STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Yudan Zou and Tom Lee Cheuk Lam

Lecture 3

Recap last lecture

Proof of (ii): First, we consider the case $k = 1$. Since $e^{s|x|} \leq e^{sx} + e^{-sx}$, we conclude that $|X|$ has an mgf that is finite in the neighborhood of 0, say $(-c, c)$ for once $c > 0$.
By using the inequality:

$$\left| e^{itx} \left\{ e^{iax} - \sum_{j=0}^{n} \frac{(iax)^j}{j!} \right\} \right| \leq \frac{|ax|^{n+1}}{(n+1)!}$$

We can write

$$\left| \phi_X(t+a) - \sum_{j=0}^{n} \frac{a^j}{j!} E\left\{ (iX)^j e^{iX} \right\} \right| \leq \frac{|a|^{n+1} E|X|^{n+1}}{(n+1)!}$$

which imples that for any $t \in \mathbb{R}$,

$$\phi_X(t+a) = \sum_{j=0}^{\infty} \frac{\phi_X^{(j)}(t)}{j!} a^j, \quad \text{for } |a| < c. \tag{*}$$

Similarly,(*) also holds for Y. That is, $\phi_Y(t+a) = \sum_{j=0}^{\infty} \{\phi_Y^{(j)}(t) a^j / j!\}$. Under the assumption that $m_X = m_Y < \infty$ in a neighbourhood of 0, X and Y have the same moment of all orders. Since $\phi_X^{(j)}(0) = \phi_Y^{(j)}(0)$ for all $j = 1, 2, ...$, which and * with $t = 0$ imply that ϕ_X and ϕ_Y are the same on the interval $(-c, c)$ and have the identical derivatives there.

Consider $t = c - \epsilon$ and $-c + \epsilon$ for an arbitrary small $\epsilon > 0$ in * and the result will follow in that ϕ_X and ϕ_Y will also agree on $(-2c + \epsilon, 2c - \epsilon)$ and hence on $(-2c, 2c)$. By the same argument, ϕ_X and ϕ_Y are the same on $(-3c, 3c)$ and so on. Hence $\phi_X(t) = \phi_Y(t)$ for all t and by (i), $F_X = F_Y$.

For the general case of $k > 2$, if $F_X \neq F_Y$, then part(i) concludes that there exists $t \in \mathbb{R}$ such that $\phi_X \neq \phi_Y$. Then $\phi_{t^T X}(1) \neq \phi_{t^T Y}(1)$, which implies that $F_{t^T X} \neq F_{t^T Y}$.But $m_X = m_Y < \infty$ in a neighborhood of $0 \in \mathbb{R}$ and by the result for $k = 1$, $F_{t^T X} = F_{t^T Y}$, this shows that $F_X = F_Y$.

2 10 ways of viewing a random variable (Cont'd)

2.9 Way # 9: Conditional probability

In our undergraduate study,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

where by convention $P(B \mid A) = 0$ when $P(A) = 0$. But the definition breaks down for uncountable χ. If $\nu \ll \mu$, then there exists a non-negative function φ such that

$$\nu(A) = \int_A \varphi d\,\mu, \quad \text{for any } A \in \mathcal{A}.$$

For example, we have (X, Y) with joint density $f(x, y)$ and X with marginal density $g(x)$, then the conditional density

$$\varphi(y|x) = \frac{f(x,y)}{g(x)}$$

Alternatively, we write $\varphi(x) = E(Y|X)$, which can be interpret as a random variable which takes the value $E(Y|X = x)$ with $P(X = x)$ (see STAT 5050).

2.10 Way # 10: Tail behavior

For a scalar random variable $X \sim F$, we say X has an exponential tail if

$$\lim_{a\to\infty} \frac{-\log(1 - F(a))}{Ca^r} = 1, \quad \text{for some } C > 0, r > 0$$

and an algebraic tail if

$$\lim_{a\to\infty} \frac{-log(1 - F(a))}{m \log a} = 1, \quad \text{for some } m > 0$$

Example 1. *Here are some examples:*

1. *Exponential:* $F(a) = 1 - e^{-\lambda a} \to c = \lambda, r = 1$
2. *Gaussian:* $F(a) = ... \to c = 2, r = 2$
3. *Student-t:* $m = \nu$ *(heavy-tail distributions/ extreme value theory)*

3 Sufficiency Principle

3.1 Introduction

Suppose $X_1, ...X_n \sim P_\theta$ for any unknown parameter $\theta \in \Omega, \Omega \subseteq \mathbb{R}^k$. Using n numbers $X_1, ...X_n$ to store the information and make inference about k features θ may waste storage space. Even worse, if n is large, the raw data $X_1, ..., X_n$ will become difficult to interpret. Therefore, we would like to produce a lower dimensional summary without losing information about θ (Data reduction).

