Outline

1) Review
2) Sufficiency
3) Factorization Theorem
**Sufficiency**

**Motivation**: Coin flipping

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

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

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

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

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

**Example (Cont'd)**

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

\[ = \frac{\theta^{\sum x_i} (1-\theta)^{n-\sum x_i} 1\{\sum x_i = t\}}{\theta^t (1-\theta)^{n-t} \binom{n}{t}} \]

\[ = 1\{\sum x_i = t\} \binom{n}{t} \]

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

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

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

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

[Avoids counterexamples from changing \( \rho_{\theta_0}(x) \) some \( \theta_0, x_0 \)]

Rigorous proof in Keener 6.4
**Proof (discrete $X$):** Assume wlog $\mu = \# \text{ on } X$

\[
(\Leftarrow) \quad P_\theta(X=x \mid T=t) = \frac{P_\theta(X=x, T(x)=t)}{P_\theta(T(x)=t)} \\
= \frac{g_\theta(t) h(x) 1\{T(x)=t\}}{\sum_{T(z)=t} g_\theta(t) h(z)}
\]

\[
(\Rightarrow) \quad \text{Assume } T(X) \text{ sufficient.} \\
\text{Take } g_\theta(t) = \sum_{T(x)=t} P_\theta(x) \\
= P_\theta(T(X)=t)
\]

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

\[
h(x) = \frac{P_{\theta_0}(x)}{\sum_{T(z)=T(x)} P_{\theta_0}(z)} \\
= P_{\theta_0}(X=x \mid T(X)=T(x)) \\
\]  

Then,

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

$\blacksquare$
Interpretations of Sufficiency

\(X\) is informative about \(\Theta\) only because its distribution depends on \(\Theta\).

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

1) Generate \(T\): distribution dep. on \(\Theta\)
2) Generate \(X \mid T\): does not dep. on \(\Theta\)

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 \(\tilde{X} \sim P(X \mid T)\) and it would be just as good as \(X\) since \(\tilde{X} \sim P_0\)

In graphical model form:

- Step 1: \(\Theta \rightarrow T(X)\)
- Step 2: \(T(X) \rightarrow X\)
- \(\tilde{X}\) (Fake step 2)

No reason to pay any attention

Just as good as \(X\)
Examples

Ex. Exponential Families

\[ p_\theta(x) = \frac{e^{\gamma(x', \tau \alpha)} - B(\theta)}{g_\theta(T(x))} h(x) \]

Ex. Uniform location family

\[ X_1, \ldots, X_n \overset{iid}{=} U[\theta, \theta+1] \]

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

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

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

\[ \Rightarrow (X_{(1)}, X_{(n)}) \text{ is sufficient.} \]
\[ \text{Ex. } X_1, \ldots, X_n \overset{iid}{\sim} P_0^{(\Theta)} \text{ for any model} \]

\[ P^{(\Theta)} = \{ P_0^{(\Theta)} : \Theta \in \Theta \} \text{ on } X \subseteq \mathbb{R} \]

\( P_0 \) is invariant to perms. of \( X=(X_1, \ldots, X_n) \)

\[ \Rightarrow \text{ All permutations of } x \text{ are equally likely} \]

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

For more general \( X \) 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 $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \tau)$

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

exponential family with $T(x) = x$

$T(X) = \sum X_i$ sufficient

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

$S(X) = (X_1, \ldots, X_n)$ too

$X = (X_1, \ldots, X_n)$ too

Which can be recovered from which others?

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

These can be compressed further.
Prop: If $T(X)$ is sufficient and $T(X) = f(S(X))$ then $S(X)$ is sufficient.

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

Definition: $T(X)$ is minimal sufficient if:
1) $T(X)$ is sufficient
2) For any other sufficient $S(X)$,
   $T(X) = f(S(X))$ for some $f$ (a.s. in $P$)

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

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

function of $\theta$ with parameter $\theta$

function of $x$ determines which function

data $X$ determines which function

The **log-likelihood function** is its log:

$$l(\theta; x) = \log \text{Lik}(\theta; x)$$

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

If $T(X)$ is sufficient then

$$\text{Lik}(\theta; x) = q_{\theta}(T(x)) \ h(x)$$

$T$ determines the scaling

"shape"

HW 2: Likelihood ratios

$$\left( \frac{\text{Lik}(\theta_1; x)}{\text{Lik}(\theta_2; x)} \right)_{\theta_1, \theta_2 \in \Theta}$$
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

(don’t forget to check!)

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

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

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

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

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

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

\( \Rightarrow \) order stats are minimal suff.
Minimal sufficiency for exp. fam.s

Suppose \( P_\xi(x) = e^{\xi' T(x) - A(\xi)} h(x) \)

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

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

Suppose \( x \) and \( \gamma \) give same likelihood shape:
\[ l(\gamma; x) = l(\gamma; y) = \text{const}(x, y) \]

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

\[ \Rightarrow \quad T(x) = T(y) \quad \text{or} \]
\[ T(x) - T(\gamma) \perp \text{Span} \{ \xi, -\xi_2 : \xi \in \Xi \} \]

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

(That is, if \( \Xi \) 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

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

\[ \rho_\theta(x) = e^{\gamma(\theta)^T(x) - 2(\theta) \cdot h(x)} \quad \theta \in \Theta \]

\[ T(X) \text{ minimal if } \operatorname{Span} \{ \gamma(\theta_1) - \gamma(\theta_2) : \theta_1, \theta_2 \in \Theta \} = \mathbb{R}^3 \]