Model: A family of distributions

$$\{ P_\theta : \theta \in \Theta \}.$$

$P_\theta(B)$ is prob. of event B when the parameter takes the value θ.

P_θ is described by giving a (joint) pdf or pmf $f(x|\theta)$.

Experiment: Observe $\underset{\text{(data)}}{X} \sim P_\theta$, θ unknown.

Goal: Make inference about θ.

Joint distn. of independent rv's:

If $X = (X_1, \ldots, X_n)$
and $X_1, \ldots, X_n$ are independent
with $X_i \sim$ pdf $g_i(x_i|\theta)$,
then joint pdf is

$$\underset{\downarrow \atop (x_1,\ldots,x_n)}{f(x|\theta)} = \prod_{i=1}^{n} g_i(x_i|\theta)$$

For iid rv's $g_1 = \cdots = g_n \equiv g$.

Types of Models (in this course)

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

1. Random Sample
   $X_1, \ldots, X_n$ are iid

2. Regression model
   $X_1, \ldots, X_n$ are independent
   (but not necessarily identically distributed)
   (The distn. of X_i may depend on covariates z_i.)

Models with Correlated Data

(not covered)

Time Series (e.g. ARMA)
Spatial and Space-Time data
Repeated measures
etc.

Random Sample Models

Example:

$X_1, X_2, \ldots, X_n$ iid Poisson(λ), λ unknown.

Here we have:

$X = (X_1, X_2, \ldots, X_n)$

$\theta = \lambda$

$\Theta = \{\lambda : \lambda > 0\}$

P_θ is described by the joint pmf

$$f(x|\lambda) = f(x_1, \ldots, x_n|\lambda)$$

$$= \prod_{i=1}^{n} g(x_i|\lambda)$$

where g is the Poisson(λ) pmf

$$g(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for $x = 0, 1, 2, \ldots$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \quad \text{for } x \in \{0, 1, 2, \ldots\}^n$$

($= 0$ otherwise).

Example:

$X_1, X_2, \ldots, X_n$ iid $N(\mu, \sigma^2)$ with μ, σ^2 unknown.

Here we have:

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

$\theta = (\mu, \sigma^2)$

$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$

P_θ is described by the joint pdf

$$f(x|\mu,\sigma^2) = \prod_{i=1}^{n} g(x_i|\mu,\sigma^2)$$

$\hookrightarrow = (x_1, \ldots, x_n)$

$$\left(\begin{array}{l}\text{where } g \text{ is the } N(\mu,\sigma^2) \text{ pdf} \\ g(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\end{array}\right)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

Sufficient Statistics

$X \sim P_\theta$, θ unknown.

What part (or function) of the data X is essential for inference about θ?

Example

Suppose $X_1, \dots, X_n$ iid Bernoulli(p).
(independent tosses of a coin)

Intuitively
$$T = \sum_{i=1}^{n} X_i = \text{\# of heads}$$
contains all the info. about p in data.

We need to formalize this.

Definition: ($X \sim P_\theta$, θ unknown)

The statistic $T = T(X)$ is a ***sufficient statistic*** for θ if the conditional distn. of X given T does *not* depend on the unknown parameter θ.

Abbrev: T is SS if $\mathcal{L}(X|T)$ is same for all θ.

Motivation for the Definition of a Sufficient Statistic

Suppose $X \sim P_\theta$, $\theta \in \Theta$, θ unknown.
Let $T = T(X)$ be any statistic.

We can imagine that the data X is generated hierarchically:

① First generate $T \sim \mathcal{L}(T)$.

② Then generate $X \sim \mathcal{L}(X \mid T)$.

If T is a suff. stat. for θ, then $\mathcal{L}(X \mid T)$ does <u>not</u> depend on θ and step ② can be carried out without knowing θ.

Since, given T, the data X can be generated with<u>out</u> knowing θ, the data X supplies no further information about θ beyond what is already contained in T.

<u>Notation</u>: $X \sim P_\theta$, $\theta \in \Theta$, θ unknown.

If $T = T(X)$ is suff. stat. for θ, then
T contains all the info. about θ in X
in this sense:

If X is discarded, but we keep $T = T(X)$,
we can "fake" the data (with<u>out</u> knowing θ)
by generating X^* from $\mathcal{L}(X|T)$.

X^* has the same distn. as X $(X^* \sim P_\theta)$
and the same value of the sufficient statistic
$(T(X^*) = T(X))$
and can be used for any purpose we would
use the real data for.