3.2 Statistic and Sufficiency Principle

- **Statistic**: A statistic $T : \mathcal{X}^n \longrightarrow \mathcal{T}^m$ is a function of the data $X_1, ..., X_n$ and *free of any unknown parameter*.
- **Sufficiency Principle**: A statistic $T = T(X_1, ..., X_n)$ is sufficient for a model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if for any $t = T(x_1, ..., x_n)$, the conditional distribution $X_{1:n} \mid T(x_{1:n}) = t$ is free of θ.
  * The n-dimensional statistic $X_{1:n} = (X_1, ..., X_n)^T$ is a *trivial sufficient statistic* for $\mathcal{P}$.

Example 2. $T(X_1, ..., X_n) = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ *(sample mean), and* $T(X_1, ..., X_n) = S_n^2 = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ *(sample variance) are a statistic.*
* *If μ is unknown, then the population variance $\sigma^2 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$ is <u>not</u> a statistic.*

Example 3. *Let* $X_1, ..., X_n \overset{iid}{\sim} Bern(\theta)$ *for any* $\theta \in (0, 1)$*. Let* $T = T(X_{1:n}) = \sum_{i=1}^n X_i$*. Consider*

- *Case 1:* $\sum_{i=1}^n x_i \neq t, P_\theta(x_{1:n} \mid t) = 0.$

- *Case 2:* $\sum_{i=1}^n x_i = t$. *Consider* $\{X_{1:n} = x_{1:n}, T = t\} = \{X_{1:n} = x_{1:n}\}$ *as knowing all data* $x_{1:n}$ *gives more information than knowing* $t = T(x_{1:n})$. *Note that* $T \sim Bin(n, \theta)$, *we have*

$$\begin{aligned}
P_\theta(x_{1:n} \mid t) &= \frac{P_\theta(x_{1:n}, t)}{P_\theta(t)} \\
&= \frac{P_\theta(x_{1:n})}{P_\theta(t)} \qquad \frac{\textit{A likelihood function}}{\textit{Binomial distribution}} \\
&= \frac{\prod_{i=1}^n \{\theta^{x_i}(1-\theta)^{1-x_i}\}}{\binom{n}{t}\theta^t(1-\theta)^{1-t}} = \binom{n}{t}^{-1}
\end{aligned}$$

Hence, for any cases, $P_\theta(x_{1:n} \mid t)$ *is free of* θ, *so* $T(x_{1:n}) = \sum_{i=1}^n x_i$ *is a sufficient statistic for* $\mathcal{P} = Bern(\theta)$.

Example 4. *(Order Statistics) Let* $X_{1:n} \overset{iid}{\sim} P_\theta \in \mathcal{P}$ *for any model* $\mathcal{P}$, *then the order statistics* $T = (x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)})^T$ *are sufficient. To see why* T *is sufficient, note that given* T, *the possible values of* X *are in* $n!$ *permutations of* T. *By symmetry, we can see that each of their permutation has an equal probability of* $\frac{1}{n!}$

$$\begin{aligned}
p_\theta(X_1 = X_{(1)}, X_2 = X_{(2)}, ..., X_n = X_{(n)}) &= \frac{1}{n!} \\
p_\theta(X_1 = X_{(2)}, X_2 = X_{(1)}, ..., X_n = X_{(n)}) &= \frac{1}{n!} \\
&\cdots \\
p_\theta(X_1 = X_{(n)}, X_2 = X_{(n-1)}, ..., X_n = X_{(1)}) &= \frac{1}{n!}
\end{aligned}$$

Hence $X_{1:n} = x_{1:n} \mid T = t = \frac{1}{n!} \perp\!\!\!\perp \theta$ *thus* $T = (x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)})^\top$ *is a sufficient statistic.*

Theorem 1. *If* $X \sim P_\theta \in \mathcal{P}$ *and* $T = T(X)$ *is a sufficient statistic for* $\mathcal{P}$, *then for any decision procedure* δ, *there exists a (possibly randomized) decision procedure of equal risk that depends on* X *only through* $T = T(X)$ *only.*

