Lecture 21: Sufficient, Ancillary, and Complete Statistics

Mathematical Statistics I, MATH 60061/70061

Tuesday December 7, 2021

Reference: Casella & Berger, 6.2.4

Complete statistic

Let $\{f_T(t \mid \theta); \theta \in \Theta\}$ be a family of PDFs (or PMFs) for a statistic $T = T(\boldsymbol{X})$. The family is called **complete** if the following condition holds:

$$E_\theta(g(T)) = 0, \ \forall \theta \in \Theta \implies P_\theta(g(T) = 0) = 1, \ \forall \theta \in \Theta.$$

In other words, $g(T) = 0$ almost surely for all $\theta \in \Theta$. Equivalently, $T(\boldsymbol{X})$ is called a **complete statistic**.

This means, the only function of T that is an unbiased estimator of zero is the function that is zero itself (with probability 1).

Binomial complete sufficient statistic

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Bern}(\theta)$ with parameter $0 < \theta < 1$.
Then $T(\boldsymbol{X}) = X_1 + \cdots + X_n$ is a complete statistic.

Binomial complete sufficient statistic

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Bern}(\theta)$ with parameter $0 < \theta < 1$. Then $T(\boldsymbol{X}) = X_1 + \cdots + X_n$ is a complete statistic.

We know $T \sim \mathrm{Bin}(n, \theta)$. Suppose $E_\theta(g(T)) = 0, \ \forall \theta \in (0,1)$. It suffices to show that $P_\theta(g(T) = 0) = 1$ for all $\theta \in (0,1)$. Write

$$E_\theta(g(T)) = \sum_{t=0}^{n} g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = (1-\theta)^n \sum_{t=0}^{n} g(t) \binom{n}{t} r^t,$$

where $r = \theta/(1-\theta)$. For $E_\theta(g(T)) = 0$, it must be that

$$\sum_{t=0}^{n} g(t) \binom{n}{t} r^t = 0.$$

Since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$, for $t = 0, 1, \ldots, n$. Therefore, $P_\theta(g(T) = 0) = 1$ for all $\theta \in (0,1)$ and $T(\boldsymbol{X})$ is a complete statistic.

Basu's Theorem

If $T = T(\boldsymbol{X})$ is a complete and sufficient statistic, then $T(\boldsymbol{X})$ is independent of every ancillary statistic S.

Basu's Theorem

If $T = T(\boldsymbol{X})$ is a complete and sufficient statistic, then $T(\boldsymbol{X})$ is independent of every ancillary statistic S.

We give a proof only for discrete distributions.

Let $S(\boldsymbol{X})$ be any ancillary statistic. Then $P(S(\boldsymbol{X}) = s)$ does not depend on θ since $S(\boldsymbol{X})$ is ancillary. Also, the conditional probability,

$$P(S(\boldsymbol{X}) = s \mid T(\boldsymbol{X}) = t) = P(\boldsymbol{X} \in \{\boldsymbol{x} : S(\boldsymbol{x}) = s\} \mid T(\boldsymbol{X} = t)),$$

does not depend on θ because $T(\boldsymbol{X})$ is a sufficient statistic.

To show that $S(\boldsymbol{X})$ and $T(\boldsymbol{X})$ are independent, it suffices to show that

$$P(S(\boldsymbol{X}) = s \mid T(\boldsymbol{X}) = t) = P(S(\boldsymbol{X}) = s)$$

for all $t \in \mathcal{T}$.

Now,

$$P(S(\boldsymbol{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\boldsymbol{X}) = s \mid T(\boldsymbol{X}) = t) P_\theta(T(\boldsymbol{X}) = t).$$

Also, since $\sum_{t \in \mathcal{T}} P_\theta(T(\boldsymbol{X}) = t) = 1$, we can write

$$P(S(\boldsymbol{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\boldsymbol{X}) = s) P_\theta(T(\boldsymbol{X}) = t).$$

Therefore, if we define the statistic

$$g(t) = P(S(\boldsymbol{X}) = s \mid T(\boldsymbol{X}) = t) - P(S(\boldsymbol{X}) = s),$$

the above two equations show that

$$E_\theta(g(T)) = \sum_{t \in \mathcal{T}} g(t) P_\theta(T(\boldsymbol{X}) = t) = 0,$$

for all θ. Since $T(\boldsymbol{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible $t \in \mathcal{T}$.

Example

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(0, \theta)$, where $\theta > 0$. Then $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent.

Example

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(0, \theta)$, where $\theta > 0$. Then $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent.

We know $X_{(n)}$ is sufficient for θ (Lecture 19). If we can show that $T(\boldsymbol{X}) = X_{(n)}$ is complete and $S(\boldsymbol{X}) = X_{(1)}/X_{(n)}$ is ancillary, then the result will follow from Basu's Theorem.

Example

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(0, \theta)$, where $\theta > 0$. Then $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent.

We know $X_{(n)}$ is sufficient for θ (Lecture 19). If we can show that $T(\boldsymbol{X}) = X_{(n)}$ is complete and $S(\boldsymbol{X}) = X_{(1)}/X_{(n)}$ is ancillary, then the result will follow from Basu's Theorem.

