STAT 653 - Notes
Introduction to Mathematical Statistics

September 26, 2025

Contents

1 Statistical Model

Example. A coin is tossed n times. The data available is $X = (X_1, X_2, \dots, X_n)$, where $X_i \in \{0, 1\}$. The assumptions are:

1. outcomes are independent.
2. $P(X_i = 1) = \theta \in \Theta$ where θ is an unknown parameter and Θ is the parameter space. In this case $\Theta = [0, 1]$.

We need to estimate θ based on the data $X = (X_1, X_2, \ldots, X_n)$, where X_i are random variables before the experiment is conducted.

So we need to find an estimator $T(X_1, X_2, \ldots, X_n)$ of $\theta \in \Theta$.

Possible Estimators

1. $T_1 := T_1(X_1, X_2, \ldots, X_n) = \overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

   Remark. (a) $\mathbb{E}(T_1) = \mathbb{E}(\overline{X}_n) = \mathbb{E}(X_1) = \theta$ for all $\theta \in \Theta$ then T_1 is unbiased estimator of θ.

   (b) $\lim_{n\to\infty} P(|\overline{X}_n - \theta| > \epsilon) = 0$ for all $\epsilon > 0$.

   Definition. In general, if $\lim\limits_{n\to\infty} P(|T(X_1, \ldots, X_n) - \theta|\epsilon) = 0$ for all $\epsilon > 0$ and for all $\theta \in \Theta$, then we call $T(X_1, \ldots, X_n)$ **consistent**.

2. $T_2(X_1, \ldots, X_n) := X_1$, where $X_1 \in \{0, 1\}$. Then $\mathbb{E}(T_2) = \mathbb{E}(X_1) = \theta$ for all $\theta \in \Theta$.

   T_2 is unbiased but is not <u>consistent</u>.

3.
$$\begin{aligned} T_3 &:= T_3(X_1, \ldots, X_n) \\ &= \sqrt{\frac{1}{\lfloor \frac{n}{2} \rfloor}\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} X_{2i}X_{2i-1}} \end{aligned}$$

   T_3 is biased because

$$\begin{aligned} \mathbb{E}(T_3) &\leq \sqrt{\frac{1}{\lfloor \frac{n}{2} \rfloor}\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} X_{2i}X_{2i-1}} \\ &= \theta \qquad \forall \theta \in \Theta \end{aligned}$$

Example. Suppose $X_1, X_2, \ldots, X_n$ are independent and have uniform$[0, \theta]$, where $theta \in \Theta = \mathbb{R}_+$. So $\Theta = \{\theta : \theta > 0\}$.

Possible Estimators

1. $T_1(X_1, \ldots, X_n) = 2\overline{X}_n$
2. $T_2(X_1, \ldots, X_n) = X_{(n)}$ (max)
3. $T_3(X_1, \ldots, X_n) = c_n X_{(n)}$

   Correct the max by a constant so it is unbiased.

Example. We want to receive a shipment of oranges and suspect that part of them rot off. To check the shipment we draw a random sample without replacement of size n from the shipment (population) of size N.

Let θ be the proportion of bad oranges in the population. So $\Theta = \{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N}{N}\}$.

Let

$$X_i = \begin{cases} 0 & \text{if good} \\ 1 & \text{if bad} \end{cases}$$

for $i = 1, 2, \ldots, n$ and let $X = (X_1, X_2, \ldots, X_n)$.

Let $T_1(X) = \sum_{i=1}^{n} X_i$. Then T_1 has a hypergeometric distribution. So

$$P_\theta(X_1 = k) = \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}}$$

for $k \in \{\max(0, n - (N - N\theta), \ldots, \min(n, N\theta))\}$

2 The Likelihood Function

$$X \sim P_\theta, \qquad \theta \in \Theta$$

We have 2 cases for now (discrete and continuous):

(R1) P_θ is defined by a joint pdf $f_X(x;\theta)$ for all $\theta \in \Theta$.

(R2) P_θ is defined by a joint pmf $P(X = x;\theta)$ for all $\theta \in \Theta$.

Definition. Let P_θ, $\theta \in \Theta$ be a model satisfying (R1) or (R2). Then the function

$$L(x;\theta) = \begin{cases} f_X(x;\theta) & \text{if (R1)} \\ P(X = x;\theta) & \text{if (R2)} \end{cases}.$$

Example. Not (R1) and not (R2).
Let

$$X \sim N(\theta, 1) \qquad \theta \in \Theta = \mathbb{R}.$$

We observe $Y = \max(0, X)$,

$$Y = \begin{cases} 0 & \text{if } X \leq 0 \\ X & \text{if } X > 0 \end{cases} = XI(X > 0)$$

where $I(\cdot)$ is the indicator function.

