1 Probability Model

Model: A family of distributions $\{P_\theta : \theta \in \Theta\}$.
$P_\theta(B)$ is the probability of the event B when the parameter takes the value θ.
P_θ is described by giving a joint pdf or pmf $f(x \mid \theta)$.
Experiment: Observe $X(\text{data}) \sim P_\theta$, θ unknown.
Goal: Make inference about θ.
Joint distribution of independent rv's: If $X = (X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ are independent with $X_i \sim g_i(x_i \mid \theta)$, then the joint pdf is $f(x \mid \theta) = \prod_{i=1}^n g_i(x_i \mid \theta)$ where $x = (x_1, \ldots, x_n)$. For iid random variables $g_1 = \cdots = g_n = g$.

1.1 Types of models to be discussed in the course

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

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 distribution of X_i may depend on covariates z_i)

1.1.1 Random Sample Models

Example: Let $X_1, X_2, \ldots, X_n$ 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

$$f(x \mid \lambda) = f(x_1, \ldots, x_n \mid \lambda) = \prod_{i=1}^n g(x_i \mid \lambda)$$

where g is the Poisson(λ) pmf $g(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0, 1, 2, \ldots$. Hence

$$f(x \mid \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

for $x \in \{0, 1, 2, \ldots\}^n$.
Example: Let $X_1, X_2, \ldots, X_n$ iid N(μ, σ^2), with μ and σ^2 unknown. Here we have: $X =$

$(X_1, X_2, \ldots, X_n)$, $\theta = (\mu, \sigma^2)$, $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$, P_θ is described by the joint pmf

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

where g is the $\mathrm{N}(\mu, \sigma^2)$ pdf $g(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$. Hence

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

2 Sufficient Statistic

Let $X \sim P_\theta, \theta$ unknown. What part (or function) of the data X is essential for inference about θ?
Example: Suppose $X_1, \ldots, X_n$ iid Bernoulli(p) (independent tosses of a coin). Intuitively,

$$T = \sum_{i=1}^{n} X_i = \#\text{ of heads}$$

contains all the information about p in the data. We need to formalize this.
Let $X \sim P_\theta, \theta$ unknown.

Definition 1. *The statistic $T = T(X)$ is a sufficient statistic for θ if the conditional distribution of X given T does not depend on the unknown parameter θ.*
Abbreviation: T is SS if $\mathcal{L}(X \mid T)$ is same for all θ, where $\mathcal{L}$ stands for law *or* distribution.

2.1 Motivation for the definition

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 as follows:

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

If T is a sufficient statistic for θ, then $\mathcal{L}(X \mid T)$ does not depend on θ and Step 2 can be carried out without knowing θ. Since, given T, the data X can be generated without knowing θ, the data X supplies no further information about θ beyond what is already contained in T.

Notation: $X \sim P_\theta$, $\theta \in \Theta$, θ unknown. If $T = T(X)$ is a sufficient statistic for θ, then T contains all the information about θ in X in the sense that if X is discarded, but we keep $T = T(X)$, we can "fake" the data (without knowing θ) by generating X^* from $\mathcal{L}(X \mid T)$. X^* has the same distribution 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 $U(X) \stackrel{d}{=} U(X^*)$ for all θ.
Cautionary Note: If the model is correct ($X \sim P_\theta$) and $T(X)$ is sufficient for θ, then can ignore data X and just use $T(X)$ for inference about θ. BUT if we are not sure that 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). $T = \sum_{i=1}^n X_i$ is a sufficient statistic for p.
Possible Model violations: The trial might be correlated as not independent. The success probability p might not be constant from trial to trial. These model violations cannot be investigated using the sufficient statistic. This can be only done by further investigation with the data.

2.2 Examples of Sufficient Statistic

1. $X = (X_1, X_2) \sim$ iid Poisson(λ). $T = X_1 + X_2$ is a sufficient statistic for λ because

$$\begin{aligned} P_\lambda(X_1 = x_1, X_2 = x_2 \mid T = t) &= \frac{P_\lambda(X_1 = x_2, X_2 = x_2, \overbrace{T = t}^{\text{redundant if } t = x_1 + x_2})}{P_\lambda(T = t)} \\ &= \begin{cases} \frac{P_\lambda(X_1 = x_2, X_2 = x_2)}{P_\lambda(T = t)}, & \text{if } t = x_1 + x_2 \\ 0 \quad \text{if } t \neq x_1 + x_2 \end{cases} \end{aligned}$$

