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(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 $U(X) \overset{d}{=} U(X^*)$ 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

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

This follows from the fact that for discrete distributions P_θ,

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

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 $\text{Unif}(0,t)$. 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 $X_2^* = t - X_1^*$.

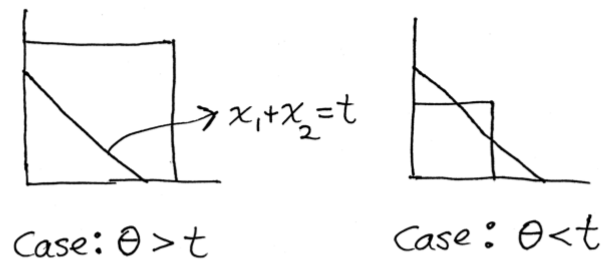
The real and fake data have the same value of the sufficient statistic: $X_1 + X_2 = t = X_1^* + X_2^*$.

4. <u>Extension to previous Example:</u> If $X = (X_1, X_2, \ldots, X_n)$ are iid $\text{Expo}(\beta)$, 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. $\square$

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,

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

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.*

Proof. (When X is discrete)
Notation: $T = T(X), t = T(x)$.

First, Assume T is a sufficient statistic for θ. Then the pmf $f(x \mid \theta)$ can be written as

$$
\begin{aligned}
f(x \mid \theta) &= \underbrace{P_\theta(T=t)}_{\text{This is a function of } t \text{ and } \theta. \text{ Call it } g(t \mid \theta)} \cdot \underbrace{P_\theta(X=x \mid T=t)}_{\text{This depends on x, but not } \theta \text{ (by defn. of suff. stat. Call it } h(x)} \\
&= g(t \mid \theta)h(x).
\end{aligned}
$$

Hence FC is true.
Next Assume FC is true.
Then

$$
\begin{aligned}
P_\theta(X=x \mid T=t) &= \frac{P_\theta(X=x)}{P_\theta(T=t)} \quad (\text{since } \{X=x\} \subset \{T=t\}) \\
&= \frac{f(x \mid \theta)}{\sum_{z:T(z)=t} f(z \mid \theta)} = \frac{g(t \mid \theta)h(x)}{\sum_{z:T(z)=t} g(t \mid \theta)h(z)} \\
&= \frac{h(x)}{\sum_{z:T(z)=t} h(z)}
\end{aligned}
$$

which does not involve θ. □

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.

2. Simple Linear Regression: Let

$$X_i = \beta_0 + \beta_1 z_i + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma_0^2) \quad i = 1, \ldots, n$$

where $z_i, i = 1, \ldots, n$ are known constants.
Alternative statement of the model:

$$
\begin{aligned}
X_1, X_2, \ldots, X_n &\quad \text{independent} \\
X_i &\sim \mathrm{N}(\beta_0 + \beta_1 z_i, \sigma_0^2).
\end{aligned}
$$

Data is $X = (X_1, X_2, \ldots, X_n)$. $(z_1, z_2, \ldots, z_n)$ are known constants. Unknown parameter is $\theta = (\beta_0, \beta_1) \in \mathbb{R}^2$. What are the sufficient statistics for this model? Use FC.

$$
\begin{aligned}
f(x \mid \theta) &= \prod_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_i-\beta_0-\beta_1 z_i)^2/2\sigma_0^2}}_{\mathrm{N}(\beta_0+\beta_1 z_i, \sigma_0^2)\text{ density}}. \\
&= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} \underbrace{(x_i - \beta_0 - \beta_1 z_i)^2}_{S} \right\}.
\end{aligned}
$$

Here

$$
\begin{aligned}
S &= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 z_i) + \sum_{i=1}^{n} (\beta_0 + \beta_1 z_i)^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2\beta_0 \sum_{i=1}^{n} x_i - 2\beta_1 \sum_{i=1}^{n} x_i z_i + \sum_{i=1}^{n} (\beta_0 + \beta_1 z_i)^2.
\end{aligned}
$$

