition: Let $X \sim P_\theta, \theta \in \Theta$ and let Π be the Prior distribution on Θ. The statistic $T(X)$ is *Bayes sufficient* for Π if

$\Pi(\theta \mid X = x)$, the Posterior distribution of θ given X

is the same as

$\Pi(\theta \mid T(X) = t(x))$, the Posterior distribution of θ given $T(X)$

for all x .

Theorem 1.5.2 (Kolmogorov). If $T(X)$ is sufficient for θ, then it is Bayes sufficient for every prior distribution Π.
Proof Problem 1.5.14.

Minimal Sufficiency

Issue: Probability models often admit many sufficient statistics.
Suppose $X = (X_1, \ldots, X_n)$ where X_i are iid $P_\theta, \theta \in \Theta$.

- $T(X) = (X_1, \ldots, X_n)$ is (trivially) sufficient
- $T'(X) = (X_{[1]}, X_{[2]}, \ldots, X_{[n]})$ where $X_[j] = j$-th smallest $\{X_i\}$ (j-th order statistic) is sufficient
- $T'(X)$ provides a greater reduction of the data.
- If the X_i are iid $N(\theta, 1)$ then $T'' = \bar{X}$ is sufficient.

Definition A statistic $T(X)$ is *minimally sufficient* if it is sufficient and provides a greater reduction of the data than any other sufficient statistic. If $S(X)$ is any sufficient statistic, then there exists a mapping r:

$$T(X) = r(S(X))$$

Minimal Sufficiency: Example

Example 1.5.1 (continued). $X_1, \ldots, X_n$ are iid *Bernoulli*(θ) and $T = \sum_1^n X_i$ is sufficient.
Let $S(X)$ be any other sufficient statistic. By the factorization therem:

$$p(x \mid \theta) = g(S(x), \theta)h(x),$$

for some functions $g(\cdot, \cdot)$ and $h(\cdot)$. Using the pmf of X we have

$$\theta^T (1-\theta)^{(n-T)} = g(S(x), \theta)h(x), \text{ for all } \theta \in [0,1]$$

Fix any two values of θ, say θ_1 and θ_2 and take the ratio of the pmfs:

$$(\theta_1/\theta_2)^T [(1-\theta_1)/(1-\theta_2)]^{n-T} = g(S(x), \theta_1)/g(S(x), \theta_2)$$

Take logarithm of both sides and solve for T. E.g., $\theta_1 = 2/3$ and $\theta_2 = 1/3$

$$T = r(S(X)) = log[2^n g(S(x), \theta_1)/g(S(x), \theta_2)]/2log2.$$

The Likelihood Function

Definition For $X \sim P_\theta, \theta \in \Theta$ let $p(x \mid \theta)$ be the pmf or density function. The *likelihood function* L for a given observed data value $X = x$ is

$$L_x(\theta) = p(x \mid \theta), \theta \in \Theta$$

The function $L : \mathcal{X}$ to $\mathcal{T}$, the function class

$$\mathcal{T} = \{f : \theta \rightarrow p(x \mid \theta), x \in \mathcal{X}\}$$

Theorem (Dynkin, Lehmann, and Scheffe)
Suppose there exists θ_0:

$$\{x : p(x \mid \theta) > 0\} \subset \{x : p(x \mid \theta_0) > 0\} \text{ for all } \theta.$$

Define: $\Lambda_x(\cdot) = \frac{L_x(\cdot)}{L_x(\theta_0)} : \Theta \rightarrow R.$

Then $\Lambda_x(\cdot)$ is the function-valued statistic that is minimal sufficient.

Proof Problem 1.5.12

Note: As a function, $\Lambda_x(\cdot)$ at θ has value $p(x \mid \theta)/p(x \mid \theta_0)$, the ratio of likelihoods at θ and at θ_0.

Sufficient Statistics and Ancillary Statistics

Suppose $X \sim P_\theta, \theta \in \Theta$ and that $T(X)$ is a sufficient statistic.
Consider a 1:1 mapping of X which includes the sufficient statistic

$$X \rightarrow (T(X), S(X)).$$

Because the mapping is 1:1, we can recover X given $T(X) = t$ and $S(X) = s$.

- $T(X)$ is sufficient for θ, so $S(X)$ is irrelevant so long as $\mathcal{P}=\{P_\theta, \theta \in \Theta\}$ is valid.

Using $S(X)$ to Evaluate Validity of $\mathcal{P}$

- Example 1.5.5: $X = (X_1, \ldots, X_n)$ iid $N(\theta, 1)$
  $$T(X) = \bar{X} \text{ and } S(X) = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$$
  To evaluate the validity of $Var(X_i) \equiv 1$, we need $S(X)$.
- Example 1.5.4: $X = (X_1, \ldots, X_n)$ iid $N(\theta, \sigma^2)$
  $$T(X) = (\textstyle\sum_1^n X_i, \sum_1^n X_i^2) \text{ or equivalently}$$
  $T(X) = (\bar{X}, s^2)$
  To evaluate the validity of the Normal Assumption, we need $S(X)$. (See Problem 1.5.13)