$F_\theta(t) = P(Y \leq t)$ for all $t \in \mathbb{R}$.

Example. Back to oranges example where $X = (X_1, X_2, \ldots, X_n)$ is the data and $\Theta = \{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N}{N}\}$. Let $T(X) = \sum_{i=1}^{n} X_i$. Then

$$\begin{aligned} L(x;\theta) &= P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= P_\theta\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T(X) = \sum_{i=1}^{n} x_i\right) \\ &= P_\theta\left(T(X) = \sum_{i=1}^{n} x_i\right) P\left(X_1 = x_1, \ldots, X_n = x_n \,\middle|\, T(X) = \sum_{i=1}^{n} x_i\right). \end{aligned}$$

Now define $K_n = \sum_{i=1}^{n} x_i$. For example, if $n = 5$ and we observed $(1, 0, 0, 1, 1)$ then

$$K = \sum_{i=1}^{5} x_i = 3.$$

Since there are 10 possibilities for which entries are 1 versus 0, $\binom{5}{3} = 10$. Because all possible combinations of 1 and 0 are possible we can use symmetry to calculate the probability of any particular sequence of 1 and 0 as $1/\binom{5}{3}$. We use this reasoning below to derive the expression on the right.

Then

$$L(x;\theta) = \frac{\binom{N\theta}{K_n}\binom{N-N\theta}{n-K_n}}{\binom{N}{n}} \times \frac{1}{\binom{n}{K_n}}.$$

3 Identifiability of Statistical Models

Definition. Let $X \sim P_\theta$, $\theta \in \Theta$. A model P_θ, $\theta \in \Theta$ is <u>identifiable</u> if for any pair (θ, θ') such that $\theta \neq \theta'$ and $\theta, \theta' \in \Theta$, then $P_\theta \neq P_{\theta'}$.

Remark. This means that there is an event A, such that $P_\theta(A) \neq P_{\theta'}$ where $\theta \neq \theta'$.

R(1) For $\theta \neq \theta'$, $f(x;\theta) \neq f(x;\theta')$ for any neighborhood of x (an open ball $B(x, r)$ centered at x).

By open ball we mean $B(x, r) = \{y : |x - y| < \epsilon\}$ where $|v| = (\sum_{i=1}^{n} v_i^2)^{1/2}$ (euclidean norm).

R(2) Discrete support, for some x $P_\theta(X = x) \neq P_{\theta'}(X = x)$ where $\theta \neq \theta'$.

Example. Suppose we observe $X_1, X_2, \ldots, X_n$ where $X_i = \theta \cdot Z_i \sim N(0, \theta^2)$ and $Z_i \sim N(0,1)$ and $\theta \in \Theta = \mathbb{R} \setminus \{0\}$.

If $\theta_1 = 1 \neq -1 = \theta_2$, then

$$L(x_1, x_2, \ldots, x_n; \theta = 1) = L(x_1, x_2, \ldots, x_n; \theta = -1)$$

for any $x = (x_1, \ldots, x_n)$.

Result. The model $\{P_\theta, \theta \in \Theta\}$ is identifiable if there exists a statistic $T(X)$ ($X \sim P_\theta, \theta \in \Theta$) where expectation is a one-to-one function of $\theta \in \Theta$, i.e., such that

$$\forall(\theta, \theta'), \quad \theta \neq \theta' \implies \mathbb{E}_\theta(T(X)) \neq \mathbb{E}_{\theta'}(T(X)) \tag{1}$$

Proof. We use proof by contradiction. Suppose that (1) holds, but there exists $\theta \neq \theta'$ such that $P_\theta = P_{\theta'}$. If so, then $\mathbb{E}_\theta(T(X)) = \mathbb{E}_{\theta'}(T(X))$, which contradicts (1). □

In the previous example, $\theta = 1$, $\theta' = -1$.

Example. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ where $\theta \in \Theta = [0,1]$. We will show that θ is identifiable using the definition and also the above result.

