Outline

1) Review
2) Sufficiency
3) Factorization Theorem
Sufficiency

Motivation: Coin flipping

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

\[ \Rightarrow X \sim \prod \theta^x_i (1-\theta)^{1-x_i} \quad \text{on } \{0,1\}^n \]

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

\[ = \theta^t (1-\theta)^{n-t} \binom{n}{t} \quad \text{on } \{0,\ldots,n\}^3 \]

\((X_1, \ldots, X_n) \Rightarrow T(X)\) 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 \( P_\theta(X | T) \) does not depend on \( \theta \)

Example (Cont'd)

\[ P_\theta(X = x \mid T = t) = \frac{\frac{P_\theta(X = x, T = t)}{P_\theta(T = t)}} = \frac{\theta^{\sum X_i} (1-\theta)^{n-\sum X_i} 1\{\sum X_i = t\}}{\theta^t (1-\theta)^{n-t} \binom{n}{t}} \]

\[ = \frac{1\{\sum X_i = t\}}{\binom{n}{t}} \]

So given \( T(X) = t \), \( X \) is uniform on all seqs with \( \sum X_i = t \)
Factorization Theorem

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

Theorem (Factorization Theorem)

Let \( S = \{ P_\theta : \theta \in \Theta \} \) be a model with densities \( p_\theta(x) \) wrt common measure \( \mu \).

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

\[
p_\theta(x) = g_\theta(T(x)) h(x)
\]

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

[Avoids counterexamples from changing \( p_\theta(x) \) some \( \theta, x_0 \)]

Rigorous proof in Keener 6.4
**Proof (discrete \( x \)):** Assume wlog \( \mu = \# \) on \( X \)

\[
\left( \leftarrow \right) \quad P_{\theta}(X = x \mid T = t) = \frac{P_{\theta}(X = x, T(x) = t)}{P_{\theta}(T(x) = t)} = \frac{g_{\theta}(t) h(x) \mathbb{1}_{\{T(x) = t\}}}{\sum_{T(z) = t} g_{\theta}(t) h(z)}
\]

\[
\left( \rightarrow \right) \quad \text{Assume } T(x) \text{ sufficient.}
\]

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

\[
= P_{\theta}(T(X) = t)
\]

For any \( \theta \in \Theta \), let

\[
h(x) = \frac{P_{\theta}(x)}{\sum_{T(z) = T(x)} P_{\theta}(z)}
\]

\[
= P_{\theta}(X = x \mid T(X) = T(x))
\]

Then,

\[
g_{\theta}(T(x)) h(x) = P_{\theta}(T = T(x)) P_{\theta}(X = x \mid T = T(x)) = P_{\theta}(X = x)
\]

\( \Box \)
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|T : does not dep on Θ

Sufficiency Principle

If T(X) is sufficient for P then any
statistical procedure should depend on X only
through T(X)

In fact, we could throw away X and generate
a new Č ~ P(X|T) and it would
be just as good as X since Č ~ P

In graphical model form:

[Diagram showing the process with Θ, T(X), and X nodes, and arrows indicating the steps and dependencies.]
Examples

Ex. Exponential Families

\[ p_\theta(x) = \frac{\gamma(\theta, \tau(x)) - B(\theta)}{g_\theta(\tau(x)) h(x)} \]

Ex. Uniform location family

\[ X_1, \ldots, X_n \sim iid \quad U[\theta, \theta + 1] \]

\[ = 1\{\theta \leq x \leq \theta + 1\} \]

\[ p_\theta(x) = \prod_{i=1}^{n} 1\{\theta \leq x_i \leq \theta + 1\} \]

\[ = 1\{\theta \leq x_{(1)} \leq \theta + 1\} \quad 1\{x_{(n)} \leq \theta + 1\} \]

\[ \Rightarrow (X_{(1)}, X_{(n)}) \text{ is sufficient.} \]
Ex. $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta^{(\cdot)}$ for any model

\[ P^{(\cdot)} = \{P^{(\cdot)} : \theta \in \Theta\} \text{ on } \mathcal{X} \subseteq \mathbb{R} \]