Note that $\mathrm{Unif}(0, \theta)$ is a scale family,

$$f_X(x \mid \theta) = \frac{1}{\theta} I(0 < x < \theta) = \frac{1}{\theta} f_Z\left(\frac{x}{\theta}\right),$$

where $f_Z = I(0 < x < 1)$ is the standard Uniform PDF. Also, $S(\boldsymbol{X})$ is scale invariant, since for $W_i = cX_i$, $i = 1, \ldots, n$,

$$S(\boldsymbol{W}) = \frac{W_{(1)}}{W_{(n)}} = \frac{cX_{(1)}}{cX_{(n)}} = \frac{X_{(1)}}{X_{(n)}} = S(\boldsymbol{X}).$$

Therefore, $S(\boldsymbol{X})$ is ancillary.

The PDF of $T(\boldsymbol{X}) = X_{(n)}$ is given by

$$\begin{aligned} f_T(t) &= n f_X(t)[F_X(t)]^{n-1} \\ &= n\frac{1}{\theta} I(0 < t < \theta) \left(\frac{t}{\theta}\right)^{n-1} \\ &= \frac{nt^{n-1}}{\theta^n} I(0 < t < \theta). \end{aligned}$$

Suppose $E_\theta(g(T)) = 0$ for all $\theta > 0$, i.e.,

$$\int_0^\theta g(t)\frac{nt^{n-1}}{\theta^n} dt = 0 \quad \forall\theta > 0.$$

This implies that for all $\theta > 0$,

$$\int_0^\theta g(t)t^{n-1}dt = 0 \implies \frac{d}{d\theta}\int_0^\theta g(t)t^{n-1}dt = 0 \implies g(\theta)\theta^{n-1} = 0.$$

Therefore,

$$E_\theta(g(T)) = 0 \implies P_\theta(g(T) = 0) = 1.$$

So, $T(\boldsymbol{X}) = X_{(n)}$ is a complete statistic.

Complete statistics in the Exponential family

Suppose $X_1, \ldots, X_n$ are iid from the **Exponential family**

$$f_X(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $d \leq k$. Then

$$\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is sufficient for $\boldsymbol{\theta}$. If the natural parameter space

$$\{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k) : \eta_j = w_j(\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$$

contains an open set in $\mathbb{R}^k$, $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ is **complete**. For the most part, this means:

- $\boldsymbol{T}(\boldsymbol{X})$ is complete if $d = k$ (full Exponential family)
- $\boldsymbol{T}(\boldsymbol{X})$ is not complete if $d < k$ (curved Exponential family)

Independence between Normal sample mean and variance

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Both parameters are unknown.

We showed in Lecture 14 that $\bar{X}$ and S^2 are independent. An easier way to this result with Basu's Theorem:

Consider the $\mathcal{N}(\mu, \sigma_0^2)$ family, where σ_0^2 is fixed and known. The PDF of $X \sim \mathcal{N}(\mu, \sigma_0^2)$ is

$$\begin{aligned} f_X(x \mid \mu) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x-\mu)^2/2\sigma_0^2} I(x \in \mathbb{R}) \\ &= \frac{I(x \in \mathbb{R}) e^{-x^2/2\sigma_0^2}}{\sqrt{2\pi\sigma_0^2}} e^{-\mu^2/2\sigma_0^2} e^{(\mu/\sigma_0^2)x} \\ &= h(x)c(\mu)\exp\{w_1(\mu)t_1(x)\}. \end{aligned}$$

The statistic $T = T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic. Because $d = k = 1$, T is complete.

The $\mathcal{N}(\mu, \sigma_0^2)$ family is a location family:

$$f_X(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x-\mu)^2/2\sigma_0^2} I(x \in \mathbb{R}) = f_Z(x - \mu),$$

where $f_Z(z)$ is the $\mathcal{N}(0, \sigma_0^2)$ PDF. Let $W_i = X_i + c$ for $i = 1, \ldots, n$. Clearly, $\bar{W} = \bar{X} + c$ and

$$S(\boldsymbol{W}) = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = S(\boldsymbol{X}).$$

So, $S(\boldsymbol{X}) = S^2$ is location invariant and hence is ancillary.

Therefore, by Basu's Theorem, $\bar{X}$ and S^2 are independent in the $\mathcal{N}(\mu, \sigma_0^2)$ family. Since we fixed $\sigma^2 = \sigma_0^2$ arbitrarily, this same argument holds for all σ_0^2 fixed.

So, this independence result holds for all choices of σ^2 and hence for the full $\mathcal{N}(\mu, \sigma^2)$ family.

Complete and minimal sufficient statistics

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

The converse is not true.

Recall that for an iid sample $X_1, \ldots, X_n$ from $\mathrm{Unif}(\theta, \theta + 1)$, $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. However, $\boldsymbol{T}$ cannot be complete because $\boldsymbol{T}$ and the sample range $X_{(n)} - X_{(1)}$ are not independent, where the latter is location invariant and hence ancillary in this model.