Let θ and θ' be arbitrary and suppose $\theta \neq \theta'$ and θ',$\theta \in \Theta$. Also suppose $X = (1, 1, \ldots, 1)$. Then

$$\begin{aligned} P_\theta(X_1, X_2, \ldots, X_n) &= \theta^n \\ P_{\theta'}(X_1 = 1, \ldots, X_n) &= (\theta')^n. \end{aligned}$$

Since $\theta \in [0,1]$ then $\theta^n \neq (\theta')^n$ and the model is identifiable.

Now take a statistic $T(X_1, \ldots, X_n) = X_1$ (or we could take $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ or $T(X_1, \ldots, X_n) = \sum_{i=1}^n \overline{X}_n$).

For any $(\theta, \theta') \in \Theta$, if $\theta \neq \theta'$ then $\mathbb{E}_\theta(\overline{X}_n) = \theta \neq \theta' = \mathbb{E}_{\theta'}(\overline{X}_n)$. Then by the above result the model is identifiable.

Example.

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

Part 1) Let $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^2$. Then

$$L(x_1, \ldots, x_n; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}} I(\mu \in \mathbb{R}) I(\sigma^2 > 0).$$

It is difficult in this case to use the definition to show identifiability in this case, but we can use the previous result.

We are given $X = (X_1, X_2, \ldots, X_N)$. Let

$$T(X) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right)$$

Then

$$\begin{aligned} \mathbb{E}_\theta(T) &= (n\mu, n(\sigma^2 + \mu^2)), \\ \mathbb{E}_{\theta'}(T) &= (n\mu', n(\sigma^{2'} + (\mu')^2) \end{aligned}$$

Thus, if $(\theta, \theta^2) \in \Theta$ then

$$\forall \theta \neq \theta' \implies \mathbb{E}_\theta(T(X)) \neq \mathbb{E}_{\theta'}(T(X)).$$

If $\theta \neq \theta'$ then $\mu \neq \mu'$ or $\sigma^2 \neq \sigma^{2'}$ or $\mu \neq \mu'$ and $\sigma^2 \neq \sigma^{2'}$. In all three cases then $\mathbb{E}_{\prime}(T(X)) \neq \mathbb{E}_{\theta'}(T(X))$.

Part 2) Suppose we observe only $Y_1, \ldots, Y_n$ where

$$Y_i = \begin{cases} +1 & \text{if } X_i \geq 0 \\ -1 & \text{if } X_i < 0. \end{cases}$$

Since $Y_i = g(X_i)$ and the X_i's are independent, then the Y_i's are also independent.

Then the likelihood function is

$$\begin{aligned} L(y_i, \ldots, y_n; \theta) &= \prod_{i=1}^{n} P(Y_i = y_i; \theta) \\ &= \prod_{i=1}^{n} \left[I(y_i = 1) P(X_i \geq 0) + I(y_i = -1) P(X_i < 0) \right]. \end{aligned}$$

Now note that

$$\begin{aligned} &P(X_i \geq 0) = 1 - P(X_i < 0) = 1 - \Phi\left(-\frac{\mu}{\sigma}\right) = \Phi\left(\frac{\mu}{\sigma}\right) \\ &P(X_i < 0) = \Phi\left(-\frac{\mu}{\sigma}\right) \end{aligned}$$

so that only the ratio μ/σ matters for the the likelihood.

Now let $\theta = (3,9) \neq (4,16) = \theta'$. For θ we have $\mu/\sigma = 3/3 = 1$ and for θ' we have $\mu/\sigma = 4/4 = 1$. Thus we have

$$\theta = (3,9) \neq (4,16) = \theta' \implies L(y;\theta) = L(y;\theta')$$

and so the model is not identifiable. For any $y = (y_1, \ldots, y_n)$ we have $L(y;\theta) = L(y;\theta')$ and thus the model is not identifiable.

Remark. Above we used the fact that for a general normal random variable $N(\mu, \sigma^2)$, $F(x) = \Phi((x-\mu)/\sigma)$.

4 Sufficient Statistic

Definition. Let $X \sim P_\theta$, $\theta \in \Theta$ and we observe data $X = (X_1, \ldots, X_n)$. A statistic $T(X)$ is **sufficient** for the model $\{P_\theta, \theta \in \Theta\}$ if the conditional distribution of $X \mid T(X)$ does not depend on θ.

Remark. Consider the following 2 stage procedure. Assume $T(X)$ is a sufficient statistic for the model $\{P_\theta, \theta \in \Theta\}$.