To illustrate the concept of randomization, suppose, given an independent source of randomness, say $U \sim Unif(0, 1)$, we can always generate a new data set $X' = f(T(X), U)$ from the conditional distribution $p(X \mid T(X))$ and define a randomized procedure

$$\delta^*(X, U) \equiv \delta\{f(T(X), U)\} - \delta(X') \overset{d}{=} \delta(X)$$

Example 5. *Suppose X and Y are independent with common density*

$$f_\theta(x) = \begin{cases} \theta \exp(-\theta x) & \textit{for } x \geq 0 \\ 0 & \textit{otherwise}, \end{cases}$$

and let $U \sim unif(0, 1)$ *and independent of* X, Y. *Define* $T = X + Y$ *and define*

$$\tilde{X} = UT \text{ and } \tilde{Y} = (1 - U)T.$$

Let us find the joint density of $\tilde{X}$ and $\tilde{Y}$. The density of T is needed, and this can be found by smoothing. Because X and Y are independent,

$$\begin{aligned} P(T \le t \mid Y = y) &= P(X + Y \le t \mid Y = y) \\ &= \mathbb{E}\Big\{ I(X + Y \le t) \mid Y = y \Big\} \\ &= \int I(X + Y \le t) dF_X(x) \\ &= F_X(t - y). \end{aligned}$$

So $P(T \le t \mid Y) = F_X(t - Y)$ and

$$F_T(t) = P(T \le t) = \mathbb{E}\Big\{ F_X(t - Y) \Big\}.$$

This formula holds generally. Specializing to our specific problem, $F_X(t-Y) = 1 - \exp\big\{ - \theta(t-Y) \big\}$ on $Y < t$ and is zero on $Y \ge t$. Writing the expected value of this variable as an integral against the density of Y, for $t \ge 0$,

$$F_T(t) = \int_0^t \Big[1 - \exp\big\{ - \theta(t - y) \big\} \Big] \theta \exp(-\theta y) dy = 1 - \exp(-\theta t) - t\theta \exp(-\theta t)$$

Taking derivative, T has density

$$p_T(t) = F_T^{'}(t) = \begin{cases} t\theta^2 \exp(-\theta t) & \textit{for } t \ge 0 \\ 0 & \textit{otherwise}. \end{cases}$$

Because T and U are independent, they have the joint density

$$p_\theta(t, u) = \begin{cases} t\theta^2 exp(-\theta t) & \textit{for } t \ge 0, u \in (0,1) \\ 0 & \textit{otherwise} \end{cases}$$

From this,

$$p\left(\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \in B \right) = \int \int I\{tu, t(1-u)\} p_\theta(t, u) du dt$$

Changing variables to $x = ut, du = dx/t$ in the inner integral, and reversing the order of integration using Fubini's theorem,

$$p\left(\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \in B \right) = \int \int I\{x, t - x\} t^{-1} p_\theta(t, \tfrac{x}{t}) dt dx$$

Now a change of variables to $y = t - x$ in the inner integral gives

$$p\left(\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \in B \right) = \int \int I\{x, t\} (x - y)^{-1} p_\theta(x + y, \tfrac{x}{x+y}) dy dx$$

Thus $\tilde{X}$ and $\tilde{Y}$ have joint density

$$\frac{p_\theta(x + y, \frac{x}{x+y})}{x + y} = \begin{cases} \theta^2 exp\big\{ - \theta(x + y) \big\} & \textit{for } x, y \ge 0 \\ 0 & \textit{otherwise} \end{cases}$$

This density is the same as the joint density of X and Y, and so this calculation shows that the joint distribution of $\tilde{X}$ and $\tilde{Y}$ is the same as the joint distribution of X and Y. Considered as data that provide information about θ, the pair $(\tilde{X}, \tilde{Y})$ should be just as informative as (X, Y).

3.3 Neyman-Fisher Factorization Theorem

Suppose each $P_\theta \in \mathcal{P}$ has density $p(x_{1:n};\theta)$ with respect to a common σ -finite measure μ. That is, $dP_\theta/d\mu = p(x_{1:n};\theta)$, then $T = T(X_{1:n})$ is sufficient if and only if for any $\theta \in \Theta, x_{1:n} \in \mathcal{X}^n$,

$$p(x_{1:n};\theta) = g_\theta(T(x_{1:n}))h(x_{1:n})$$

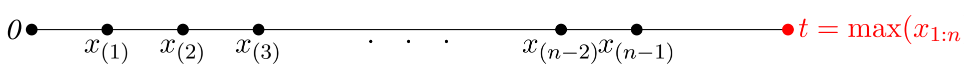
for some functions g_θ, h.
* A necessary and sufficient condition for $T(x_{1:n})$ to be sufficient is that the density $p(x_{1:n};\theta)$ can be factorized into two components, one of which depends on both $\theta, T(x_{1:n})$, and another one is free of θ.