Example: If $U(X)$ is an estimator of θ,
then $U(X^*)$ is another estimator of θ
which performs just as well since

$$U(X) \overset{d}{=} U(X^*) \quad \text{for all } \theta.$$

<u>Cautionary Remark</u>: $X \sim P_\theta$

<u>If the model is correct</u> and $T(X)$ is
sufficient for θ, then can ignore data X
and just use $T(X)$ for inference about θ.

<u>BUT</u> if we are not sure the model is correct, X may contain valuable information about model correctness not contained in $T(X)$.

Example:

$X_1, X_2, \ldots, X_n$ iid Bernoulli(p)

$T = \sum X_i$ is suff. stat. for p.

Possible model violations:

Trials might be correlated.

p (Prob. success) might not be constant.

etc.

These model violations can<u>not</u> be investigated using the suff. stat. You need the data.

Example: $X = (X_1, X_2)$ i.i.d. Poisson(λ).

$T = X_1 + X_2$ is a sufficient statistic for λ.

$$P_\lambda(X_1 = x_1, X_2 = x_2 \mid T = t)$$

$$= \frac{P_\lambda(X_1 = x_1, X_2 = x_2, T = t)}{P_\lambda(T = t)}$$

(redundant if $t = x_1 + x_2$)

$$= \begin{cases} \dfrac{P_\lambda(X_1 = x_1, X_2 = x_2)}{P_\lambda(T = t)} & \text{if } t = x_1 + x_2 \\ 0 & \text{if } t \neq x_1 + x_2 \end{cases}$$

Remark: In general (discrete) case

$$P_\theta(X = x \mid T(X) = t)$$

$$= \begin{cases} \dfrac{P_\theta(X = x)}{P_\theta(T(X) = t)} & \text{if } T(x) = t \\ 0 & \text{otherwise} \end{cases}$$

(Assume $t = x_1 + x_2$.)

$$= \frac{\dfrac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \dfrac{\lambda^{x_2} e^{-\lambda}}{x_2!}}{\dfrac{(2\lambda)^t e^{-2\lambda}}{t!}} \quad \text{(since } T \sim \text{Poisson}(2\lambda)\text{)}$$

$$= \frac{\lambda^t e^{-2\lambda} / x_1!\, x_2!}{2^t \lambda^t e^{-2\lambda}/t!} = \binom{t}{x_1} \frac{1}{2^t}$$

which does <u>not</u> involve λ.
Thus, T is suff. stat. for λ.

Note that

$$P(X_1 = x_1 | T = t) = \binom{t}{x_1} \left(\frac{1}{2}\right)^{x_1} \left(\frac{1}{2}\right)^{t - x_1}$$

for $x_1 = 0, 1, \ldots, t$.

Thus $\mathcal{L}(X_1 | T = t)$ is Binomial$(t, \frac{1}{2})$.

Given $T = t$, we may generate fake data (X_1^*, X_2^*) (without knowing λ) which has the same distn. as the real data:

① Generate $X_1^* \sim \text{Binomial}(t, \frac{1}{2})$.

(Toss a fair coin t times.
Count the number of heads.)

② Set $X_2^* = t - X_1^*$.

The real and fake data have the same value of the sufficient statistic:

$$X_1 + X_2 = t = X_1^* + X_2^*.$$

Extension: If $X = (X_1, X_2, \ldots, X_n)$ are i.i.d. Poisson(λ), then $T = X_1 + X_2 + \cdots + X_n$ is a suff. stat. for λ.

Moreover,

$$P(X_1 = x_1, \ldots, X_n = x_n \mid T = t) = \frac{t!}{x_1! \cdots x_n!} \left(\frac{1}{n}\right)^t$$

$$= \binom{t}{x_1, \ldots, x_n} \left(\frac{1}{n}\right)^{x_1} \cdots \left(\frac{1}{n}\right)^{x_n}$$

so that $\mathcal{L}(X \mid T = t)$ is Multinomial with t trials and n categories with equal probability $\frac{1}{n}$. (See Section 4.6.)

Example: $X = (X_1, X_2)$ i.i.d. Expon (β).
$T = X_1 + X_2$ is suff. stat. for β.

We need $\mathcal{L}(X_1, X_2 \mid T = t)$.

Suffices to get $\mathcal{L}(X_1 \mid T = t)$.

How to do this?

① Find joint density $f_{X_1,T}(x_1, t)$.

