STAT5010 Advanced Statistical Inference **Fall 2023**

Lecture 1: Introduction, Sufficiency and Exponetial families

Lecturer: Tony Sit *Scribe: Peiming Lai and Zhengyao Sun*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Inference Problem

1. You are given a collection of probability mesures $\{P_\theta : \theta \in \Theta\}$ on a sample space $(\mathcal{X}, \mathcal{F})$, where $\mathcal{X}$ is a set and $\mathcal{F}$ is a σ-field on $\mathcal{X}$.
2. Observe $X \sim P_\theta$ for some $\theta \in \Theta$.
3. Infer θ from X.

Let $L(\theta, \delta(X))$ be the loss in estimating θ by $\delta(X)$, an estimator. Define $R(\theta, \delta) = E_{X\sim P_\theta} L(\theta, \delta)$ to be the risk functon of the estimator δ.

Example 1.1.1 *Let* $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1),\ \Theta = \mathbb{R},\ \mathcal{X} = \mathbb{R}^n$

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

$$P_\theta(A) = \frac{1}{\left(\sqrt{2\pi}\right)^n} \int_A e^{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2}} dx_1 \ldots dx_n$$

$$L(\theta, \delta(X)) = (\theta - \delta(x))^2.\ \textit{Proposed estimators: } \left\{ \begin{array}{l} \delta_1(X) = \bar{X} \\ \delta_2(X) = 0 \end{array} \right.$$

$$c.f.\ \left\{ \begin{array}{l} R(\theta, \delta_1) = E(\bar{X} - \theta)^2 = \frac{1}{n} \\ R(\theta, \delta_2) = E\left(\theta^2\right) = \theta^2 \end{array} \right.$$

* *To rule out estimators like* δ_2, *we need some strategies.*

Strategy 1 (Unbiasedness)

Definition 1.1 (Unbiasedness) *We say* $\delta(X)$ *is unbiased for* θ *if* $E_{X\sim p_\theta}(\delta(X)) = \theta, \forall \theta \in \Theta$

Since $E(\delta_1(x)) = E\left(\frac{1}{n}\sum_{i=1}^n x_i\right) = \theta$ whereas $E(\delta_2(x)) = 0$, we shall show later that δ_1 is the "best" amongst the class of all unbiased estimators in this problem.

Strategy 2 (Minimaxity)

We can look at $\sup_{\theta\in\Theta} R(\theta, \delta)$ for comparison and $\delta_{minimax} = \arg\min_\delta \sup_\theta R(\theta, \delta)$. In our example, $\sup_{\theta\in\mathbb{R}} R(\theta, \delta_1) = \frac{1}{n}, \sup_{\theta\in\mathbb{R}} R(\theta, \delta_2) = +\infty$. We shall show that δ_1 is the best minimax estimator for this problem.

Strategy 3 (Bayes / Average Risk Optmiality)

Assume θ is random and has a distribution π. In this case, we may compare estinitators via Bayes risk, which is defined as $E_{\theta\sim\pi}R(\theta,\delta)$.

In our example, let $\pi \sim N(\mu,\tau)$

Bayes risk of δ_1 is

$$E_{\theta\sim\pi}R(\theta,\delta_1) = E_{\theta\sim\pi}\left(\frac{1}{n}\right) = \frac{1}{n}$$

Bayes risk of δ_2 is

$$E_{\theta\sim\pi}R(\theta,\delta_2) = E_{\theta\sim\pi}\left(\theta^2\right) = \mu^2+\tau$$

In this case, we shall show that there is a third estimator δ_3 which is the "best".

Strategy 4 What happens when n is large?

In this case, by WLLN, $\delta_1(X) = \bar{X}_n \xrightarrow{p} \theta$. Also, $\delta_2 \xrightarrow{p} 0$. We shall analyse the asymptotic normality in more details if time allows.