(1) Suppose we observed data from $X \sim P_\theta$, $\theta \in \Theta$. Now calculate $T(X)$, keep it and discard X.

(2) Generate X' from conditional distribution $X \mid T(X)$.

For any $\theta \in \Theta$ calculate marginal distribution of new X'. Then

$$\begin{aligned} P_\theta(X' = x) &= \sum_t P_\theta(X' = x \mid T(X) = t) P_\theta(T(X) = t) \\ &= \sum_t P_\theta(X = x \mid T(X) = t) P_\theta(T(X) = t) \\ &= P_\theta(X = x) \end{aligned}$$

for any X.

Example. Let $X = (X_1, X_2, \ldots, X_n) \overset{iid}{\sim} \text{Bernoulli}(\theta)$ where $\theta \in \Theta = (0,1)$. Let $T(X) = \sum_{i=1}^n X_i \overset{iid}{\sim} \text{Binomial}(n, \theta)$.

Then

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid T(X) = t) = \begin{cases} 0 & \text{if } t \neq \sum_{i=1}^n x_i \\ * & \text{if } t = \sum_{i=1}^n x_i \end{cases}$$

where

$$* = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

which does not depend on θ.

Thus the $X \mid T(X)$ has a discrete uniform distribution,

$$(X_1, \ldots, X_n) \mid T(X) = t \sim \text{uniform}\left\{x_1, \ldots, x_n : x_i \in \{0,1\} \text{ and } \sum_{i=1}^{n} x_i = t\right\}$$

Remark. In the above example $\sum_{i=1}^{n-1} X_i$ is not a sufficient statistic. To see this note that

$$\mathbb{E}\left(X \,\middle|\, \sum_{i=1}^{n-1} X_i = t\right) = \theta$$

which implies that the conditional distribution depends on θ.

Example.

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1) \qquad \theta \in \Theta = \mathbb{R}$$

Let $T(X) = \sum_{i=1}^{n} X_i = \overline{X}_n$. Then

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \,\Bigg|\, \overline{X}_n = t \;\sim N\left(\begin{bmatrix} t \\ t \\ \vdots \\ t \end{bmatrix}, \begin{bmatrix} 1-\frac{1}{n}, & -\frac{1}{n}, & \ldots, & -\frac{1}{n} \\ -\frac{1}{n}, & 1-\frac{1}{n}, & \ldots, & -\frac{1}{n} \\ \vdots & & & \\ -\frac{1}{n}, & \ldots, & -\frac{1}{n}, & 1-\frac{1}{n} \end{bmatrix}\right)$$

where the multivariate normal distribution on the right does not depend on θ. Thus $\overline{X}_n$ is sufficient for this model.

5 Fisher-Neyman Factorization Theorem

Consider the model $X \sim P_\theta$, $\theta \in \Theta$. Then $T(X)$ is sufficient statistic for P_θ if and only if there exists functions $g(\theta, t)$ and $h(x)$ (with appropriate domains) such that

$$L(x;\theta) = g(\theta, T(x))h(x) \qquad \forall x; \forall \theta \in \Theta$$

Proof. (I) Sufficient condition

Assume holds and we must show that $T(X)$ is sufficient. We do only the discrete case.

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

where $*$ is

$$\begin{aligned} * = \frac{P_\theta(X = x)}{P_\theta(T(x) = t)} &= \frac{P_\theta(X = x)}{\sum\limits_{y:T(y)=t} P_\theta(X = y)} \\ &= \frac{g(\theta, T(x) = t)h(x)}{\sum\limits_{y:T(y)=t} g(\theta, T(x) = t)h(y)} \\ &= \frac{h(x)}{\sum\limits_{y:T(y)=t} h(y)}. \end{aligned}$$

Since the final expression above does not depend on θ, which implies that $T(X)$ is a sufficient statistic for the given model.

(II) Necessary Condition

Now assume that $T(X)$ is a sufficient statistic for the given model. Then

$$\begin{aligned} P_\theta(X = x) &= P_\theta(X = x, T(x) = T(x)) \\ &= P(X = x \mid T(x) = t_x) P_\theta(T(x) = t_x) \\ &= h(x) g(\theta, t_x). \end{aligned}$$

$\square$

Example. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ where $X_i \in \{0, 1\}$. Then the likelihood is

$$\begin{aligned} L(x; \theta) &= P_\theta(X_1 = x_1, \ldots, X_n = x_n) \\ &= \theta^{\sum\limits_{i=1}^n x_i} (1 - \theta)^{n - \sum\limits_{i=1}^n x_i} \prod_{i=1}^n I(x_i \in \{0, 1\}) \\ &= g\left(\theta, T(x) = \sum_{i=1}^n x_i\right) h(x). \end{aligned}$$

Thus, $T(X) = \sum_{i=1}^n X_i$ is sufficient for this model.