② Then get conditional density

$$f_{X_1|T}(x_1|t) = \frac{f_{X_1,T}(x_1,t)}{f_T(t)}.$$

Step ①: Use the transformation

$$\begin{aligned} U &= X_1 \\ T &= X_1 + X_2 \end{aligned} \quad \Rightarrow \quad \begin{aligned} X_1 &= U \\ X_2 &= T - U \end{aligned}$$

with Jacobian $J = 1$

$$f_{U,T}(u,t) = f_{X_1,X_2}(u, t-u)\,|J|$$

$$= \frac{1}{\beta} e^{-u/\beta} \cdot \frac{1}{\beta} e^{-(t-u)/\beta} \cdot 1$$

$$= \frac{1}{\beta^2} e^{-t/\beta} \quad \text{for } 0 \le u \le t < \infty .$$

Step ②:

$T = X_1 + X_2 \sim \text{Gamma}(\alpha = 2, \beta)$

so that $f_T(t) = \dfrac{t e^{-t/\beta}}{\beta^2}$ for $t \geq 0$.

Alternative integrate over x_1 in the joint density $f_{X_1,T}(x_1, t)$ to get $f_T(t)$.

Now

$$f_{X_1|T}(x_1|t) = \frac{\frac{1}{\beta^2} e^{-t/\beta} \, I(0 \leq x_1 \leq t)}{\frac{t e^{-t/\beta}}{\beta^2}}$$

$$= \frac{1}{t} I(0 \leq x_1 \leq t)$$

which does <u>not</u> involve β.

Thus $T = X_1 + X_2$ is suff. stat. for β.

Moreover, $\mathcal{L}(X_1 | T = t)$ is Uniform$(0, t)$.

(This can be seen intuitively by noting that $f_{X_1,X_2}(x_1, x_2) = \frac{1}{\beta^2} e^{-(x_1+x_2)/\beta}$ is <u>constant</u>)

(on the line segment
$$\{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0, x_1 + x_2 = t\}.)$$

Thus, given $T = t$, we may generate fake data (X_1^*, X_2^*) (without knowing β) which has the same distn. as the real data:

① Generate $X_1^* \sim \text{Uniform}(0, t)$.
② Set $X_2^* = t - X_1^*$.

The real and fake data have the same value of the suff. stat.: $X_1 + X_2 = t = X_1^* + X_2^*$.

Extension: If $X = (X_1, \ldots, X_n)$ iid $\text{Expon}(\beta)$, then $T = X_1 + \cdots + X_n$ is suff. stat. for β and $\mathcal{L}(X \mid T = t)$ is a uniform distn. on the simplex
$$\{(x_1, \ldots, x_n) : x_1 + \cdots + x_n = t \text{ and } x_i \geq 0 \ \forall i\}.$$

Example: $X=(X_1,X_2)$ iid Uniform$(0,\theta)$.

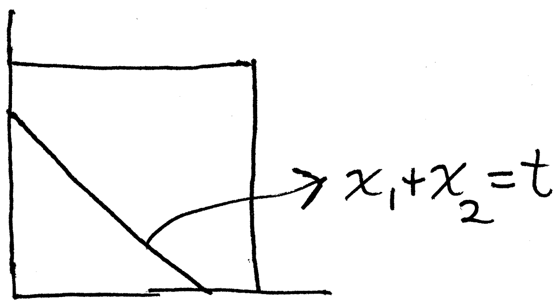
$T=X_1+X_2$ is <u>not</u> suff. stat. for θ.

Proof: We must show that $\mathcal{L}(X_1,X_2\mid T)$ depends on θ.

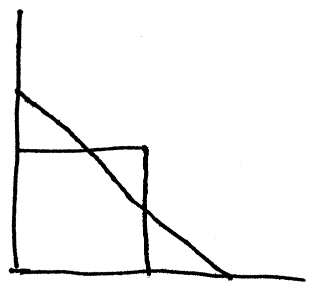
The support of (X_1,X_2) is $[0,\theta]^2$.

Given $T=t$, we know (X_1,X_2) lies on the line $\mathcal{L}=\{(x_1,x_2): x_1+x_2=t\}$.

Thus, the support of $\mathcal{L}(X_1,X_2\mid T=t)$ is $\mathcal{L}\cap[0,\theta]^2$ which is drawn below for two different values of θ.



Case: $\theta>t$



Case: $\theta<t$

The support of $\mathcal{L}(X_1,X_2\mid T=t)$ varies with θ. This shows that $\mathcal{L}(X_1,X_2\mid T=t)$ depends on θ.