Plus this back into the exponential and rearrange to get

$$
\begin{aligned}
f(x \mid \theta) &= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left\{ -\frac{1}{2\sigma_0^2}\left(-2\beta_0 \sum_{i=1}^{n} x_i - 2\beta_1 \sum_{i=1}^{n} x_i z_i + \sum_{i=1}^{n} (\beta_0 + \beta_1 z_i)^2 \right) \right\} \\
\times \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} x_i^2 \right\} \\
&= g\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i z_i, \beta_0, \beta_1 \right) h(x) \\
&= g(T(x), \theta) h(x)
\end{aligned}
$$

where $T(x) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i z_i)$ and

$$
g(t, \theta) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left\{ -\frac{1}{2\sigma^2}\left(-2\beta_0 t_1 - 2\beta_1 t_2 + \sum_{i=1}^{n} (\beta_0 + \beta_1 z_i)^2 \right) \right\}
$$

with $t = (t_1, t_2)$ and $h(x) = \exp\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} x_i^2 \}$.

3. <u>Continuation of Simple Linear Regression Example:</u> What if the variance σ^2 is unknown? Now $\theta = (\beta_0, \beta_1, \sigma^2)$ and $\Theta = \mathbb{R}^2 \times (0, \infty)$. (Change σ_0^2 to σ^2 in the earlier

formulas to indicate this). Now $\exp\big\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\big\}$ is not a function of x, but depends also on θ. So we now factor the joint density as

$$\begin{aligned} f(x \mid \theta) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\beta_0\sum_{i=1}^n x_i - 2\beta_1\sum_{i=1}^n x_i z_i + \sum_{i=1}^n(\beta_0+\beta_1 z_i)^2\right)\right\}\cdot 1. \\ &= g(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i, \sum_{i=1}^n x_i z_i, \beta_0, \beta_1, \sigma^2)h(x) \\ &= g(T(x), \theta)h(x) \end{aligned}$$

where

$$\begin{aligned} T(x) &= \left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i, \sum_{i=1}^n x_i z_i\right) = (t_1, t_2, t_3) \\ g(t, \theta) &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left(t_1 - 2\beta_0 t_2 - 2\beta_1 t_3 + \sum_{i=1}^n(\beta_0+\beta_1 z_i)^2\right)\right\} \end{aligned}$$

and $h(x) = 1$. According to FC, $T(X) = (\sum_i X_i^2, \sum_i X_i, \sum_i z_i X_i)$ is a sufficient statistic for $\theta = (\beta_0, \beta_1, \sigma^2)$.

4. <u>Discussion on the preceding examples:</u> We have described two models. The model with σ^2 known (i.e., $\sigma^2 = \sigma_0^2$) can be regarded as a subset of the model where σ^2 is unknown.

$$\begin{aligned} \Theta_1 &= \{(\beta_0, \beta_1, \sigma^2) : \sigma^2 = \sigma_0^2\} = \mathbb{R}^2 \times \{\sigma_0^2\}. \\ \Theta_2 &= \{(\beta_0, \beta_1, \sigma^2) : \sigma^2 > 0\} = \mathbb{R}^2 \times (0, \infty). \end{aligned}$$

$\Theta_1 \subset \Theta_2$. The sufficient statistics we found for these two models were different:

$$\begin{aligned} T_1 &\equiv (\sum_i X_i, \sum_i z_i X_i) \quad \text{is SS for } \Theta_1. \\ T_2 &\equiv (\sum_i X_i^2, \sum_i X_i, \sum_i z_i X_i) \quad \text{is SS for } \Theta_2. \end{aligned}$$

<u>Note:</u> T_2 is also a SS for Θ_1, but it is not "minimal".

