STATS 300A: Theory of Statistics I **Autumn 2016/17**

Lecture 3: October 4

Lecturer: Joseph Romano *Scribes: Kenneth Tay*

3.1 Minimal Sufficiency

Definition 3.1 *A sufficient statistic $T = T(X)$ is **minimal sufficient** if for any other sufficient statistic T', T is a function of T'. (In other words, if $T'(x) = T'(y)$, then $T(x) = T(y)$.)*

Intuitively, a minimal sufficient statistic represents the greatest possible reduction of the data.

Theorem 3.2 *Suppose that X has density $p_\theta(x)$ w.r.t. some measure μ. Suppose that a statistic T has the following property: for $x, y \in \mathcal{S}$ (the sample space from which data is drawn), the ratio $\dfrac{p_\theta(x)}{p_\theta(y)}$ does not depend on θ if and only if $T(x) = T(y)$.*

Then T is minimal sufficient.

Proof: To make the argument easier, assume that the densities have common support across all θ. (The argument can be modified in the case without common support.)

First, we show that T is sufficient. Let $\mathcal{T} = \{t : t = T(x) \text{ for some } x \in \mathcal{S}\}$ (i.e. the image of the sample space S under T). Define partition sets $A_t := \{x : T(x) = t\}$. For each A_t, choose an element $x_t \in A_t$.

Note that x and $x_{T(x)}$ belong to the same A_t by construction. Hence, $T(x) = T\left(x_{T(x)}\right)$. By the assumption on T, we can conclude that

$$\frac{p_\theta(x)}{p_\theta(x_{T(x)})}$$

does not depend on θ. Hence, we can define

$$h(x) = \frac{p_\theta(x)}{p_\theta(x_{T(x)})}$$

as a function of x. Letting $g(t, \theta) = p_\theta(x_t)$, we can write the density of x as

$$p_\theta(x) = \frac{p_\theta(x) p_\theta(x_{T(x)})}{p_\theta(x_{T(x)})} = h(x) \cdot g(T(x), \theta).$$

By the Fisher-Neyman Factorization Theorem, we conclude that T is sufficient.

To show minimality: let T' be any other sufficient statistic. By the Factorization Theorem, $\exists g', h'$ such that

$$p_\theta(x) = g'(T'(x), \theta) h'(x).$$

Let x, y satisfy $T'(x) = T'(y)$. Then

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{g'(T'(x),\theta)h'(x)}{g'(T'(y),\theta)h'(y)} = \frac{h'(x)}{h'(y)},$$

which does not depend on θ. Hence, by assumption on T, we have $T(x) = T(y)$.

Since this holds for all x and y, T is minimal sufficient. ■

3.1.1 Example

Let $X_1, \ldots, X_n$ iid, $X_i \sim \mathcal{N}(\mu, \sigma^2)$. For $x = (x_1, \ldots x_n)$ and $y = (y_1, \ldots y_n)$, we have

$$\begin{aligned}\frac{p(x,(\mu,\sigma^2))}{p(y,(\mu,\sigma^2))} &= \frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right]}{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right]} \\ &= \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2-(y_i-\mu)^2\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\left(\sum x_i^2-\sum y_i^2\right)+\frac{\mu}{\sigma^2}\left(\sum x_i-\sum y_i\right)\right].\end{aligned}$$

- If $\mu = 0$ and σ^2 unknown, then $T = \sum X_i^2$ is minimal sufficient.
- If σ^2 fixed and known, μ unknown, then $T = \bar{X}$ is minimal sufficient.
- If both are unknown, then $T = \left(\sum X_i, \sum X_i^2\right)$ is minimal sufficient.

3.2 Convex Functions

Definition 3.3 *A real-valued function ϕ defined on an open interval $I = (a, b)$ (with possibly $a = -\infty$ and/or $b = +\infty$) is **convex** if for any $a < x < y < b$ and any $0 < \gamma < 1$,*

$$\phi(\gamma x + (1-\gamma)y) \leq \gamma\phi(x) + (1-\gamma)\phi(y).$$

*If the inequality above is strict, then ϕ is **strictly convex**.*

Theorem 3.4 (Jensen's Inequality) *If ϕ is convex on an open interval I and X is a random variable satisfying*

$$P(X \in I) = 1 \qquad \text{and} \qquad \mathbb{E}|X| < \infty,$$

then

$$\phi(\mathbb{E}X) \leq \mathbb{E}\left[\phi(X)\right].$$

If ϕ is strictly convex, then the inequality above is strict unless X is constant with probability 1.

Proof: Let $t = \mathbb{E}X$. Let $y = L(x)$ be the equation of the tangent line to ϕ at the point $x = t$, i.e.

$$L(t) = \phi(t) \qquad \text{and} \qquad L(x) \leq \phi(x) \text{ for } x \in I.$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\phi(X)\right] &\geq \mathbb{E}L(X) && \text{(by definition of } L) \\
&= L(\mathbb{E}X) && \text{(as } L \text{ is linear)} \\
&= L(t) \\
&= \phi(t) \\
&= \phi(\mathbb{E}X).
\end{aligned}$$

■

Example of a convex loss function: $L(\theta, d) = |d - g(\theta)|^p$.

Example of a non-convex loss function: $L(\theta, d) = \begin{cases} 1 & \text{if } |d - g(\theta)| > \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$

3.3 Ancillary and Complete Statistics

Definition 3.5 *A statistic $V = V(X)$ is* ***ancillary*** *if its distribution does not depend on θ.*

V is ***first-order ancillary*** *if $E_\theta V(X)$ does not depend on θ.*

In some sense, an ancillary statistic contains ”useless” information, in that knowing it does not tell us anything about θ. “First order ancillary” is a weaker notion of “ancillary”: an ancillary statistic is clearly first-order ancillary, but the reverse need not be true.

