Sufficiency

9/5/2023

Outline

1) Review
2) Sufficiency
3) Factorization Theorem

Sufficiency

Motivation: Coin flipping

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$

$$\Rightarrow X \sim \prod_i \theta^{x_i}(1-\theta)^{1-x_i} \quad \text{on} \quad \{0,1\}^n$$

Then $T(X) = \sum X_i \sim \text{Binom}(n, \theta)$

$$= \theta^t (1-\theta)^{n-t} \binom{n}{t} \quad \text{on} \quad \{0, \ldots, n\}$$

$(X_1, \ldots, X_n) \rightsquigarrow T(X)$ is throwing away data. How do we justify this?

In exp. fam. lingo, $T(X)$ is the "sufficient statistic" for X. Today we'll see why we call it that.

Definition Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model for data X. $T(X)$ is sufficient for $\mathcal{P}$ if $P_\theta(X \mid T)$ does not depend on θ

Example (Cont'd)

$$\mathbb{P}_\theta(X = x \mid T = t) = \frac{\mathbb{P}_\theta(X = x, T = t)}{\mathbb{P}_\theta(T = t)}$$

$$= \frac{\cancel{\theta^{\sum x_i}} \, \cancel{(1-\theta)^{n-\sum x_i}} \, 1\{\sum x_i = t\}}{\cancel{\theta^t} \, \cancel{(1-\theta)^{n-t}} \binom{n}{t}}$$

$$= 1\{\sum x_i = t\} \Big/ \binom{n}{t}$$

So given $T(X) = t$, X is uniform on all seq.s with $\sum x_i = t$

Factorization Theorem

Often, we can identify sufficient stats by inspecting the density.

Theorem (Factorization Theorem)

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a model with densities $p_\theta(x)$ wrt common measure μ.

$T(x)$ is sufficient iff there exist $g_\theta(t)$, $h(x)$ with

$$p_\theta(x) = g_\theta(T(x))\, h(x)$$

for μ-almost-every x: $\mu(\{x : p_\theta(x) \neq g_\theta(T(x)) \cdot h(x)\}) = 0$

[Avoids counterexamples from changing $p_{\theta_0}(x_0)$ some θ_0, x_0]

Rigorous proof in Keener 6.4

<u>Proof (discrete $\mathcal{X}$)</u>: Assume wlog $\mu = \#$ on $\mathcal{X}$

$$(\Leftarrow)\quad \mathbb{P}_\theta(X=x \mid T=t) = \frac{\mathbb{P}_\theta(X=x,\ T(x)=t)}{\mathbb{P}_\theta(T(x)=t)}$$

$$= \frac{\cancel{g_\theta(t)}\, h(x)\, 1\{T(x)=t\}}{\sum_{T(z)=t} \cancel{g_\theta(t)}\, h(z)}$$

$(\Rightarrow)$ Assume $T(x)$ sufficient.

Take $g_\theta(t) = \sum_{T(x)=t} p_\theta(x)$

$$= \mathbb{P}_\theta(T(X)=t)$$

For any $\theta_0 \in \Theta$, let

$$h(x) = p_{\theta_0}(x) \Big/ \sum_{T(z)=T(x)} p_{\theta_0}(z)$$

$$= \mathbb{P}_{\cancel{\theta_0}}(X=x \mid T(X)=T(x))$$

no dep. on θ

Then,

$$g_\theta(T(x))\, h(x) = \mathbb{P}_\theta(T=T(x))\, \mathbb{P}(X=x \mid T=T(x))$$

$$= \mathbb{P}_\theta(X=x)$$

$\boxtimes$

Interpretations of Sufficiency

X is informative about θ only because its distribution depends on θ.