Example. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Then the likelihood is

$$
\begin{aligned}
L(x;\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} \prod_{i=1}^{n} I(-\infty < x_i < \infty) \\
&= e^{\theta(\sum_{i=1}^{n} x_i) - \frac{n\theta^2}{2}} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \prod_{i=1}^{n} I(-\infty < x_i < \infty) \\
&= \left[e^{\theta n \overline{X}_n - \frac{n\theta^2}{2}}\right] \left[\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \prod_{i=1}^{n} I(-\infty < x_i < \infty)\right] \\
&= g(\theta, \overline{X}_n) h(x)
\end{aligned}
$$

Thus, $\overline{X}_n$ is a sufficient statistic for this model.

Example.

$$
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad -1 < \rho < 1, \quad -\infty < x_1, x_2, < \infty
$$

Suppose we have on observation $x = (x_1, x_2)$. Then the likelihood is

$$
\begin{aligned}
L(x;\rho) &= \frac{1}{2\pi\sqrt{1-\rho}} e^{-\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)}} I(-\infty < x_1, x_2 < \infty) \\
&= g(\rho; T(x) = (x_1^2 + x_2^2, x_1 x_2)) h(x)
\end{aligned}
$$

where $h(x) = 1$. Now suppose we have n observations

$$
x = \left(\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix}, \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix}, \ldots, \begin{bmatrix} x_{1n} \\ x_{2n} \end{bmatrix}\right)
$$

where each vector in x is independent of all others. Then the sufficient statistic $T(X)$ is

$$
T(x) = \left(\sum_{j=1}^{n} (x_{1j}^2 + x_{2j}^2), \sum_{j=1}^{n} x_{1j} x_{2j}\right).
$$

6 Exchangable Random Variables

Definition. The random variables $X_1, X_2, \ldots, X_n$ are **exchangable** random variables if

$$
(X_1, \ldots, X_n) \sim (X_{\pi(1)}, \ldots, X_{\pi(n)})
$$

for any permutation $\pi(1), \ldots, \pi(n)$ of integers $1, 2, \ldots, n$.

Remark. If $X_1, \ldots, X_n$ are identically and independently distributed $\implies X_1, \ldots, X_n$ are exchangable. However, $X_1, \ldots, X_n$ are exchangable $\not\Longrightarrow X_1, \ldots, X_n$ are identically and independently distributed.

Example.

$$P(X_1 = x_1, X_2 = x_2) = P(X_2 = x_2, X_1 = x_1)$$

Example.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad -1 < \rho < 1, \quad -\infty < x_1, x_2, < \infty$$

Suppose we have on observation $x = (x_1, x_2)$.
Then $f_{x_1,x_2}(x_1, x_2; \rho) = f_{x_1,x_2}(x_2, x_1; \rho)$ so that $(X_1, X_2) \sim (X_2, X_1)$.

Result. If $(X_1, \ldots, X_n) \sim P_\theta$, $\theta \in \Theta$ are exchangable random variables then

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

is a sufficient statistic for P_θ, $\theta \in \Theta$ where $T(X)$ is the vector of order statistics.

Proof. Let $(X_1, \ldots, X_n) \sim P_\theta$, $\theta \in \Theta$ are exchangable random variables. Let $y_1 \leq y_2 \leq \ldots \leq y_n$ be the observed order statistics. Then

$$\begin{aligned} &P_\theta(X_1 = x_1, X_2, = x_2, \ldots, X_n = x_n \mid X_{(1)} = y_1, \ldots, X_{(n)} = y_n) \\ &\qquad = \begin{cases} * & \text{if } \{x_1, \ldots x_n\} = \{y_1, \ldots, y_n\} \\ 0 & \text{if } \{x_1, \ldots, x_n\} \neq \{y_1, \ldots, y_n\} \end{cases} \end{aligned}$$