1.2 Sufficiency

Definition 1.2 (Statistic) *A statistic T is a measurable function form $(\mathcal{X},\mathcal{F})$ to $\left(\mathbb{R}^k,\mathcal{B}_{\mathbb{R}^k}\right)$.*

Definition 1.3 (Sufficient Statistic) *A statistic T is said to be sufficient for θ (or $\{P_\theta : \theta\in\Theta\}$) of the conditional distribution of $(X \mid T)$ is free of $\theta, \forall\theta\in\Theta$.*

Definition 1.4 (Conditional Distribution)

1. Suppose (X,Y) are discrete random variables with a probability mass function $P(x,y)$ on a countable set $\mathcal{X}$. Then, the conditional distribution of X given $Y=y$ has a pmf, given by $P(X=x\mid Y=y)$

$$= \frac{p(x,y)}{\sum_{(z,y)\in\mathcal{X}} p(z,y)}$$

2. if (X,Y) has a joint probability density function $p(x,y)$ w.r.t. Lebesgue measure, then $(X\mid Y=y)$ has a pdf w.r.t. Lebesgue measure, given by

$$\frac{p(x,y)}{\int_{-\infty}^{\infty} p(z,y)dz}$$

In general, given (X,Y) a random vector in $\mathbb{R}^2$, for every $y\in\mathbb{R}$, one can define a distribution function $F_Y(\cdot)$ satisfying:

$$E_Y\left(F_Y(x)I(Y\in B)\right) = P(x\leqslant x, Y\in B) \quad \forall B \leq \mathcal{B}_{\mathbb{R}}$$

Example 1.2.1

1. Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(0,1), X_{n+1},\ldots,X_{2n} \overset{\text{iid}}{\sim} N(0,1)$. In this case $(X_1,\ldots,X_n)$ is sufficient for θ. Given $(X_1=x_1,\ldots,X_n=X_n)$ the distribution of $(X_{n+1}\ldots,X_{2n})$ has a density w.r.t. Lebesgue measure given by $f(x_{n+1},\ldots,x_{2n}) = \frac{1}{(\sqrt{2\pi})^n}e^{-\sum_{i=n+1}^{2n}x_i^2/2}$.
The joint distribution of $(X_1,\ldots,X_{2n}\mid X_1=x_1\ldots,X_n=x_n)$ is $\delta_{x_1}\ldots\delta_{X_n}\times N(0,1)$.

2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bin}(1, \theta)$, $\Theta = (0, 1)$,then we have $P(X_i = 1) = \theta$, $P(X_i = 0) = 1 - \theta$, for $i = 1, \ldots, n$ In this case $T(X) = \sum_{i=1}^n x_i$ is sufficient for θ. (Ex)
$$P(X_1 = x_1, \ldots, X_n = x_n \mid T(X)) \perp\!\!\!\perp \theta$$
3. Let $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N(\theta, 1)$, $(x_1, \ldots, x_n)$ and $T(x) = \sum_{i=1}^n x_i$ are both sufficient for θ. Actually given $T = t$,
$$(X_1, \ldots, X_n) \overset{\textbf{(Ex)}}{\sim} N\left(\left(\begin{array}{c} \frac{t}{n} \\ \vdots \\ \frac{t}{n} \end{array}\right), \left(\begin{array}{cccc} 1 - \frac{1}{n}, & -\frac{1}{n}, & \cdots & -\frac{1}{n} \\ -\frac{1}{n}, & 1 - \frac{1}{n}, & \cdots & \vdots \\ \vdots & \vdots & \ddots & 1 - \frac{1}{n} \end{array}\right)\right)$$

Definition 1.5 (Neyman-Fisher Factorisation Criterion, NFFC) *Suppose $\{P_\theta : \theta \in \Theta\}$ is a collection of probability measures on $(\mathcal{X}, F)$, which are dominated by a σ-finite measure γ. Let $X \sim P_\theta$ for some $\theta \in \Theta$, then T is sufficient for $\theta \Leftrightarrow P_\theta(x) = g_\theta(T(x))h(x)$ a.s. γ for some $g_\theta(\cdot)$ and $h(\cdot)$, where $P_\theta(\cdot) = \frac{dP_\theta}{\partial\gamma}, P_\theta(A) = \int_A P_\theta(x) d\gamma$ (a.s. γ means: $\gamma\{X : p_\theta(x) \neq g_\theta(T(x))h(x)\} = 0$)*

Proof: Assuming γ is a counting measure as a countable set $\mathcal{X}$ (i.e. X is discrete) Let the family of pmf's be given by $\{P_\theta : \theta \in \Theta\}$ discrete)

