Biostatistics 602 - Statistical Inference
Lecture 06
Basu's Theorem

Hyun Min Kang

January 29th, 2013

Last Lecture

1. What is a complete statistic?
2. Why it is called as "complete statistic"?
3. Can the same statistic be both complete and incomplete statistics, depending on the parameter space?
4. What is the relationship between complete and sufficient statistics?
5. Is a minimal sufficient statistic always complete?

Complete Statistics

Definition

- Let $\mathcal{T} = \{f_T(t|\theta), \theta \in \Omega\}$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$.
- The family of probability distributions is called *complete* if
- $E[g(T)|\theta] = 0$ for all θ implies $\Pr[g(T) = 0|\theta] = 1$ for all θ.
  - In other words, $g(T) = 0$ almost surely.
- Equivalently, $T(\mathbf{X})$ is called a *complete statistic*

Example - Poisson distribution

When parameter space is limited - NOT complete

- Suppose $\mathcal{T} = \left\{f_T : f_T(t|\lambda) = \frac{\lambda^t e^{-\lambda}}{t!}\right\}$ for $t \in \{0, 1, 2, \cdots\}$. Let $\lambda \in \Omega = \{1, 2\}$. This family is NOT complete

With full parameter space - complete

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda), \lambda > 0$.
- $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a complete statistic.

Example from Stigler (1972) Am. Stat.

Problem

Let X is a uniform random sample from $\{1, \cdots, \theta\}$ where $\theta \in \Omega = \mathbb{N}$. Is $T(X) = X$ a complete statistic?

Solution

Consider a function $g(T)$ such that $E[g(T)|\theta] = 0$ for all $\theta \in \mathbb{N}$.
Note that $f_X(x) = \frac{1}{\theta} I(x \in \{1, \cdots, \theta\}) = \frac{1}{\theta} I_{\mathbb{N}_\theta}(x)$.

$$
\begin{aligned}
E[g(T)|\theta] &= E[g(X)|\theta] = \sum_{x=1}^{\theta} \frac{1}{\theta} g(x) = \frac{1}{\theta} \sum_{x=1}^{\theta} g(x) = 0 \\
\sum_{x=1}^{\theta} g(x) &= 0
\end{aligned}
$$

Solution (cont'd)

for all $\theta \in \mathbb{N}$, which implies

- if $\theta = 1$, $\sum_{x=1}^{\theta} g(x) = g(1) = 0$
- if $\theta = 2$, $\sum_{x=1}^{\theta} g(x) = g(1) + g(2) = g(2) = 0$.
- $\vdots$
- if $\theta = k$, $\sum_{x=1}^{\theta} g(x) = g(1) + \cdots + g(k-1) = g(k) = 0$.

Therefore, $g(x) = 0$ for all $x \in \mathbb{N}$, and $T(X) = X$ is a complete statistic for $\theta \in \Omega = \mathbb{N}$.

Is the previous example barely complete?

Modified Problem

Let X is a uniform random sample from $\{1, \cdots, \theta\}$ where $\theta \in \Omega = \mathbb{N} - \{n\}$. Is $T(X) = X$ a complete statistic?

Solution

Define a nonzero $g(x)$ as follows

$$
\begin{aligned}
g(x) &= \begin{cases} 1 & x = n \\ -1 & x = n+1 \\ 0 & \text{otherwise} \end{cases} \\
E[g(T)|\theta] &= \frac{1}{\theta} \sum_{x=1}^{\theta} g(x) = \begin{cases} 0 & \theta \neq n \\ \frac{1}{\theta} & \theta = n \end{cases}
\end{aligned}
$$

Because Ω does not include n, $g(x) = 0$ for all $\theta \in \Omega = \mathbb{N} - \{n\}$, and $T(X) = X$ is not a complete statistic.

Last Lecture : Ancillary and Complete Statistics

Problem

- Let $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(\theta, \theta+1)$, $\theta \in \mathbb{R}$.
- Is $\mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(n)})$ a complete statistic?

A Simple Proof

- We know that $R = X_{(n)} - X_{(1)}$ is an ancillary statistic, which do not depend on θ.
- Define $g(\mathbf{T}) = X_{(n)} - X_{(1)} - E(R)$. Note that $E(R)$ is constant to θ.
- Then $E[g(\mathbf{T})|\theta] = E(R) - E(R) = 0$, so T is not a complete statistic.