   This follows from the fact that for discrete distributions P_θ,

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

   Assuming $t = x_1 + x_2$,

$$\begin{aligned} P_\lambda(X_1 = x_1, X_2 = x_2 \mid T = t) &= \frac{\frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!}}{\frac{(2\lambda)^t e^{-2\lambda}}{t!}} (\text{Since } T \sim \text{Poisson}(2\lambda)) \\ &= \frac{\binom{t}{x_1}}{2^t} \end{aligned}$$

   which does not involve λ. Thus, T is a sufficient statistic for λ. Note that

$$P(X_1 = x_1 \mid T = t) = \binom{t}{x_1}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{2}\right)^{t-x_1}, x_1 = 0, 1, \ldots, t.$$

Thus $\mathcal{L}(X_1 \mid T = t)$ is Binomial$(t, 1/2)$. Given $T = t$, we may generate fake data X_1^*, X_2^* without knowing λ which has the same distribution as the real data:

(a) Generate $X_1^* \sim$ Binomial$(t, 1/2)$. (Toss a fair coin t times and count the number of heads).

(b) Set $X_2^* = t - X_1^*$.

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

2. Extension to previous Example: If $X = (X_1, X_2, \ldots, X_n)$ are iid Poisson(λ), then $T = X_1 + X_2 + \cdots + X_n$ is a sufficient statistic for λ. Moreover

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n \mid T = t) &= \frac{t!}{x_1! x_2! \cdots x_n!}\left(\frac{1}{n}\right)^t \\ &= \binom{t}{x_1, \ldots, x_n}\left(\frac{1}{n}\right)^{x_1} \cdots \left(\frac{1}{n}\right)^{x_n} \end{aligned}$$

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

3. $X = (X_1, X_2)$ iid Expo(β). Then $T = X_1 + X_2$ is a sufficient statistic for β.
To derive this, we need to calculate $\mathcal{L}(X_1, X_2 \mid T = t)$. It suffices to get $\mathcal{L}(X_1 \mid T = t)$ since $X_2 = t - X_1$. How to do this?

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

(b) Then get conditional density

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

Continuing with the steps,

(a) Use the transformation

$$U = X_1, T = X_1 + X_2 \quad \Rightarrow \quad X_1 = U, X_2 = T - U$$

with Jacobian $J = 1$. Then

$$\begin{aligned} f_{U,T}(u,t) &= f_{X_1,X_2}(u, t-u)|J| \\ &= \frac{1}{\beta}e^{-u/\beta} \cdot \frac{1}{\beta}e^{-(t-u)/\beta} \cdot 1 \\ &= \frac{1}{\beta^2}e^{-t/\beta}, \quad \text{for} \quad 0 \le u \le t < \infty. \end{aligned}$$

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

$$f_T(t) = \frac{te^{-t/\beta}}{\beta^2}, \quad t \ge 0.$$

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

$$\begin{aligned} f_{X_1|T}(x_1 \mid t) &= \frac{\frac{1}{\beta^2}e^{-t/\beta}I(0 \le x_1 \le t)}{\frac{te^{-t/\beta}}{\beta^2}} \\ &= \frac{1}{t}I(0 \le x_1 \le t) \end{aligned}$$

which does <u>not</u> involve β.

Thus $T = X_1 + X_2$ is a sufficient statistic for β.
Moreover, $\mathcal{L}(X_1 \mid T = t)$ is Unif$(0, t)$. This can also 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 constant on the line segment

$$\{(x_1, x_2) : x_1 \ge 0, x_2 \ge 0, x_1 + x_2 = t\}$$

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

(a) Generate $X_1^* \sim \text{Unif}(0, t)$.
(b) Set $X_2^* = t - X_1^*$.

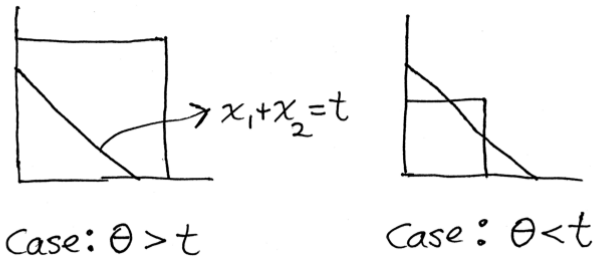
The real and fake data have the same value of the sufficient statistic: $X_1 + X_2 = t = X_1^* + X_2^*$.