where

$$\begin{aligned} * &= \frac{P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P_\theta(X_{(1)} = y_1, \ldots, X_{(n)} = y_n)} \\ &= \frac{P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{\sum\limits_{\substack{\text{all} \\ \text{possible} \\ \text{permutations}}} P_\theta(X_1 = y_{\pi(1)}, \ldots, X_n = y_{\pi(n)})} \\ &= \frac{P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{\sum\limits_{\substack{\text{all} \\ \text{possible} \\ \text{permutations}}} P_\theta(X_1 = x_{\pi(1)}, \ldots, X_n = x_{\pi(n)})} \\ &= \frac{P_\theta(X_1 = x_1, \ldots, X_n = x_n)}{n! P_\theta(X_1 = x_1, \ldots, X_n = x_n)} \\ &= \frac{1}{n!}. \end{aligned}$$

Note that $1/n!$ does not depend on θ for all $\theta \in \Theta$. Therefore, by definition, the vector of order statistics is sufficient for model P_θ, $\theta \in \Theta$. □

Example.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad -1 < \rho < 1, \quad -\infty < x_1, x_2, < \infty$$

Suppose we have on observation $x = (x_1, x_2)$.
From the previous result we have $T(x) = (x_{(1)}, x_{(2)})$ is sufficient.

We know already that $T_1(x) = (x_1^2 + x_2^2, x_1 x_2)$ is sufficient and we just showed that $T_2(x) = (x_{(1)}, x_{(2)})$ is also sufficient.

Then T_1 is a function of T_2. So if we know T_2 then we can calculate T_1, but not vice versa. This leads to the next definition.

7 Minimal Sufficient Statistic

Definition. Let $X \sim P_\theta$, $\theta \in \Theta$. The sufficient statistic, $S(X)$, is minimal sufficient if there is a function of any other sufficient statistic. This means that for any sufficient statistic $T(X)$ there exists a function $f : S(X) = f(T(X))$.

We can use the following lemma to check for minimal sufficiency.

Lemma. *Let $X \sim P_\theta$, $\theta \in \Theta$. A sufficient statistic $S(X)$ is minimal sufficient if*

$$\frac{L(x;\theta)}{L(y;\theta)} \text{ does not depend on } \theta \implies S(x) = S(y), \quad \forall \theta \in \Theta.$$

Proof. Let x, y be such that $T(x) = T(y)$. Suppose the implication in the lemma holds and let $T(X)$ be a sufficient statistic. Then

$$\frac{L(x;\theta)}{L(y;\theta)} = \frac{g(\theta, T(x))h(x)}{g(\theta, T(y))h(y)} = \frac{h(x)}{h(y)}$$

which does not depend on *theta*, for all $\theta \in \Theta$. Then this implies that $S(x) = S(y)$.

Recall that $T(x)$ is arbitrary. Above we showed that $T(x) = T(y) \implies S(x) = S(y)$. Now choose another sufficient statistic, $T_1(X)$. Then $T_1(x) = T_1(y) \implies S(x) = S(y)$. Then if follows that

$$S(x) = f(T(x)).$$

We won't write the formal proof because it will take a while. □

Remark. (1) If $S(x) = S(y)$ then

$$\frac{g(\theta; S(x))h(x)}{g(\theta, S(y))h(y)} = \frac{h(x)}{h(y)}$$

which does not depend on θ for all $\theta \in \Theta$.

(2) $A \implies B$ is equivalent to $B^c \implies A^c$.

(3) The meaning of the left hand side of the implication in the above lemma is

$$L(x;\theta) = c(x,y)L(y;\theta), \quad \forall \theta \in \Theta.$$

Example. Let $X = (X_1, \ldots, X_n)$ where $X_i \overset{iid}{\sim} N(\theta, 1)$ where $\theta \in \Theta = \mathbb{R}$.

8 Ancillary Statistic

Definition. Let $X \sim P_\theta$, $\theta \in \Theta$.

(1) A statistic, $A(X)$, whose distribution does not depend on θ is called **ancillary**.

(2) If $\mathbb{E}_\theta(A^*(X))$ does not depend on θ then $A^*(X)$ is **first-order ancillary**.

Remark. Ancillary $\implies$ first-order ancillary, but first-order ancillary $\not\Longrightarrow$ ancillary.

Example. Let $X_1, X_2 X_n \overset{iid}{\sim} N(\theta, 1), \quad \theta \in \Theta = \mathbb{R}$.