We can think of the data as being generated in two stages:

1) Generate T : distribution dep. on θ
2) Generate $X \mid T$: does <u>not</u> dep on θ

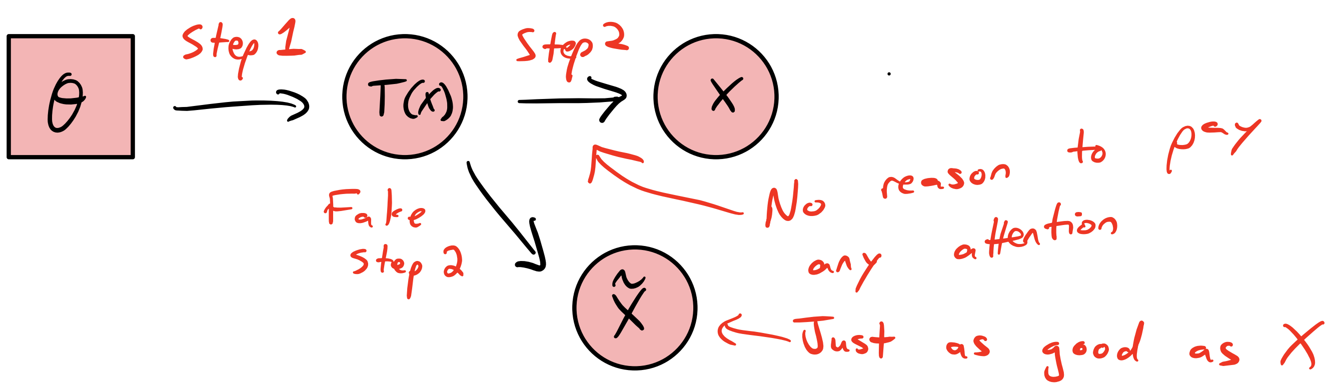
Sufficiency Principle

If $T(X)$ is sufficient for $\mathcal{P}$ then any statistical procedure should depend on X only through $T(X)$

In fact, we could throw away X and generate a new $\tilde{X} \sim P(X \mid T)$ and it would be just as good as X since $\tilde{X} \sim P_\theta$

no θ

In graphical model form:



Examples

Ex. Exponential Families

$$p_\theta(x) = \underbrace{e^{\eta(\theta)' T(x) - B(\theta)}}_{g_\theta(T(x))} \underbrace{h(x)}_{h(x)}$$

Ex. Uniform location family

$$X_1, \dots, X_n \overset{iid}{\sim} U[\theta, \theta+1]$$

$$= 1\{\theta \le x \le \theta + 1\}$$

$$p_\theta(x) = \prod_{i=1}^{n} 1\{\theta \le x_i \le \theta + 1\}$$

$$= 1\{\theta \le X_{(1)}\} \, 1\{X_{(n)} \le \theta + 1\}$$

$\Rightarrow$ $(X_{(1)}, X_{(n)})$ is sufficient.

Order Statistics / Empirical Distribution

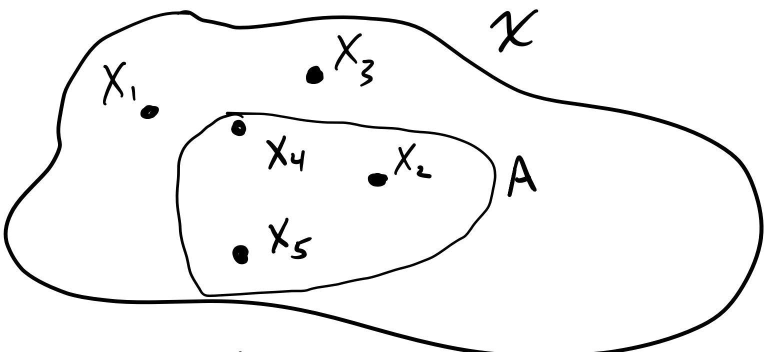
Ex. $X_1, \dots, X_n \overset{iid}{\sim} P_\theta^{(1)}$ for any model
$\mathcal{P}^{(1)} = \{P_\theta^{(1)} : \theta \in \Theta\}$ on $\mathcal{X} \subseteq \mathbb{R}$

P_θ is invariant to perm.s of $X = (X_1, \dots, X_n)$

$\Rightarrow$ All permutations of x are equally likely

$\Rightarrow$ order statistics $(X_{(i)})_{i=1}^n$ ($X_{(k)}$ = k^{th} smallest) are sufficient. [Note $(X_i)_{i=1}^n \rightsquigarrow (X_{(i)})_{i=1}^n$ loses information, specifically the orig. ordering]