4. <u>Extension to previous Example:</u> If $X = (X_1, X_2, \ldots, X_n)$ are iid Expo(β), then $T = X_1 + X_2 + \cdots + X_n$ is a sufficient statistic for β and $\mathcal{L}(X \mid T = t)$ is a uniform distribution on the simplex

$$\{(x_1, \ldots, x_n) : x_1 + \cdots + x_n = t, x_i \ge 0 \,\forall\, i\}.$$

5. $X = (X_1, X_2)$ iid Unif$(0, \theta)$. Then $T = X_1 + X_2$ is <u>not</u> sufficient statistic 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 $\mathcal{L}(X_1, X_2 \mid T)$ is $\mathcal{L} \cap [0, \theta]^2$ which is drawn below for two different values of θ. The support of $\mathcal{L}(X_1, X_2 \mid T = t)$ varies with θ. This shows



   that $\mathcal{L}(X_1, X_2 \mid T)$ depends on θ. □

6. If $X_1, \ldots, X_n$ iid Bernoulli(p), then $T = \sum_{i=1}^n X_i$ is a sufficient statistic for p.
   First: What is the joint pmf of $X_1, \ldots, X_n$? Note that

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

   where $q = 1 - p$. In general,

$$\begin{aligned} P(X = x) = P(X_1 = x_1, \ldots, X_n = x_n) &= \prod_{i=1}^n p^{x_i} q^{1-x_i} = p^{\sum_{i=1}^n x_i} q^{\sum_{i=1}^n (1-x_i)} \\ &= p^t q^{n-t} = p^{T(x)} q^{n-T(x)}, \end{aligned}$$

   where $T(x) = t = \sum_{i=1}^n x_i$. Next, we derive $\mathcal{L}(X \mid T)$. We will use the notation $T(X) = \sum_{i=1}^n X_i = T$ and $T(x) = \sum_{i=1}^n x_i$. Recall that for discrete distributions P_θ,

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

   Assume $T(x) = \sum_{i=1}^n x_i = t, \theta = p$. Then

$$\begin{aligned} P_\theta(X = x \mid T(X) = t) &= \frac{P_\theta(X = x)}{P_\theta(T(X) = t)} \\ &= \frac{p^t q^{n-t}}{\binom{n}{t} p^t q^{n-t}} = \frac{1}{\binom{n}{t}} \end{aligned}$$

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 1s and $n - t$ 0s. There are $\binom{n}{t}$ such sequences.
Summary: Given $T = X_1 + \cdots + X_n = t$, all possible sequences of t 1s and $n - t$ 0s are equally likely.
Algorithm for generating from $\mathcal{L}(X_1, \ldots, X_n \mid T = t)$:

(a) Put t 1s and $n - t$ 0s in an urn.

(b) 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 $(X_1, \ldots, X_n)$:

$$\sum_{i=1}^{n} X_i^* = t = \sum_{i=1}^{n} X_i$$

2.3 Sufficient conditions for sufficiency

Sometimes finding sufficient statistic might be time-consuming and cumbersome if one proceeds directly from the definition. We need an easy to verifiable sufficient condition to find a sufficient statistic. Suppose $X \sim P_\theta, \theta \in \Theta$.
Theorem 6.2.2
T(X) is a sufficient statistic for θ iff for all x

$$\frac{f_X(x \mid \theta)}{f_T(T(x) \mid \theta)}$$

is constant as a function of θ.
Notation: $f_X(x \mid \theta)$ is pdf (or pmf) of X. $f_T(t \mid \theta)$ is pdf (or pmf) of $T = T(X)$.
Factorization Criterion (FC): There exist functions $h(x)$ and $g(t \mid \theta)$ such that

$$f(x \mid \theta) = g(T(x) \mid \theta)h(x)$$

for all x and θ.

Theorem 1. *$T(X)$ is a sufficient statistic for θ iff the factorization criterion is satisfied.*

2.4 Applications of FC

1. Let $X = (X_1, \ldots, X_n)$ iid Poisson(λ). The joint pmf is

$$\begin{aligned} f(x \mid \lambda) &= f(x_1, \ldots, x_n \mid \lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_i x_i} e^{-n\lambda}}{\prod_i x_i!} \\ &= \left(\lambda^{\sum_i x_i} e^{-n\lambda}\right)\left(\frac{1}{\prod_i x_i!}\right) \\ &= g(t(x) \mid \lambda) h(x) \end{aligned}$$

   where $T(x) = \sum_i x_i$, $g(t \mid \lambda) = \lambda^t e^{-n\lambda}, h(x) = \frac{1}{\prod_i x_i!}$ Thus, by FC, $T(X) = \sum_i X_i$ is a sufficient statistic for λ.