Useful Fact 1 : Ancillary and Complete Statistics

Fact

For a statistic $T(\mathbf{X})$, If a non-constant function of T, say $r(T)$ is ancillary, then $T(\mathbf{X})$ cannot be complete

Proof

Define $g(T) = r(T) - E[r(T)]$, which does not depend on the parameter θ because $r(T)$ is ancillary. Then $E[g(T)|\theta] = 0$ for a non-zero function $g(T)$, and $T(\mathbf{X})$ is not a complete statistic.

Useful Fact 2 : Arbitrary Function of Complete Statistics

Fact

If $T(\mathbf{X})$ is a complete statistic, then a function of T, say $T^* = r(T)$ is also complete.

Proof

$$E[g(T^*)|\theta] = E[g \circ r(T)|\theta]$$

Assume that $E[g(T^*)|\theta] = 0$ for all θ, then $E[g \circ r(T)|\theta] = 0$ holds for all θ too. Because $T(\mathbf{X})$ is a complete statistic, $\Pr[g \circ r(T) = 0] = 1, \ \forall \theta \in \Omega$. Therefore $\Pr[g(T^*) = 0] = 1$, and T^* is a complete statistic.

Theorem 6.2.28 - Lehman and Schefle (1950)

The textbook version

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

Paraphrased version

Any complete, and sufficient statistic is also a minimal sufficient statistic

The converse is NOT true

A minimal sufficient statistic is not necessarily complete. (Recall the example in the last lecture).

Basu's Theorem

Theorem 6.2.24

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

Proof strategy - for discrete case

Suppose that $S(\mathbf{X})$ is an ancillary statistic. We want to show that

$$\Pr(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = \Pr(S(\mathbf{X}) = s), \ \forall t \in \mathcal{T}$$

Alternatively, we can show that

$$\begin{aligned} \Pr(T(\mathbf{X}) = t | S(\mathbf{X}) = s) &= \Pr(T(\mathbf{X}) = t) \\ \Pr(T(\mathbf{X}) = t \wedge S(\mathbf{X}) = s) &= \Pr(T(\mathbf{X}) = t) \Pr(S(\mathbf{X}) = s) \end{aligned}$$

Proof of Basu's Theorem

- As $S(\mathbf{X})$ is ancillary, by definition, it does not depend on θ.
- As $T(\mathbf{X})$ is sufficient, by definition, $f_{\mathbf{X}}(\mathbf{X}|T(\mathbf{X}))$ is independent of θ.
- Because $S(\mathbf{X})$ is a function of $\mathbf{X}$, $\Pr(S(\mathbf{X})|T(\mathbf{X}))$ is also independent of θ.
- We need to show that $\Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = \Pr(S(\mathbf{X}) = s), \; \forall t \in \mathcal{T}$.

Proof of Basu's Theorem (cont'd)

$$\Pr(S(\mathbf{X}) = s|\theta) = \sum_{t\in\mathcal{T}} \Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t)\Pr(T(\mathbf{X}) = t|\theta) \quad (1)$$

$$\Pr(S(\mathbf{X}) = s|\theta) = \Pr(S(\mathbf{X}) = s)\sum_{t\in\mathcal{T}} \Pr(T(\mathbf{X}) = t|\theta) \quad (2)$$

$$= \sum_{t\in\mathcal{T}} \Pr(S(\mathbf{X}) = s)\Pr(T(\mathbf{X}) = t|\theta) \quad (3)$$

Define $g(t) = \Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - \Pr(S(\mathbf{X}) = s)$. Taking (1)-(3),

$$\sum_{t\in\mathcal{T}} [\Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - \Pr(S(\mathbf{X}) = s)]\Pr(T(\mathbf{X}) = t|\theta) = 0$$

$$\sum_{t\in\mathcal{T}} g(t)\Pr(T(\mathbf{X}) = t|\theta) = E[g(T(\mathbf{X}))|\theta] = 0$$

$T(\mathbf{X})$ is complete, so $g(t) = 0$ almost surely for all possible $t \in \mathcal{T}$. Therefore, $S(\mathbf{X})$ is independent of $T(\mathbf{X})$.

Application of Basu's Theorem

Problem

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, \theta)$.
- Calculate $E\left[\frac{X_{(1)}}{X_{(n)}}\right]$ and $E\left[\frac{X_{(1)}+X_{(2)}}{X_{(n)}}\right]$

A strategy for the solution

- We know that $X_{(n)}$ is sufficient statistic.
- We know that $X_{(n)}$ is complete, too.
- We can easily show that $X_{(1)}/X_{(n)}$ is an ancillary statistic.
- Then we can leverage Basu's Theorem for the calculation.