5. <u>Sufficient statistic for random samples from various families of normal distributions:</u> Let $X = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are iid $\mathrm{N}(\mu, \sigma^2)$. Consider different families of normal distributions.

$$\begin{aligned} \Theta_1 = \{(\mu, \sigma^2) : \sigma^2 > 0\} &\quad \text{(all normal distributions)} \\ \Theta_2 = \{(\mu, \sigma^2) : \sigma^2 = \sigma_0^2\} &\quad \text{(known variance)} \\ \Theta_3 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\} &\quad \text{(known mean)} \end{aligned}$$

For each space, the "obvious" sufficient statistic is different. In all case, the joint pdf of X is given by

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (x_i-\mu)^2 \right\} \end{aligned} \tag{1}$$

$\underline{\Theta_3:}$ Here $\mu = \mu_0$, (a known value), so the "unknown" parameter is $\theta = \sigma^2$. The joint pdf may be factored as

$$\begin{aligned} f(x \mid \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (x_i-\mu_0)^2 \right\} \\ &= g\big(\sum_i (x_i-\mu_0)^2, \sigma^2\big) h(x) \\ &= g\big(T_3(x), \sigma^2\big) h(x), \end{aligned}$$

where $T_3(x) \equiv \sum_{i=1}^{n} (x_i-\mu_0)^2$ so that $T_3 = T_3(X) \equiv \sum_i (X_i-\mu_0)^2$ is a SS for Θ_3. Note: T_3 is not even a statistic if μ is unknown (i.e., not fixed). For the rest (Θ_1 and Θ_2), we modify (1) by substituting

$$\sum_{i=1}^{n} (x_i-\mu)^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2,$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. (This is an identity valid for all $x_1, x_2, \ldots, x_n$ and μ). Substituting in (1) and breaking up the exponential yields

$$f(x \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^2} \right\} \exp\left\{ -\frac{n(\bar{x}-\mu)^2}{2\sigma^2} \right\} \tag{2}$$

$\underline{\Theta_2:}$ Here $\sigma^2 = \sigma_0^2$, (a known value), so the "unknown" parameter is $\theta = \mu$. Factoring the joint pdf (2) as

$$\begin{aligned} f(x \mid \mu) &= \left[(2\pi\sigma_0^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_i (x_i-\bar{x})^2 \right\} \right] \left[\exp\left\{ -\frac{n(\bar{x}-\mu)^2}{2\sigma_0^2} \right\} \right] \\ &= h(x) g(\bar{x}, \mu) = h(x) g(T_2(x), \mu) \end{aligned}$$

where $T_2(x) \equiv \bar{x}$. This shows that $T_2 = T_2(X) = \bar{X}$ is a SS for θ_2.

$\underline{\Theta_1:}$ Here both μ and σ^2 are unknown so $\theta = (\mu, \sigma^2)$. It is clear that (2) may be written as

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= g(\bar{x}, \sum_i (X_i-\bar{x})^2, \mu, \sigma^2) \cdot 1 \\ &= g(T_1(x), \theta) h(x) \end{aligned}$$

where $T_1(x) = (\bar{x}, \sum_i (x_i - \bar{x})^2)$ so that $T_1 = T_1(X) = (\bar{X}, \sum_i (X_i - \bar{X})^2)$ is a SS for Θ_1.
Note: T_1 is also a SS for Θ_2 and Θ_3, neither T_2 or T_3 is a SS for Θ_1.

2.5 General Facts about SS

1. If $T = T(X)$ is a SS for $\theta \in \Theta_A$, and $\Theta_B \subset \Theta_A$, then T is SS for $\theta \in \Theta_B$.

   Proof. If $\mathcal{L}(X \mid T)$ is constant for $\theta \in \Theta_A$, then it is constant for $\theta \in \Theta_B$. □

2. If T is a SS (for $\theta \in \Theta$) and $T = \phi(U)$ where $U = U(X)$, then U is also a SS (for $\theta \in \Theta$).

