Sufficiency

Outline

1) Sufficiency
2) Factorization Theorem
3) Examples
4) Minimal sufficiency

Three models for coin flipping

Model 3 $\quad X_{i,j} \overset{ind}{\sim} \text{Bernoulli}(\theta_{i,t}) \quad \begin{matrix} i=1,\dots,48 \\ j=1,\dots,n_i \end{matrix} \quad \theta_{i,j} \searrow \text{ in } j$

Model 2 $\quad X_{i+} \overset{ind.}{\sim} \text{Binom}(n_i, \theta_i) \qquad X_{i+} = \sum_{j=1}^{n_i} X_{i,j}$

Model 1 $\quad X_{++} \overset{ind.}{\sim} \text{Binom}(n, \theta) \qquad X_{++} = \sum_{i=1}^{48} \sum_{j=1}^{n_i} X_{i,j}$

most assumptions $\qquad$ fewest assumptions

These models are nested: $\mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \mathcal{P}_3$

Data keeps getting compressed too...
are we losing anything by doing this?

Answer No. X_{++} is a sufficient statistic for $\mathcal{P}_1$, and $(X_{1+},\dots,X_{48+})$ is also sufficient for $\mathcal{P}_2$

Def A statistic $T(X)$ is any function of data X

Sufficiency

Def A statistic $T(X)$ is sufficient for model $\mathcal{P}$ if the conditional distribution of $X \mid T(X)$ is the same for all $P \in \mathcal{P}$

Check definition for $T(X) = X_{++}$ in $\mathcal{P}_1$:

$$p_\theta(x) = \prod_{i=1}^{48} \prod_{j=1}^{n_i} \theta^{x_{ij}} (1-\theta)^{1-x_{ij}}$$

$$= \theta^{x_{++}} (1-\theta)^{n-x_{++}}$$

(why no $\binom{n}{x_{++}}$?)

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

$$= \frac{1\{x_{++} = t\} \cdot \theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}}$$

$$= 1\{x_{++} = t\} / \binom{n}{t}$$

Intuition Suppose we believe Model 1.

Big/small X_{++} more likely with big/small θ

But once we know $X_{++} = 178{,}079$, all data sets X with that many same-side flips are equally likely, regardless of θ

Not true in Models 2 & 3 $\Rightarrow X_{++}$ no longer sufficient

Factorization Theorem

Usually, we can recognize 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 $g_\theta(t)$, $h(x) \geq 0$ with

$$p_\theta(x) = g_\theta(T(x))\, h(x) \quad (\text{for } \mu\text{-a.e. } x)$$

Note we could absorb h into μ as density

(define new base measure ν, $\nu(A) = \int_A h(x)\,d\mu(x)$)

$\Rightarrow \mathcal{P}$ has densities $p_\theta(x) = g_\theta(T(x))$ wrt ν

Interp: after changing base measure, density depends on x only through $T(x)$

(Can't absorb $g_\theta(T(x))$ into μ: depends on θ)

Proof (discrete $\mathcal{X}$): 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, let

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

$$h(x) = \mathbb{P}(X = x \mid T(X) = T(x))$$

no dep. on θ

$$\Rightarrow \quad g_\theta(T(x))\, h(x) = \mathbb{P}_\theta(T(X) = T(x) \text{ and } X = x)$$

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

Proof similar for general densities

requires care about conditioning

Examples

Ex. Uniform location family

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

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

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

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

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

Ex. Normal location family

$$X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$$

$$p_\theta(x) = (2\pi)^{-n/2} \prod_{i=1}^n e^{-(x_i-\theta)^2/2}$$

$$= e^{\theta \sum_i x_i - n\theta^2/2} \cdot \frac{e^{-\sum x_i^2/2}}{(2\pi)^{n/2}}$$

(collect factors with no dep. on θ)

$\Longrightarrow$ $\sum X_i$ is sufficient

Ex. Poisson family

$$X_1, \dots, X_n \overset{iid}{\sim} \text{Pois}(\theta) = \frac{\theta^x e^{-\theta}}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

$$p_\theta(x) = \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \theta^{\sum x_i} e^{-n\theta} \cdot \frac{1}{\prod x_i!}$$

$\Longrightarrow$ $\sum X_i$ is sufficient

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

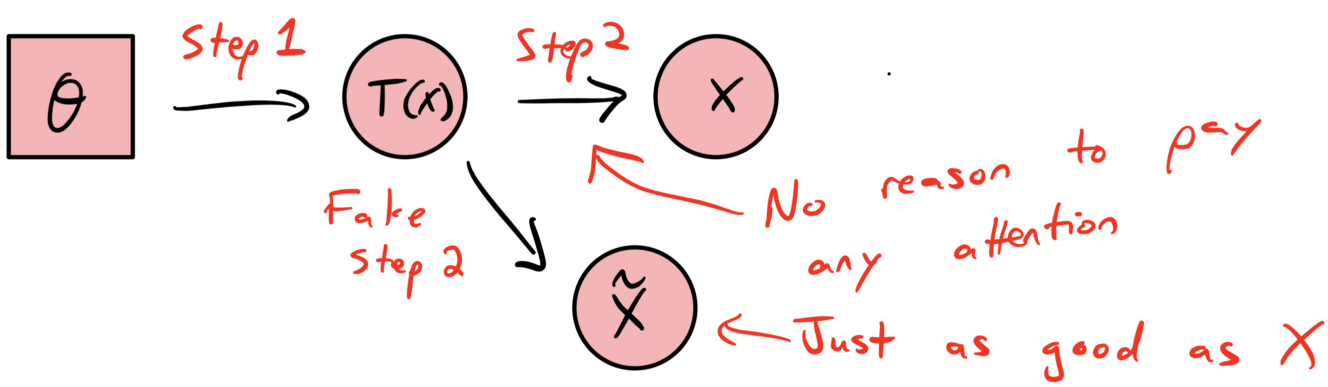
Sufficiency Principle

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:



Order Statistics

For $x_1, \dots, x_n \in \mathbb{R}$, define order statistics

$$\min_i x_i = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} = \max_i x_i$$

Ex (iid sampling on $\mathbb{R}$) $X_1, \dots, X_n \overset{iid}{\sim} P_\theta$,

any model $\mathcal{P} = \{ P_\theta^n : \theta \in \Theta \}$ on $\mathcal{X} \subseteq \mathbb{R}$

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

$\Rightarrow$ All permutations of x are equally likely

$\Rightarrow$ Order statistics $S(X) = (X_{(i)})_{i=1}^n$ sufficient

$X \rightsquigarrow S(X)$ forgets orig. ordering of observations

Empirical Distribution

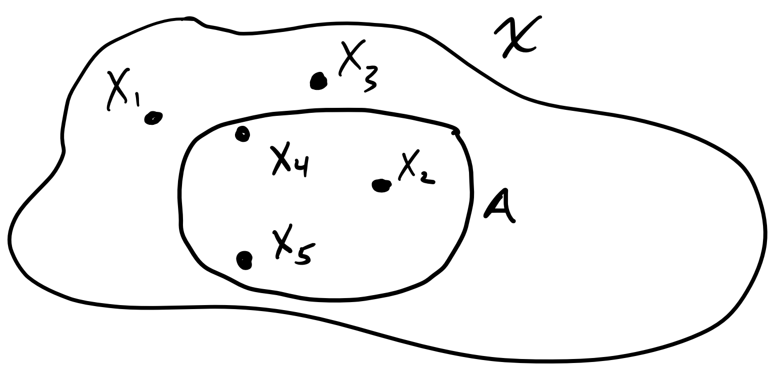
Order statistics depend on total ordering of $\mathcal{X}$

What about more general sample space?

Define Dirac measure $\delta_x(A) = 1\{x \in A\}$

Empirical distribution $\hat{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\cdot)$

random measure on $\mathcal{X}$, determined by sample



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

Ex (iid sampling) $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$

any model $\mathcal{P} = \{P_\theta^n : \theta \in \Theta\}$ on any $\mathcal{X}$

$\hat{P}_n$ is sufficient

$X \rightsquigarrow \hat{P}_n$ records which values observed, how many times

Minimal Sufficiency

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

$T(X) = \sum X_i$ sufficient

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

$S(X) = (X_{(1)}, \dots, X_{(n)})$ too

$X = (X_1, \dots, X_n)$ 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:
$$\begin{aligned} p_\theta(x) &= g_\theta(T(x))\, h(x) \\ &= (g_\theta \circ f)(S(x))\, h(x) \end{aligned}$$
$\boxtimes$

Definition: $T(X)$ is <u>minimal sufficient</u> if

1) $T(X)$ is sufficient
2) For any other sufficient $S(X)$,

$$T(X) = f(S(X)) \text{ for some } f \text{ (a.s. in } \mathcal{P})$$

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

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

function of x with parameter θ

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

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

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

$$\text{“}Lik(\cdot\,; x) \propto Lik(\cdot\,; y) \Rightarrow T(x) = T(y)\text{”}$$

equivalently,

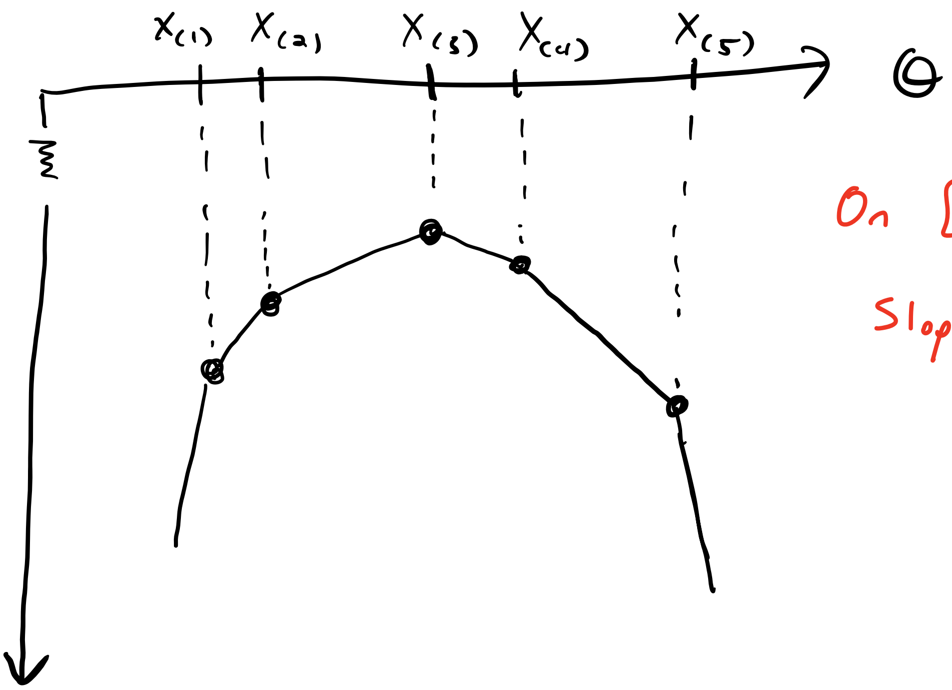
$$\text{“}\ell(\cdot\,; x) - \ell(\cdot\,; y) = \text{const}(x, y) \Rightarrow T(x) = T(y)\text{”}$$

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

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

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 $\text{Span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\} = \mathbb{R}^s$