Stat210A: Theoretical Statistics **Lecture Date: September 9, 2014**

Minimial sufficiency and completeness

Lecturer: Michael I. Jordan *Scribe: Fan Yang*

1 Some basic probability comments

Here we look at the idea of a conditional probability from a more rigorous point of view and start by establishing some basic probability results first.

1.1 Probability space, random variables

Let us assume the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a sample space with possible realizations ω, $\mathbb{P}$ a probability measure and $\mathcal{F}$ the σ-algebra containing event sets which are subsets of Ω. The probability measure assigns positive scalars to sets in $\mathcal{F}$. A random variable is a (measurable) function from the sample space to another space E (with $E = \mathbb{R}$ or $E = \mathbb{R}^n$, else called random element), i.e. $X : \Omega \to E$. For example in the case $E = \mathbb{R}$, the inverse image of $x \in \mathbb{R}$ is $\{\omega : X(\omega) = x\}$. So by the random variable you induce a probability distribution μ defined on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ and is defined by the probability measure $\mathbb{P}$ as $\mu(A) = \mathbb{P}(X^{-1}(A))$. Note: The randomness of the random variable comes from the domain of X, but X itself as a function is not random.

1.2 Conditional probabilities and expectations

Conditioning is often taught using conditional probability distributions, e.g. defining the conditional probabilities densities as $p(x|y) = \frac{p(x,y)}{p(y)}$ using Bayes' Theorem. Though probabilities are often considered to be more fundamental, especially in undergrad classes, it is not necessarily the right way to think about it. In the particular above definition when one conditions on a random variable, it is even unclear whether such an object exists.

Therefore, instead of conditional probability distributions let us look at conditional expectations which have a clear interpretation first. Conditional probability distributions can then be defined using conditional expectations. In fact expectations are "equally fundamental" since expectations can always be thought of as probabilities of indicator functions, i.e. $P(X \in A) = \mathbb{E}\, \mathbf{1}_{X \in A}$.

Let us look at the well-known tower property and the objects involved more carefully

$$\mathbb{E}\ \mathbb{E}\,(X|Y) = \mathbb{E}\, X$$

The conditional expectation $\mathbb{E}\,(X|Y)$ itself is a random variable, thus a measurable function which acts on the sample space, i.e. $\mathbb{E}\,(X|Y)(\omega)$ and is constant on the sets $\{\omega \in \Omega : Y(\omega) = y\}$. If you take the expectation of this random variable (take the weighted average over the sample space) you obtain the expectation of X.

The conditional distribution on the random variable Y is also a random variable defined as

$$P(X \in A|Y) = \mathbb{E}\,(\mathbf{1}_{X \in A}|Y)$$

The total probability can then be computed as

$$\begin{aligned} P(X \in A) &= \mathbb{E}\ P(X \in A|Y) \\ &= \int P(X \in A|y)\mathrm{d}P_Y(y) \end{aligned}$$

2 Sufficiency and Completeness

Why is sufficiency important? Originally it was intended to save disk space on the computer by only saving the sufficient statistics for estimating parameters from the data. Nowadays this problem is obsolete since disk space is often not the limiting factor. Still sufficient statistics turn out to be useful for answering the question: How does the data relate to the parameter? We will e.g. see that MLE comes from sufficiency.

2.1 Minimal sufficiency

We have defined sufficient statistics for family of distributions $\mathcal{P}$ last lecture. The random variable/vector X itself is always a sufficient statistic. The question is: Can I reduce it further? This motivates the notion of a *minimal sufficient* statistic.

Let T be sufficient. Suppose that $T = f(\tilde{T})$ (with f measurable, in general many-to-one), then $\tilde{T}$ is sufficient, i.e.

$$p_\theta(x) = h(x)g_\theta(T(x)) = h(x)\tilde{g}_\theta(\tilde{T}(x))$$

with $\tilde{g}_\theta = g_\theta \circ f$.

Definition 1 (Minimal sufficiency, Def 3.9 in Keener.)**.** A statistic T is minimal sufficient if T is sufficient and for all sufficient $\tilde{T}$, there exists a function f s.t. $T = f(\tilde{T})$ (a.e. $\mathcal{P}$). Here (a.e. $\mathcal{P}$) means that the set where it fails is a null set for every $P \in \mathcal{P}$.

Showing minimal sufficiency is non-trivial. A general trick is use the following fact: The likelihood, viewed as a function of θ yields the "likelihood shape", which in turn is minimal sufficient. The following theorem makes this explicit

Theorem 2 (Thm 3.11 in Keener)**.** *Suppose that the family $\mathcal{P}$ is dominated (i.e. are absolutely continuous with respect to some measure μ) and that there is a sufficient statistic T. Then (by the factorization theorem) we can write for the density*

$$p_\theta(x) = g_\theta(T(x))h(x) \qquad (a.e.\ \mu)$$

If $p_\theta(x) \propto_\theta p_\theta(y)$ (proportional when viewed as a function of θ) implies $T(x) = T(y)$, then T is minimal sufficient.

The intuition behind it is that if the observations x, y are actually indistinguishable from the point of view of inference (likelihood shape), then if T is minimal, it should not make a distinction between x, y either.