A sufficient statistic for this model is

$$S(X) = X_1 + X_2 \sim N(2\theta, 2).$$

An ancillary statistic for this model is

$$A(X) = X_1 - X_2 \sim N(0, 2).$$

Example.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad -1 < \rho < 1$$

In this example $\theta = \rho$. A sufficient statistic for this model is

$$S(X) = (X_1^2 + X_2^2, X_1 X_2).$$

Note $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$ are not independent.

Then two ancillary statistics are

$$\begin{aligned} A_1(X) &= X_1 \\ A_2(X) &= X_2. \end{aligned}$$

A first-order ancillary statistic is

$$A^*(X) = X_1^2 + X_2^2$$

because $\mathbb{E}_\theta(A^*(X)) = 2$.

Another ancillary statistic is

$$A(X) = I(-1 \leq X_1 < 1) + I(-1 \leq X_2 \leq 1).$$

9 Scale and Location Family

9.1 Location Family

Let $X \sim F$ where F is some distribution function that does not depend on any unknown parameters, e.g., $N(0,1)$. Let $\theta \in \Theta = \mathbb{R}$ and define

$$Y = X + \theta.$$

Then Y is in a location family. The distribution function for Y is

$$F_Y(t) = P(Y \leq t) = P(X \leq t - \theta) = F_X(t - \theta).$$

Remark. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ where F is a distribution function that does not depend on any unknown parameters. Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables from a location family.

Then

$$\begin{aligned} Y_1 - Y_2 &= (X_1 + \theta) - (X + \theta) \\ &= X_1 - X_2. \end{aligned}$$

Thus, $Y_1 - Y_2$ is an ancillary statistic. Also,

$$\begin{aligned} Y_{(j)} - Y_{(i)} &= (X_{(j)} + \theta) - (X_{(i)} + \theta) \\ &= X_{(j)} - X_{(i)}. \end{aligned}$$

Thus, $Y_{(j)} - Y_{(i)}$ is ancillary.

9.2 Scale Family

Let $X \sim F$ where F is some distribution function that does not depend on any unknown parameters, e.g., $N(0,1)$. Let $\sigma \in \mathbb{R}_+ = \Theta$. Then

$$Y = \sigma \cdot X$$

is in a scale family. The distribution function for Y is

$$F_Y(t) = P(Y \leq t) = P(\sigma X \leq t) = P\left(X \leq \frac{t}{\sigma}\right) = F_X\left(\frac{t}{\sigma}\right) \quad \forall t \in \mathbb{R}.$$

Remark. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ where F is a distribution function that does not depend on any unknown parameters. Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables from a scale family.

Then

$$\frac{Y_1}{Y_2} = \frac{\sigma X_1}{\sigma X_2} = \frac{X_1}{X_2}.$$

Thus $\dfrac{Y_1}{Y_2}$ is an ancillary statistic. Also

$$\frac{Y_{(1)}}{Y_{(2)}} = \frac{\sigma X_{(1)}}{\sigma X_{(2)}}.$$

Thus $\dfrac{Y_{(1)}}{Y_{(2)}}$ is an ancillary statistic.

9.3 Location-scale Family

Let $X \sim F$ where F is some distribution function that does not depend on any unknown parameters, e.g., $N(0, 1)$. Let $\sigma \in \mathbb{R}_+ = \Theta$ and $\theta \in \mathbb{R}$. Then

$$Y = \sigma \cdot X + \theta$$

is in a location-scale family.

Remark. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ where F is a distribution function that does not depend on any unknown parameters. Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables from a location-scale family.

Then

$$\frac{Y_1 - Y_2}{Y_3 - Y_4}$$

does not depend on σ or θ. We could also write

$$\frac{Y_1 - Y_2}{Y_2 - Y_4}.$$

10 Point Estimation

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

We want to estimate θ or some function $g(\theta)$ based on the sample X.

Example. Taxi Example

$$1, 2, \ldots, N$$

where N is an unknown parameter. Take a random sample without replacement of size $X_1, \ldots, X_n \in \{1, 2, \ldots, N\}$. We know that $X_{(n)}$ is a sufficient statistic for this model.