Example:

If $X_1, \ldots, X_n$ iid Bernoulli(p), then $T = \sum_{i=1}^{n} X_i$ is suff. stat. for p.

First: What is joint pmf of $X = (X_1, \ldots, X_n)$?

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0)$$
$$= p \cdot q \cdot p \cdot p \cdot q = p^3 q^2$$

where $q = 1 - p$.

In general,

$$P(X = x) = P(X_1 = x_1, \ldots, X_n = x_n)$$
$$= \prod_{i=1}^{n} \underbrace{p^{x_i} q^{1-x_i}}_{\substack{= p \text{ if } x_i = 1 \\ = q \text{ if } x_i = 0}} = p^{\sum x_i} q^{\sum (1 - x_i)}$$

$$= p^t q^{n-t} = p^{T(x)} q^{n - T(x)}$$

where $T(x) = t = \sum_{i=1}^{n} x_i$.

What is $\mathcal{L}(X|T)$?

Notation: $T(X) = \sum_{i=1}^{n} X_i = T$

$T(x) = \sum_{i=1}^{n} x_i$

Recall that $P_\theta(X = x \mid T(X) = t) =$

$$\begin{cases} \dfrac{P_\theta(X=x)}{P_\theta(T(X)=t)} & \text{if } T(x) = t \\ 0 & \text{if } T(x) \neq t \end{cases}$$

Assume $T(x) = \sum x_i = t$. $\theta = p$.

Then

$$P_\theta(X=x \mid T=t) = \frac{P_\theta(X=x)}{P_\theta(T=t)}$$

$$= \frac{p^t q^{n-t}}{\binom{n}{t} p^t q^{n-t}} = \frac{1}{\binom{n}{t}}$$

since $T \sim \text{Binomial}(n, p)$.

This does not involve p which proves that T is a sufficient statistic for p.

Note: The conditional probability is the same for any sequence $x = (x_1, \ldots, x_n)$ with t 1's and $n-t$ 0's. There are $\binom{n}{t}$ such sequences.

Summary: Given $T = X_1 + \cdots + X_n = t$, all possible sequences of t 1's and $n-t$ 0's are equally likely.

Algorithm for generating from $\mathcal{L}(X_1, \ldots, X_n \,|\, T = t)$:

> Put t 1's and $n-t$ 0's in an urn.
>
> Draw them out one by one (without replacement) until the urn is empty.

This makes all possible sequences equally likely. (Think about it!)

The resulting sequence $(X_1^*, \ldots, X_n^*)$ (the fake data) has the same value of the sufficient statistic as $(X_1, \ldots, X_n)$:

$$\sum_i X_i^* = t = \sum_i X_i \,.$$

Suppose $X \sim P_\theta$, $\theta \in \Theta$.

Theorem 6.2.2

$T(X)$ is a suff. stat. for θ

iff for all x, $\dfrac{f_X(x|\theta)}{f_T(T(x)|\theta)}$ is constant

as a function of θ.

Notation: $f_X(x|\theta)$ is pdf (or pmf) of X.

$f_T(t|\theta)$ is pdf (or pmf) of $T = T(X)$.

Factorization Criterion (FC)

There exist functions $h(x)$ and $g(t|\theta)$ such that

$$f(x|\theta) = g(T(x)|\theta)\, h(x)$$

for all x and θ.

Theorem: $T(X)$ is a suff. stat. for θ iff the factorization criterion is satisfied.

Example: Application of Factorization Criterion to random sample from Poisson distribution.

$X = (X_1, \ldots, X_n)$, iid Poisson(λ).

Joint pmf is

$$
\begin{aligned}
f(x \,|\, \lambda) &= f(x_1, \ldots, x_n \,|\, \lambda) \\
&= \prod_{i=1}^{n} \frac{\lambda^{x_i} \, e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_i x_i} \, e^{-n\lambda}}{\prod_i x_i!} \\
&= \left(\lambda^{\sum_i x_i} e^{-n\lambda} \right) \left(\frac{1}{\prod_i x_i!} \right) \\
&= g(T(x) \,|\, \lambda) \, h(x) \quad \text{where} \\
& \qquad T(x) = \textstyle\sum_i x_i \, , \; g(t \,|\, \lambda) = \lambda^t e^{-n\lambda} \, , \; h(x) = \dfrac{1}{\prod_i x_i!} \, .
\end{aligned}
$$

Thus (by FC) $T(X) = \sum_i X_i$ is a sufficient statistic for λ.

This is much easier than proving sufficiency directly from the definition.