For more general $\mathcal{X}$ we can say the empirical distribution $\hat{P}_n(\cdot) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(\cdot)$ is sufficient, where $\delta_{X_i}(A) = 1\{X_i \in A\}$



$\hat{P}_n(A) = \frac{3}{5}$

[Not important that it's a measure in this context; just keeps track of which values came up how many times]

Minimal Sufficiency

Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$

$$p_\theta^{(1)}(x) = \frac{1}{\sqrt{2\pi}} e^{\theta x - \theta^2/2 - x^2/2}$$

exponential family with $T(x) = x$

$$T(X) = \sum X_i \quad \text{sufficient}$$

$$\bar{X} = \frac{1}{n} \sum X_i \quad \text{also}$$

$$S(X) = (X_{(1)}, \ldots, X_{(n)}) \quad \text{too}$$

$$X = (X_1, \ldots, X_n) \quad \text{too}$$

Which can be recovered from which others?

$$\begin{tikzcd}
 & X \arrow[d] \arrow[ddl, bend right] \arrow[ddr, bend left] & \\
 & S(X) \arrow[dl] \arrow[dr] & \\
\sum X_i \arrow[rr, leftrightarrow] & & \bar{X}
\end{tikzcd}$$

these can be compressed further

These are the most compressed. Are they as compressed as possible?

Prop If $T(X)$ is sufficient and $T(x) = f(S(x))$ then $S(X)$ is sufficient

Proof: $$p_\theta(x) = g_\theta(T(x))\,h(x)$$
$$= (g_\theta \circ f)(S(x))\,h(x)$$ $\boxtimes$

Definition: $T(X)$ is minimal sufficient if

1) $T(X)$ is sufficient

2) For any other sufficient $S(X)$,
$$T(X) = f(S(X)) \quad \text{for some } f \quad (\text{a.s. in } \mathcal{P})$$

So, no matter how many more suff. stats we add to our diagram, they will all have arrows pointing to ΣX_i

Likelihood Shape is Minimal

Definition

Assume $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ has densities $p_\theta(x)$

The likelihood function is the (random) function

$$\underbrace{\mathrm{Lik}(\theta}_{\text{function of }\theta}\,;\underbrace{X)}_{\text{data } X \text{ determines which function}} = \underbrace{p_\theta(x)}_{\text{function of } x \text{ with parameter } \theta}$$

The log-likelihood function is its log:

$$\ell(\theta; X) = \log \mathrm{Lik}(\theta; X)$$

The likelihood up to scaling (or ℓ up to vertical shift) is a minimal sufficient statistic

If $T(X)$ is sufficient then

$$\mathrm{Lik}(\theta; x) = \underbrace{g_\theta(T(x))}_{T \text{ determines the "shape"}} \underbrace{h(x)}_{\text{scaling}}$$

HW 2: Likelihood ratios $\left(\dfrac{\mathrm{Lik}(\theta_1; X)}{\mathrm{Lik}(\theta_2; X)}\right)_{\theta_1, \theta_2 \in \Theta}$ minimal suff.

Recognizing Minimal Sufficient Statistics

$T(X)$ is minimal sufficient if

1) $T(X)$ is sufficient (don't forget to check!)

2) $T(x)$ can be recovered from the likelihood shape

Keener Thm 3.11 formalizes condition 2

"$Lik(\cdot\,; x) \propto Lik(\cdot\,; y) \Rightarrow T(x) = T(y)$"