   Proof. (Using FC) Since T is SS,

$$\begin{aligned} f(x \mid \theta) &= g(T(x) \mid \theta)h(x) \\ &= g(\phi(U(x)) \mid \theta)h(x) \\ &= g^*(U(x) \mid \theta)h(x) \end{aligned}$$

   where $g^*(u \mid \theta) = g(\phi(u) \mid \theta)$. Hence $U(X)$ is SS. □

3. If $T = T(X)$ is a sufficient statistics (for $\theta \in \Theta$), then $U = (S, T)$ is also a sufficient statistic for any $S = S(X)$.

   Proof. Immediate consequence of 2) by taking $\phi(s, t) = t$. With this choice of ϕ, we have $T = \phi(U) \Rightarrow U$ is SS. □

4. If $T = T(X)$ and $U = U(X)$ are related by $T = \phi(U)$ where ϕ is one-one function, then T is SS iff U is SS.

2.6 Application to random samples from various families of normal distributions:

Recall:

1. $T_1 = (\bar{X}, \sum (X_i - \bar{X})^2)$ is SS for $\Theta_1 = \{(\mu, \sigma^2) : \sigma^2 > 0\}$.
2. $T_2 = \bar{X}$ is SS for $\Theta_2 = \{(\mu, \sigma^2) : \sigma^2 = \sigma_0^2\}$.
3. $T_3 = \sum (X_i - \mu_0)^2$ is SS for $\Theta_3 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$.

Some facts:

1. T_1 is SS for $\Theta_1 \Rightarrow T_1$ is SS for Θ_2 and for Θ_3 (Follows from Fact 1 since $\Theta_1 \supset \Theta_2$ and $\Theta_1 \supset \Theta_3$.
2. T_2 is SS for $\Theta_2 \Rightarrow T_1$ is SS for Θ_2 (Follows from Fact 3).
3. T_3 is SS for Θ_3 and $T_3 = \sum(X_i - \mu_0)^2 = \sum(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 = \phi(T_1) \Rightarrow T_1$ is SS for Θ_3 (Follows from Fact 2).
4. T_1 is SS for $\Theta_1 \Rightarrow (\bar{X}, \frac{1}{n-1}\sum(X_i - \bar{X})^2)$ is SS for Θ_1 and $(\sum X_i, \sum X_i^2)$ is SS for Θ_1 (Since both of these are one-one functions of T_1 (Follows from Fact 4).

3 Minimal sufficient statistic

Definition 2. *A minimal sufficient statistic is a function of any other sufficient statistic. $T = T(X)$ is minimal sufficient if for every sufficient statistic $S = S(X)$ there exists a function ψ such that $T = \psi(S)$, that is, $T(X) = \psi(S(X))$.*

Theorem 2. *(Lehmann-Scheffe Theorem) $X \sim P_\theta, \theta \in \Theta$. $T(X)$ is a minimal sufficient statistic iff for all x, y, $T(x) = T(y)$ iff $\frac{f(x|\theta)}{f(y|\theta)}$ is constant as a function of θ.*

Remark 1. *It is difficult to show a statistic is MSS directly from the definition. For proving MSS, we usually use the Lehmann-Scheffe Theorem. However, it is often very easy to prove a statistic is not MSS using the definition. If S and T are two different sufficient statistics, and T cannot be written as a function of S, then T is not minimal.*
Example: Consider the three families of normal distributions used earlier. T_1 and T_2 are both SS for Θ_2, but T_1 clearly cannot be written as a function of T_2. Thus T_1 is not a MSS for Θ_2.
Similarly, T_1 and T_3 are both SS for Θ_3, but T_1 clearly cannot be written as a function of T_3. Thus T_1 is not a MSS for Θ_3.