Then

$$\mathbb{E}(X_{(n)}) = a_n \cdot N + b_n$$

and so we can estimate N as

$$\widehat{N} = \frac{X_{(N)} - b_n}{a_n}.$$

11 Method of Moments

$$X_1, \ldots, X_n \overset{iid}{\sim} P_\theta, \quad \theta \in \Theta$$

Let

$$\begin{aligned}
\mu_1 &= \mathbb{E}_\theta(X_1) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \mathbb{E}(m_1) \\
\mu_2 &= \mathbb{E}_\theta(X_1^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) = \mathbb{E}(m_2) \\
&\vdots \\
\mu_k &= \mathbb{E}_\theta(X_1^k) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i^k\right) = \mathbb{E}(m_k).
\end{aligned}$$

We want to estimate $g(\mu_1, \mu_2, \ldots, \mu_k)$ and with the method of moments we use $g(m_1, m_2, \ldots, m_k)$ as our estimator.

However, the method of moments can lead to estimates outside of the sample space and non-unique estimates.

Example.

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), \quad \theta \in \Theta = \left[\frac{1}{4}, \frac{3}{4}\right]$$

Then

$$\mu_1 = \mathbb{E}(X_1) = \theta$$

and

$$g(\mu_1) = g(\theta) = \theta.$$

Now we plug in

$$g(m_1) = m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \in [0,1].$$

Since $X_i \in \{0,1\}$ then this estimator could result in an estimate outside of the sample space.

Example.

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta), \quad \theta \in \mathbb{R}_+$$

Then

$$\mu_1 = \mathbb{E}(X_1) = E(m_1) = \theta \equiv g(\theta)$$

and our estimate of θ is

$$\widehat{\theta} = g(m_1) = m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = T_1(X).$$

Now μ_2 is

$$\mu_2 = \mathbb{E}(X_1^2) = \mathbb{E}(m_2) = \theta + \theta^2$$

and then

$$\begin{aligned} m_2 &= \theta + \theta^2 = T_2(X) + T_2^2(X) \\ &\implies T_2^2(X) - T_2(X) - \mu_2 = 0. \end{aligned}$$

Example.

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\theta, \beta), \quad \alpha, \beta > 0$$

$$\begin{aligned} \mu_1 &= \mathbb{E}(X_1) = \mathbb{E}(m_1) = \frac{\alpha}{\beta} \\ \mu_2 &= \mathbb{E}(X_1^2) = \mathbb{E}(m_2) = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} \end{aligned}$$

and setting the sample moments equal to the population moments we have

$$m_1 = \frac{\alpha}{\beta}$$
$$m_2 = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}$$

and we must solve these two equations for α and β to get method of moments estimates.

12 Empirical Distribution Function

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} F$$

where $F(t) = P(X_1 \le t)$, for all $t \in \mathbb{R}$. Then based on a sample $x_1, \dots, x_n$ define the empirical distribution function $F_n(t)$ as

$$\begin{aligned} F_n(t) &= \frac{\#\{x_i \le t, i = 1, \dots, n\}}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} I(x_i \le t) \\ &= T_n \end{aligned}$$

where

$$\begin{aligned} I(X_1 \le t) &\overset{\text{iid}}{\sim} \text{Bernoulli}(F(t)) \\ T_n &\overset{\text{iid}}{\sim} \text{Binoimal}(n, F(t)) \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}(F_n(t)) &= F(t), \quad \forall t \in \mathbb{R} \\ \text{Var}(F_n(t)) &= \frac{F(t)(1-F(t))}{n} \le \frac{1}{4n}. \end{aligned}$$

Note that $F_n(t)$ is a consistent estimator as

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} P(|F_n(t) - F(t)| > \epsilon) = 0.$$

13 Least Square Estimator

We have data $(X_1, Y_1), \dots, (X_n, Y_n)$. Then consider the model

$$Y_i = g_i(x_i; \theta) + \epsilon_i$$

where g is a known function, *theta* is an unknown parameter, and ϵ_i is a random error term.

Then our estimate of θ is

$$\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - g_i(x_i; \theta))^2.$$

Example.

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

14 Maximum Likelihood

$$X \sim P_\theta, \quad \theta \in \Theta$$

We define the MLE of θ as

$$\widehat{\theta}(x) = \operatorname*{argmax}_{\theta \in \Theta} L(x; \theta)$$

where $L(x;\theta)$ is the likelihood function.

Example.

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}, \quad \theta \in \Theta = [0,1]$$

The likelihood function is

$$\begin{aligned} L(x_1, \ldots, x_n; \theta) &= \sum_{i=1}^{n} x_i (1-\theta)^{n = \sum x_i} \\ &= \end{aligned}$$

Example. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} uniform[0, \theta]$. Find the MLE for θ.

We want to estimate α and β. We have