Showing that $X_{(1)}/X_{(n)}$ is Ancillary

$$f_X(x|\theta) = \frac{1}{\theta}I(0 < x < \theta)$$

Let $y = x/\theta$, then $|dx/dy| = \theta$, and $Y \sim \text{Uniform}(0, 1)$.

$$f_Y(y|\theta) = I(0 < y < 1)$$

$$\frac{X_{(1)}}{X_{(n)}} = \frac{Y_{(1)}}{Y_{(n)}}$$

Because the distribution of $Y_1, \cdots, Y_n$ does not depend on θ, $X_{(1)}/X_{(n)}$ is an ancillary statistic for θ.

Applying Basu's Theorem

- By Basu's Theorem, $X_{(1)}/X_{(n)}$ is independent of $X_{(n)}$.
- If X and Y are independent, $E(XY) = E(X)E(Y)$.

$$\begin{aligned} E[X_{(1)}] &= E\left[\frac{X_{(1)}}{X_{(n)}} X_{(n)}\right] = E\left[\frac{X_{(1)}}{X_{(n)}}\right] E\left[X_{(n)}\right] \\ E\left[\frac{X_{(1)}}{X_{(n)}}\right] &= \frac{E[X_{(1)}]}{E[X_{(n)}]} \\ &= \frac{E[\theta Y_{(1)}]}{E[\theta Y_{(n)}]} \\ &= \frac{E[Y_{(1)}]}{E[Y_{(n)}]} \end{aligned}$$

Obtaining $E[Y_{(1)}]$

$$\begin{aligned} Y &\sim \text{Uniform}(0,1) \\ f_Y(y) &= I(0 < y < 1) \\ F_Y(y) &= yI(0 < y < 1) + I(y \geq 1) \\ f_{Y_{(1)}}(y) &= \frac{n!}{(n-1)!} f_Y(y)\left[1 - F_Y(y)\right]^{n-1} I(0 < y < 1) \\ &= n(1-y)^{n-1} I(0 < y < 1) \\ Y_{(1)} &\sim \text{Beta}(1, n) \\ E[Y_{(1)}] &= \frac{1}{n+1} \end{aligned}$$

Obtaining $E[Y_{(n)}]$

$$\begin{aligned} Y &\sim \text{Uniform}(0,1) \\ f_Y(y) &= I(0 < y < 1) \\ F_Y(y) &= yI(0 < y < 1) + I(y \geq 1) \\ f_{Y_{(n)}}(y) &= \frac{n!}{(n-1)!} f_Y(y)\left[F_Y(y)\right]^{n-1} I(0 < y < 1) \\ &= ny^{n-1} I(0 < y < 1) \\ Y_{(n)} &\sim \text{Beta}(n, 1) \\ E[Y_{(n)}] &= \frac{n}{n+1} \end{aligned}$$

Therefore, $E\left[\frac{X_{(1)}}{X_{(n)}}\right] = \frac{E[Y_{(1)}]}{E[Y_{(n)}]} = \frac{1}{n}$

Obtaining $E[Y_{(2)}]$

$$\begin{aligned} Y &\sim \text{Uniform}(0,1) \\ f_Y(y) &= I(0 < y < 1) \\ F_Y(y) &= yI(0 < y < 1) + I(y \geq 1) \\ f_{Y_{(2)}}(y) &= \frac{n!}{(n-2)!}\left[1 - F_Y(y)\right]^{n-2} f_Y(y)\left[F_Y(y)\right] I(0 < y < 1) \\ &= n(n-1)y(1-y)^{n-2} I(0 < y < 1) \\ Y_{(2)} &\sim \text{Beta}(2, n-1) \\ E[Y_{(2)}] &= \frac{2}{n+1} \end{aligned}$$

Therefore, $E\left[\frac{X_{(1)}+X_{(2)}}{X_{(n)}}\right] = \frac{E[Y_{(1)}+Y_{(2)}]}{E[Y_{(n)}]} = \frac{E[Y_{(1)}]+E[Y_{(2)}]}{E[Y_{(n)}]} = \frac{3}{n}$

Summary

Today

- More on complete statistics
- Basu's Theorem

Next Lecture

- Exponential Family