Comments on the Lehmann-Scheffe Theorem

1. In situations where the support of $f(x \mid \theta)$ depends on θ, a better statement (which avoids awkward $\frac{0}{0}$'s) is: For all x, y, $T(x) = T(y)$ iff $f(x \mid \theta) = c(x, y) f(y \mid \theta)$ for all θ.
2. The "iff" can be broken down as two results
   (a) If $T(X)$ is sufficient, then for all x, y, $T(x) = T(y)$ implies $\frac{f(x|\theta)}{f(y|\theta)}$ constant in θ.
   (b) A sufficient statistic $T(X)$ is minimal if for all x, y, $\frac{f(x|\theta)}{f(y|\theta)}$ constant in θ implies $T(x) = T(y)$.

3.1 Examples for Lehmann-Scheffe Theorem

1. $X = (X_1, \ldots, X_n)$ iid $\mathrm{N}(\mu, \sigma^2)$. $T(X) = (\bar{X}, S^2)$ where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is MSS for (μ, σ^2)

2. $X = (X_1, \ldots, X_n)$ iid Uniform(α, β), $\Theta = \{(\alpha, \beta) : -\infty < \alpha < \beta < \infty\}$. $T(X) = (X_{(1)}, X_{(n)})$ is MSS for (α, β) ($X_{(1)} = \min X_i, X_{(n)} = \max X_i$). We must verify: for all x, y, $T(x) = T(y)$ iff there exists $c \neq 0$ such that $f(x \mid \theta) = cf(y \mid \theta)$ for all θ. (c does not involve θ, but can depend on x, y). In this case,

$$\begin{aligned} f(x \mid \theta) &= \prod_{i=1}^n \frac{1}{\beta - \alpha} I(\alpha \leq x_i \leq \beta) \\ &= \frac{1}{(\beta - \alpha)^n} I(x_{(1)} \geq \alpha) I(x_{(n)} \leq \beta) \end{aligned}$$

   Similarly,

$$f(y \mid \theta) = \frac{1}{(\beta - \alpha)^n} I(y_{(1)} \geq \alpha) I(y_{(n)} \leq \beta).$$

   Clearly,

$$(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$$

   implies $f(x \mid \theta) = f(y \mid \theta)$ (can take $c = 1$) for all $\theta \in \Theta$. This gives one direction. What about the other? Define

$$A(x) = \{\theta : f(x \mid \theta) > 0\}.$$

   Here $\theta = (\alpha, \beta)$ with $\alpha < \beta$. Assume that there exists $c \neq 0$ such that $f(x \mid \theta) = cf(y \mid \theta)$ for all θ. Then we must have $A(x) = A(y)$. But

$$A(x) = \{(\alpha, \beta) : \alpha \leq x_{(1)}, \beta \geq x_{(n)}\}.$$

   for any x. Thus $A(x) = A(y)$ implies $(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$ proving that $(x_{(1)}, x_{(n)})$ is MSS.
   Note: This style of argument can only work for examples similar to the uniform distribution where the support depends upon the parameter value.

3. $X = (X_1, \ldots, X_n)$ iid Uniform$(\theta, \theta + 1)$. Then $T(X) = (X_{(1)}, X_{(n)})$ is MSS for θ.
   Comments:

   (a) The dimension of the MSS does not have to be the same as the dimension of the parameter.

(b) "shrinking" the parameter space does <u>not</u> always change the MSS. When $X = (X_1, \ldots, X_n)$ iid Uniform(α, β), $\Theta_1 = \{(\alpha, \beta) : \alpha < \beta\}$ and $\Theta_2 = \{(\alpha, \beta) : \beta = \alpha + 1\}$ have the same MSS.

4. <u>Random Sample Model:</u> Suppose $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ iid $\psi(x \mid \theta)$ (pdf or pmf) where $\psi(x \mid \theta)$ is an arbitrary family of pdf's (pmf's). Then

$$T(\underset{\sim}{X}) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)}),$$

the order statistics (data arranged in increasing order) is a sufficient statistic for θ, but may not be minimal.