equivalently,

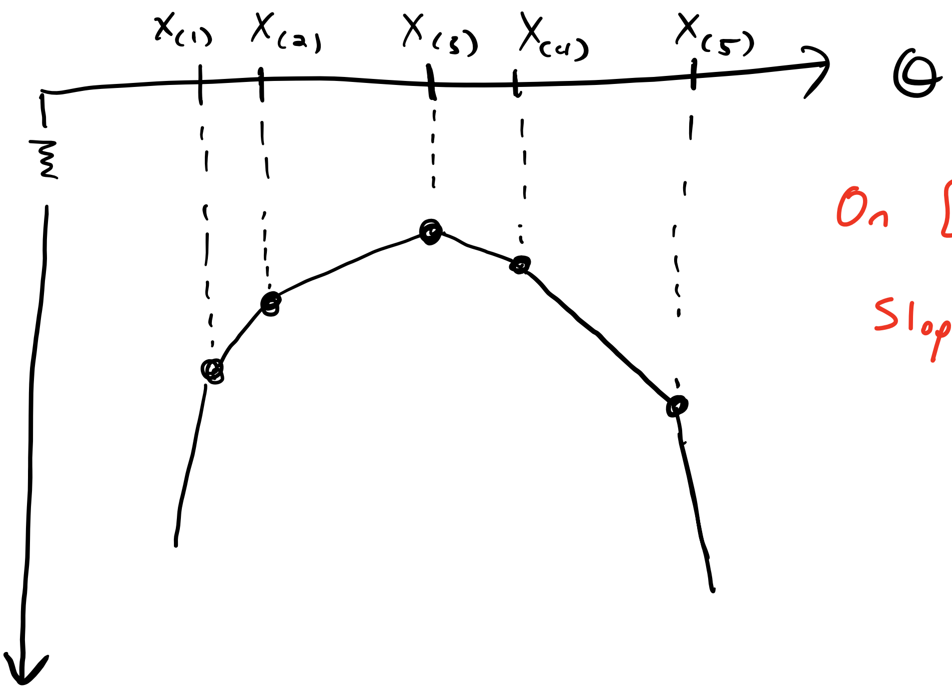
"$\ell(\cdot\,; x) - \ell(\cdot\,; y) = \text{const}(x, y) \Rightarrow T(x) = T(y)$"

Ex Laplace location family

$$X_1, \ldots, X_n \overset{iid}{\sim} p_\theta^{(1)}(x) = \frac{1}{2} e^{-|x-\theta|}$$

$$\ell(\theta; X) = -\sum_{i=1}^{n} |x_i - \theta| - n\log 2$$

Piecewise linear in θ, knots at $x_{(i)}$



On $[x_{(k)}, x_{(k+1)}]$,
Slope $= n - 2k$

$$\ell(\theta; X) = \ell(\theta; Y) + \text{const} \Leftrightarrow X, Y \text{ same order statistics}$$

$\Rightarrow$ order stats are minimal suff.

Minimal sufficiency for exp. fam.s

Suppose $p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x)$

$$\ell(\eta;X) = \underbrace{T(X)'\eta}_{\substack{\text{random linear} \\ \text{function of } \eta}} - \underbrace{A(\eta)}_{\substack{\text{deterministic} \\ \text{function of } \eta}} + \underbrace{\log h(x)}_{\text{(random) const.}}$$

Is $T(X)$ minimal? (always sufficient)

Suppose x and y give same likelihood shape:

$$\ell(\eta;x) - \ell(\eta;y) = \text{const}(x,y)$$

Then $(T(x) - T(y))'\eta = \text{const}(x,y)$ for $\eta \in \Xi$

$\Rightarrow$ $T(x) = T(y)$ <u>or</u>

$T(x) - T(y) \perp \text{Span}\{\eta_1 - \eta_2 : \eta \in \Xi\}$

If $\text{Span}\{\cdots\} = \mathbb{R}^s$, $T(X)$ is minimal

(That is, if Ξ is not contained in a lower-dim affine space)

Otherwise might not be:

If $s=2$, $\Xi = \left\{\begin{pmatrix}\theta \\ 0\end{pmatrix} : \theta \in \mathbb{R}\right\}$ then $T_1(X)$ minimal

[Can we conclude $T(X)$ is <u>not</u> minimal?]

Other parameterizations:

$$p_\theta(x) = e^{\eta(\theta)' T(x) - B(\theta)} h(x) \qquad \theta \in \Theta$$

$T(X)$ minimal if $\text{span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\} = \mathbb{R}^s$