Proof. Suppose that $\tilde{T}$ is sufficient. Then

$$p_\theta(x) = \tilde{g}_\theta(\tilde{T}(x))\tilde{h}(x)$$

We now prove by contradiction: Given $p_\theta(x) \propto_\theta p_\theta(y) \implies T(x) = T(y)$ suppose T is not minimal, thus not a function of $\tilde{T}$. Then there exists x, y such that

$$\tilde{T}(x) = \tilde{T}(y) \qquad \text{but} \qquad T(x) \neq T(y)$$

and consequently

$$p_\theta(x) = \tilde{g}_\theta(\tilde{T}(x))\tilde{h}(x) \propto_\theta \tilde{g}_\theta(\tilde{T}(y))\tilde{h}(y) = p_\theta(y)$$

But since $T(x) \neq T(y)$ the original implication is not valid which is a contradiction. □

Example 3. Let

$$X_i \overset{i.i.d}{\sim} f_\theta(x) = \frac{1+\theta x}{2} \qquad |x| \leq 1, \theta \in [-1,1]$$

Claim: The order statistics $X_{(1)}, X_{(2)}, \ldots$ (X_i sorted in increasing order) are minimal sufficient.
Proof: We write

$$p_\theta(x) = \frac{1}{2^n}\prod_{i=1}^{n}(1+\theta x_i) = \frac{1}{2^n}\prod_{i=1}^{n}(1+\theta x_{(i)})$$

which is a polynomial of θ. They have the same roots at $\{-\frac{1}{x_i}\}$. Now suppose that $p_\theta(x) \propto_\theta p_\theta(y)$, so that they have the same roots, i.e. $\frac{1}{x_{(i)}} = \frac{1}{y_{(i)}}$. Hence $T(x) = \{x_{(i)}\} = \{y_{(i)}\} = T(y)$.

Example 4 (Ex. 3.13. in Keener)**.**

$$X_i \overset{i.i.d}{\sim} \frac{1}{2}\mathrm{e}^{-|x-\theta|}$$

Note that this distribution is not from the exponential family. Again the order statistics are minimal.

Definition 5 (Full-rank)**.** An exponential family $\mathcal{P}$ is considered full-rank if the image $\eta(\Omega)$ contains an open set in $\mathbb{R}^s$.

Example 6 (Ex. 3.12. in Keener)**.** Let $\mathcal{P}$ be an s-parameter exponential family.

$$p_\theta(x) = h(x)\exp(\eta(\theta)^T T(x) - B(\theta))$$

Suppose $p_\theta(x) \propto_\theta p_\theta(y)$, which implies that

$$\mathrm{e}^{\eta(\theta)^T T(x)} \propto \mathrm{e}^{\eta(\theta)^T T(y)} \implies \eta(\theta)^T T(x) = \eta(\theta)^T T(y) + c$$

For all θ_0, θ_1 pairs in Ω we have that

$$(\eta(\theta_0) - \eta(\theta_1))^T (T(x) - T(y)) = 0 \tag{1}$$

This is an orthogonality statement, so it doesn't generally imply that $T(x) = T(y)$. However if the exponential family is full-rank, the image of $\eta(\theta_0) - \eta(\theta_1)$ for all $\theta_0, \theta_1 \in \Omega$ "includes all directions". In this case, equality (1) implies that $T(x) = T(y)$ and T is minimal sufficient.

Note: An interesting reading on exponential families is Lawrence Brown's monograph (Brown (1986)) covering duality and convexity and their relationship to exponential families.

3 Completeness

Definition 7 (Completeness)**.** A statistic T is complete for a family $\mathcal{P}$, if

$$\mathbb{E}_\theta f(T) = c \ \forall\theta \implies f(T) = c \qquad (\text{ a.e. } \mathcal{P})$$

Example 8. Let

$$X_i \overset{i.i.d}{\sim} \text{Unif}(0, \theta)$$

where $\theta \in (0, \infty)$.

$T(\max_i(X_i))$ is sufficient by the factorization theorem. To find the density we observe

$$P_\theta(T \leq t) = P_\theta(X_1 \leq t) \cdots P_\theta(X_n \leq t) = \left(\frac{t}{\theta}\right)^n$$

Therefore, by taking the derivative we arrive at

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

Suppose $\mathbb{E}_\theta f(T) = c$ for all θ. Then, because

$$\mathbb{E}_\theta(f(T) - c) = \frac{n}{\theta^n}\int_0^\theta (f(t) - c)t^{n-1}\mathrm{d}t = 0$$

holds for all θ, we have that $f(t) - c = 0$ (see fact 4 in Section 1.4. in Keener). Note that this doesn't hold for a family with the same distributions but where θ is in the range e.g. $[1, \infty)$.

Theorem 9. *If T is sufficient and complete, then it is minimal sufficient.*

Proof. Let $\tilde{T}$ be minimal sufficient. Let T and $\tilde{T}$ be bounded random variables. By minimality we have $\tilde{T} = f(T)$. Let $g(\tilde{T}) = \mathbb{E}_\theta(T|\tilde{T})$ which is not dependent on θ by sufficiency. The tower property yields $\mathbb{E}_\theta g(\tilde{T}) = \mathbb{E}_\theta(T)$. This implies that $\mathbb{E}_\theta(T - g(\tilde{T})) = 0$ and by completeness $T = g(\tilde{T})$. Because also $\tilde{T} = f(T)$, we have a one-to-one mapping between T and $\tilde{T}$. □

Definition 10 (Ancillarity)**.** *V* is ancillary, if its distribution does not depend on θ.

Theorem 11 (Basu)**.** *Let T be sufficient and complete. Let V be ancillary. This implies that T is independent of V for all θ.*

References

Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, pages i–279.