STAT 653 - Notes
Introduction to Mathematical Statistics

September 21, 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 $X = (X_1, X_2, \dots, X_n)$, where X_i are random variables before the experiment is conducted.

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

<u>**Possible Estimators**</u>

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

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 $\Theta = \{\theta : \theta > 0\}$.

<u>**Possible Estimators**</u>

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

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

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

$F_\theta(t) = P(Y \le t)$ 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

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

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

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

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

Then

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

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,

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

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

□

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

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

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

Example. Let $X_1, X_2, \ldots, X_n \stackrel{\text{iid}}{\sim} N(\theta, 1)$. 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.

6 Maximum Likelihood

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