Example 6. *Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)^T$, then we have*

$$\begin{aligned} p(x_{1:n};\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \\ &= \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right) \right\} \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \\ &= g_\theta(T(x_{1:n}))h(x_{1:n}) \end{aligned}$$

By the Neyman-Fisher factorization theorem, $T(X_{1:n}) = (\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)^T$ is a sufficient statistic for $\mathcal{P} = N(\mu, \sigma^2)$.

Example 7. *Let $X_1, ..., X_n \overset{iid}{\sim} Unif(0, \theta)$ for any $\theta > 0$. $T = T(X_{1:n}) = \max(X_{1:n})$ is a sufficient statistic for $\mathcal{P} = Unif(0, \theta)$.*
The intuition: think of $x_1, ..., x_n$ as n numbers on the real line $\mathbb{R}$, then the remaining $n-1$ numbers, given the maximum is fixed at $t = \max(x_{1:n})$, behave like $n-1$ iid random samples drawn from $Unif(0, t)$.



for some order statistics $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ of $x_1, ..., x_n$. To show that $t = \max(x_{1:n})$ is a sufficient statistic,

$$\begin{aligned} p(x_{1:n};\theta) &= \prod_{i=1}^{n} \left\{ \frac{1}{\theta} I(0 < x_i < \theta) \right\} \\ &= \frac{1}{\theta^n} I(x_{(n)} < \theta) I(0 < x_{(1)}) \\ &= g_\theta(T(x_{1:n}))h(x_{1:n}) \end{aligned}$$

By the Neyman-Fisher factorization theorem, $T = T(X_{1:n}) = \max(X_{1:n})$ is a sufficient statistic for $\mathcal{P} = Unif(0, \theta)$.

Proof of the Neyman-Fisher factorization theorem

Proof. *To begin, suppose $p_\theta \in p$ and $\theta \in \Omega$*

$$p(x;\theta) = g_\theta(T(x))h(x).$$

With respect to μ. Modifying h, we can assume without loss of generality that μ us a probability measure equivalent to the family $P = \{p_\theta : \theta \in \Omega\}$ [Equivalence referes to the situation where $\mu(N) = 0$ iff $p_\theta(N) = 0 \quad \forall \theta \in \Omega$].

Let E^* and P^* be the expectation and probability where $X \sim \mu$. Let G^* and G_θ denote marginal distribution for $T(x)$ where $X \sim \mu$ and $X \sim P_\theta$ respectively. Let Q be the conditional distribution for X given T where $X \sim \mu$.

To find the densities for T,

$$\begin{aligned} E_\theta f(T) &= \int f(T(x)) g_\theta(T(x)) h(x) d\mu(x) \\ &= E^*\{f(T) g_\theta(T) h(X)\} \\ &= \int \int f(t) g_\theta(T) h(x) dQ_t(x) dG^*(t) \\ &\triangleq \int f(t) g_\theta(t) \omega(t) dG^*(t), \end{aligned}$$

where $\omega(t) = \int h(x) dQ_t(x)$. If f is an indicator function this shows that G_θ has the density $g_\theta \omega(t)$ with respect to G^*. Next we define $\widetilde{Q}$ to have density $h/\omega(t)$ with respect to $Q(t)$, so that

$$\widetilde{Q}_t(B) = \int_B \frac{h(x)}{\omega(t)} dQ_t(x),$$

the conditional distribution of X given T under P_θ is independent of Q.

$$\begin{aligned} E_\theta \int (X, T) &= E^*\{f(X,T) g_\theta(T) h(x)\} \\ &= \iint f(x,t) g_\theta(t) h(x) dQ_t(x) dG^*(t) \\ &= \iint f(x,t) d\widetilde{Q}_t(x) dG_\theta(t) \end{aligned}$$

By the definition of conditional distribution, it shows that $\widetilde{Q}$ is a conditional distribution of X given under P_θ. Because $\widetilde{Q}$ does not depend on Q, it is sufficient statistic. (TBC)