Example: $X_1, \ldots, X_n$ iid, $X_i \sim \mathcal{N}(\theta, 1)$. Then the distribution of $X_1 - X_2$ is $\mathcal{N}(0, 1)$, and hence is ancillary.

A sufficient statistic T is “most successful” in data reduction if no non-constant function f of T is first-order ancillary, i.e.,

$$\mathbb{E}_\theta f(T) = c \text{ for all } \theta \quad \Rightarrow \quad f(T) = c \text{ for all } \theta.$$

We can formalize this in the following definition:

Definition 3.6 *A statistic T is* ***complete*** *if*

$$\mathbb{E}_\theta f(T) = 0 \textit{ for all } \theta \quad \Rightarrow \quad f = 0 \textit{ with probability 1.}$$

Theorem 3.7 (Basu's Theorem) *If T is complete and sufficient and V is ancillary, then T and V are independent.*

Proof: Let $p_A = P(V \in A)$. Since V is ancillary, p_A does not depend on θ.

Let $\eta_A(t) = P(V \in A | T = t)$. Since T is sufficient, η_A does not depend on θ as well.

For all θ, using the law of iterated expectation, we have

$$\begin{aligned} \mathbb{E}_\theta \eta_A(T) &= \mathbb{E}\left[P(V \in A|T)\right] \\ &= P(V \in A) \\ &= p_A, \\ \mathbb{E}_\theta\left[\eta_A(T) - p_A\right] &= 0. \end{aligned}$$

Since T is complete, we must have $\eta_A(T) = p_A$ with probability 1. Hence, T and V are independent. ■

3.3.1 Example

$X_1, \dots, X_n$ iid, $X_i \sim U(0, \theta)$. We know that $T = \max(X_1, \dots X_n)$ is sufficient. We will now show that it is complete.

First, compute the density of T. Since

$$P_\theta(T \leq t) = \prod_{i=1}^n P_\theta(X_i \leq t) = \left(\frac{t}{\theta}\right)^n,$$

The density of T is $\dfrac{nt^{n-1}}{\theta^n}$.

Now, assume that $\mathbb{E}_\theta f(T) = 0$ for all θ. Then

$$\begin{aligned} \int_0^\theta f(t)\frac{nt^{n-1}}{\theta^n} &= 0 \quad \forall\theta, \\ \int_0^\theta f(t)t^{n-1}dt &= 0 \quad \forall\theta. \end{aligned}$$

Taking the derivative w.r.t θ, we have $f(\theta)\theta^{n-1} = 0$ for all θ, i.e. $\theta = 0$. Thus T is complete.

3.3.2 Special case of exponential families

Theorem 3.8 *Assume that X is taken from the exponential family model*

$$p(x, \eta) = \exp\left[\sum_{i=1}^s \eta_i T_i(x) - A(\eta)\right] h(x)$$

of full rank. Then $T = (T_1(X), \dots, T_s(X))$ is complete and sufficient.

Proof: (Sketch of proof) Assume f is such that $\mathbb{E}_\eta f(T) = 0$ for all θ. Decompose f into its positive and negative parts: $f = f^+ - f^-$. Then

$$\begin{aligned} \mathbb{E}_\eta f^+(T) = \mathbb{E}_\eta f^-(T) \quad \forall\theta, \\ \int f^+(t)\exp\left[\sum \eta_i t_i\right] d\nu(t) = \int f^-(t)\exp\left[\sum \eta_i t_i\right] d\nu(t) \end{aligned}$$

for some measure ν. Handwave: The LHS is something like the characteristic function of f^+, while the RHS is that of f^-. By uniqueness of characteristic functions, we have $f^+ = f^-$, i.e. $f = 0$. ■

3.3.3 Rao-Blackwell Theorem

The Rao-Blackwell Theorem is a statement of how we can, from an existing estimator δ, construct another estimator which has better risk than δ. (Recall that risk is the "average loss" over all possible data X, and that risk is a function of θ.)

Theorem 3.9 (Rao-Blackwell Theorem) *Assume that T is a sufficient statistic.*

Assume that we have a loss function $L(\theta, d)$ which is strictly convex in d, and that $\delta(X)$ is an estimator of $g(\theta)$ with finite risk $R(\theta, \delta)$.

Let $\eta(t) = \mathbb{E}\left[\delta(X)|T(X) = t\right]$. Then

$$R(\theta, \eta) < R(\theta, \delta)$$

unless $\delta = \eta$ with probability 1 (i.e. δ was a function of T to begin with).

Proof: Fix θ. Let $\phi(d) = L(\theta, d)$.

Applying Jensen's inequality to the conditional distribution of $\delta(X)|T(X) = t$:

$$\begin{aligned} \phi(\mathbb{E}\left[\delta(X)|T(X) = t\right]) &< \mathbb{E}\left[L(\theta, \delta(X))|T(X) = t\right], \\ \phi(\eta(t)) &< \mathbb{E}\left[L(\theta, \delta(X))|T(X) = t\right]. \end{aligned}$$

Taking expectations on both sides (i.e. averaging over X), we get

$$\begin{aligned} R(\theta, \eta) &< \mathbb{E}\left[L(\theta, \delta(X))\right] \\ &= R(\theta, \delta). \end{aligned}$$

■