Proof. (Use FC)

$$\begin{aligned} f(\underset{\sim}{x} \mid \theta) = \prod_{i=1}^{n} \psi(x_i \mid \theta) &= \prod_{i=1}^{n} \psi(x_{(i)} \mid \theta) \cdot 1 \\ &= g(T(\underset{\sim}{x}) \mid \theta) h(\underset{\sim}{x}). \end{aligned}$$

□

<u>Note:</u> (assume $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$). Then

$$P(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t) = \frac{1}{n!}$$

if $\underset{\sim}{x}$ is any rearrangement of $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and 0 otherwise. All possible ordering are equally likely. To generate from $\mathcal{L}(\underset{\sim}{X} \mid T)$, place the values $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ in a hat and draw them out one by one.
<u>Comment:</u> For random sample models, the order statistics are often the SS.

5. $\underset{\sim}{X} = (X_1, \ldots, X_n)$ iid $\psi(x \mid \theta)$ with

$$\psi(x \mid \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2},$$

the Cauchy-location family. Look at

$$\frac{f(\underset{\sim}{x} \mid \theta)}{f(\underset{\sim}{y} \mid \theta)} = \frac{\prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1+(x_i-\theta)^2}}{\prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1+(y_i-\theta)^2}}$$

If $x_{(i)} = y_{(i)}$ for all i, then the ratio is a constant function of θ. Now suppose $f(\underset{\sim}{x} \mid \theta)/f(\underset{\sim}{y} \mid \theta)$ is a constant function of θ. Then

$$\prod_{i=1}^{n}(1 + (x_i - \theta)^2) = c(x, y) \prod_{i=1}^{n}(1 + (y_i - \theta)^2)$$

for some function $c(x, y)$ independent of θ. This is equivalent to

$$\prod_{i=1}^{n}(\theta^2 - 2x_i\theta + x_i^2 + 1) = c(x, y) \prod_{i=1}^{n}(\theta^2 - 2y_i\theta + y_i^2 + 1).$$

Clearly, both $\prod_{i=1}^{n}(\theta^2 - 2x_i\theta + x_i^2 + 1)$ and $\prod_{i=1}^{n}(\theta^2 - 2y_i\theta + y_i^2 + 1)$ are polynomials of degree $2n$ in θ with the same set of zeros $\mathcal{O}_L$ and $\mathcal{O}_R$. We can spell out

$$\mathcal{O}_L = \{x_i \pm i, i = 1, \ldots, n\}, \quad \mathcal{O}_R = \{y_i \pm i, i = 1, \ldots, n\},$$

where $i = \sqrt{-1}$, the imaginary root of $-1/$ Then $\mathcal{O}_L$ and $\mathcal{O}_R$ are permutations of each other. Hence $x_{(i)} = y_{(i)}$ for all $i = 1, \ldots, n$.

6. Suppose $X \sim P_\theta, \theta \in \Theta$ and P_θ has a joint pdf or pmf $f(x \mid \theta)$.
   Fact: X is a SS for θ.

   Proof. (Using FC) Define $T = T(X) = X$. (T is the identity function.) Then

   $$f(x \mid \theta) = f(x \mid \theta) \cdot 1 = g(T(x) \mid \theta) \cdot h(x)$$

   where $g = f$ and $h(x) \equiv 1$. Thus T is SS. □

   Proof. (From definition of SS)

   $$\mathcal{L}(X \mid T(X) = t) = \mathcal{L}(X \mid X = t) = \delta_t$$

   where δ_t is the probability measure which places all its mass at the point (dataset) t. □

7. Further suppose $X = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are iid from the pdf (pmf) $f(x \mid \theta)$.
   Fact: $T(X) = X = (X_1, \ldots, X_n)$ is not a MSS.

   Proof. (from definition of MSS) Let $S = S(X) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ (the order statistics). Since we have a random sample model, S is a SS. But clearly T is not a function of S. (You cannot recover the original ordering of the data given only the order statistics.) Thus T is not a MSS. □