1. $\Leftarrow$: Suppose $p_\theta(x) = g_\theta(T(x))h(x), \quad \forall x \in \mathcal{X}$. Need to show that T is sufficient.
$$P_\theta(X = x \mid T(X) = t) = \frac{P_\theta(X = x, T(X) = t)}{P(T(X) = t)} = \begin{cases} 0 & T(x) \neq t \\ \frac{P(T(X)=t)}{P(X) \neq t} & T(X) = t \end{cases}$$
$$= \begin{cases} 0 & T(x) \neq t \\ \frac{g_\theta(t)h(x)}{\sum_{y \in x : T(y) = t} g_\theta(t)h(y)} & T(x) = t \end{cases}$$
$$= \begin{cases} 0 & T(x) \neq t \\ \frac{h(x)}{\sum_{y \in x : T(y) = t} h(y)} \perp\!\!\!\perp \theta & T(x) = t \end{cases}$$

2. $\Rightarrow$: Suppose is sufficient for θ, so
$$\begin{aligned} P_\theta(X = x) = P_\theta(X = x, T(X) = t) &= P_\theta(X = x \mid T(X) = t) P_\theta(T(X) = t) \\ &\triangleq h(x) g_\theta(t) \\ &\text{as } P_\theta(X = x \mid T(X) = t) \text{ is free of } \theta \text{ by definition} \end{aligned}$$

■

Example 1.2.2

1. $X_1 \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$:
$$\begin{aligned} P_\theta(X) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n (x_i - \theta)^2/2} \\ &= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n (x_i - \bar{x})^2/2}}_{h(x)} \underbrace{e^{-n(\bar{x} - \theta)^2/2}}_{g_\theta(\bar{x}) \text{ or } g_\theta(T)} \end{aligned}$$

2. $X_1 \ldots, X_n \overset{\text{iid}}{\sim}$ *Benoulli*(θ):

$$P_\theta(X) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \underbrace{\prod_{i=1}^n I(0 \le X_i \le 1)}_{h(x)}$$
$$= (\frac{\theta}{1-\theta})^{\sum_{i=1}^n x_i}(1-\theta)^n \prod_{i=1}^n I(0 \le X_i \le 1)$$

1.3 Exponential families

Definition 1.6 *The model $\{\mathbb{P}_\theta : \theta \in \Omega\}$ forms an s-dimensional exponential family if each $\mathbb{P}_\theta$ has density of the form:*

$$p(x;\theta) = \exp\left(\sum_{i=1}^s \eta_i(\theta)T_i(x) - B(\theta)\right) h(x)$$

- $\eta_i(\theta) \in \mathbb{R}$ are called the natural parameters.
- $T_i(x) \in \mathbb{R}$ are its sufficient statistics, which follows from NFFC.
- $B(\theta)$ is the log-partition function because it is the logarithm of a normalization factor:

$$B(\theta) = \log\left(\int \exp\left(\sum_{i=1}^s \eta_i(\theta)T_i(x)\right) h(x)d\mu(x)\right) \in \mathbb{R}$$

- $h(x) \in \mathbb{R}:$ base measure.

Example 1.3.1 *Let $X_1, \ldots, X_n \overset{iid}{\sim} N\left(\theta, \sigma^2\right), \Theta = \mathbb{R} \times (0, \infty)$*

$$P_\theta(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^n (x_i-\mu)^2/\left(2\sigma^2\right)}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^n x_i^2/\left(2\sigma^2\right)} e^{+\theta\sum_{i=1}^n x_i/\sigma^2} \cdot e^{-\frac{n\theta}{2\sigma^2}}$$

$$T_1(X) = \sum_{i=1}^n x_i^2), \quad \eta_1\left(\theta, \sigma^2\right) = -\frac{1}{2\sigma^2}$$
$$T_2(X) = \sum_{i=1}^n x_i, \quad \eta_2\left(\theta, \sigma^2\right) = -\frac{\theta}{\sigma^2}$$
$$B\left(\theta, \sigma^2\right) = \frac{nx^2}{2\sigma^2} - \frac{n}{2}\log\left(2\pi\sigma^2\right), \quad h(x) = 1 = \prod_{i=1}^n (X_i \in \mathbb{R})$$

Example 1.3.2 *Example Let $x_1 \ldots, x_n \overset{iid}{\sim}$ Cauchy i.e. $P_\theta(x) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}$ is the density of P_θ, w.r.t. Lebesgue measure, In this case, $X_1, \ldots X_n$ is sufficient. $T = \left(X_{(1)}, \ldots, X_{(n)}\right)$ is sufficient, where $\left(X_{(1),\ldots,}X_{(n)}\right)$ are the order statistics of X.* $\quad (X_{(1)} \leqslant X_{(2)}) \leqslant \ldots \leqslant X(n)$