$P_\theta$ is invariant to perms of $X = (X_1, \ldots, X_n)$

$\Rightarrow$ All permutations of $x$ are equally likely

$\Rightarrow$ order statistics $(X_{(i)})_{i=1}^n$, $(X_{(i)} = k^{th}$ smallest are sufficient. [Note $(X_i)_{i=1}^n \not\sim (X_{(i)})_{i=1}^n$ 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]
Consider \( X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma) \)

\[
p_\theta(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}
\]

exponential family with \( T(x) = x \)

\[
T(X) = \sum X_i \text{ sufficient}
\]

\[
\bar{X} = \frac{1}{n} \sum X_i \text{ also}
\]

\[
S(X) = (X_1, ..., X_n) \text{ too}
\]

\[
X = (X_1, ..., X_n) \text{ too}
\]

Which can be recovered from which others?

\[
X \downarrow \quad \sum X_i \leftrightarrow \bar{X}
\]

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: \( p_\theta(x) = g_\theta(T(x)) h(x) \)
\[ = (g_\theta \circ f)(s(x)) h(x) \]
\[ \blacksquare \]

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

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

\[ T(X) = f(s(x)) \quad \text{for some } f \quad (a.s. \text{ in } \mathbb{P}) \]

So, no matter how many more suff. stats we add to our diagram, they will all have arrows pointing to \( E_X \).
Likelihood Shape is Minimal

**Definition**

Assume \( \Theta = \{ \theta : \theta \in \Theta \} \) has densities \( p_\theta(x) \). The likelihood function is the (random) function

\[
\text{Lik}(\theta; x) = p_\theta(x)
\]

The log-likelihood function is its log:

\[
\ell(\theta; x) = \log \text{Lik}(\theta; x)
\]

The likelihood up to scaling (or \( \ell \) up to vertical shift) is a minimal sufficient statistic.

If \( T(X) \) is sufficient then

\[
\text{Lik}(\theta; x) = g_\theta(T(x)) h(x)
\]

\( T \) determines the scaling "shape".

HW 2: Likelihood ratios \( \left( \frac{\text{Lik}(\theta_1; x)}{\text{Lik}(\theta_2; x)} \right) \text{ for } \theta_1, \theta_2 \in \Theta \)
Recognizing Minimal Sufficient Statistics

$T(X)$ is minimal sufficient if

1) $T(X)$ is sufficient

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

Keener Thm 3.11 formalizes condition 2

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

equivalently

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

\[ X_1, \ldots, X_n \overset{iid}{\sim} \rho_{\theta}^{(\cdot)}(x) = \frac{1}{2} e^{-|x-\theta|} \]

\[ l(\theta; x) = - \sum_{i=1}^{n} |x_i - \theta| - n \log 2 \]

Piecewise linear in \( \theta \), knots at \( x_{(i)} \)

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

\[ l(\theta; x) = l(\theta; y) + \text{const} \iff X, Y \text{ same order statistics} \]

\[ \implies \text{order stats are minimal suff.} \]
Minimal sufficiency for exp. fams

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

\[ l(\eta; x) = \frac{T(X)' \eta - A(\eta)}{\text{random linear function of } \eta} + \frac{\log h(x)}{\text{deterministic function of } \eta} + \text{(random) const.} \]

Is \( T(X) \) minimal? \textbf{(always sufficient)}

Suppose \( x \) and \( y \) give same likelihood shape:

\[ l(\eta; x) - l(\eta; y) = \text{const}(x,y) \]

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

\[ \Rightarrow \quad T(x) = T(y) \quad \text{or} \]

\[ T(x) - T(y) \perp \text{Span} \{ \eta, - \eta : \eta \in \Xi \} \]

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

\text{(That is, if } \Xi \text{ is not contained in a lower-dim affine space)}

Otherwise might not be:

If \( s = 2 \), \( \Xi = \{ (\theta) : \theta \in \mathbb{R} \} \) then \( T_1(X) \) minimal

\[[\text{Can we conclude } T(X) \text{ is not minimal?}]]\]
Other parameterizations:

\[ \rho_{\theta}(x) = e^{\gamma(\theta)T(x) - B(\theta)h(x)}, \quad \Theta \in \Theta \]

\( T(X) \) minimal if \( \delta_{\Theta} \gamma(\theta_1) - \gamma(\theta_2) : \Theta, \Theta_2 \in \Theta^2 = \mathbb{R}^3 \)

\[ \eta_2 \]

\( T(X) \) minimal

\( \gamma \quad \gamma^T T(x) \) is sufficient

\( \Rightarrow T(X) \) prob.

not minimal