Lecture 12: Completeness

Ancillary statistics

A statistic $V(X)$ is ancillary iff its distribution does not depend on any unknown quantity. A statistic $V(X)$ is first-order ancillary iff $E[V(X)]$ does not depend on any unknown quantity.

A trivial ancillary statistic is $V(X) \equiv$ a constant.

The following examples show that there exist many nontrivial ancillary statistics (non-constant ancillary statistics).

Example: location-scale families

- If $X_1,...,X_n$ is a random sample from a location family with location parameter $\mu \in \mathscr{R}$, then, for any pair (i,j), $1 \le i,j \le n$, $X_i - X_j$ is ancillary, because $X_i - X_j = (X_i - \mu) - (X_j - \mu)$ and the distribution of $(X_i - \mu, X_j - \mu)$ does not depend on any unknown parameter. Similarly, $X_{(i)} - X_{(j)}$ is ancillary, where $X_{(1)},...,X_{(n)}$ are the order statistics, and the sample variance S^2 is ancillary.
- Note that we do not even need to obtain the form of the distribution of $X_i - X_j$.

- If $X_1, ..., X_n$ is a random sample from a scale family with scale parameter $\sigma > 0$, then by the same argument we can show that, for any pair (i,j), $1 \le i,j \le n$, X_i/X_j and $X_{(i)}/X_{(j)}$ are ancillary.
- If $X_1, ..., X_n$ is a random sample from a location-scale family with parameters $\mu \in \mathscr{R}$ and $\sigma > 0$, then, for any (i,j,k), $1 \le i,j,k \le n$, $(X_i - X_k)/(X_j - X_k)$ and $(X_{(i)} - X_{(k)})/(X_{(j)} - X_{(k)})$ are ancillary.

- If $V(X)$ is a non-trivial ancillary statistic, then $\sigma(V)$ does not contain any information about the unknown population P.
- If $T(X)$ is a statistic and $V(T(X))$ is a non-trivial ancillary statistic, it indicates that the reduced data set by T contains a non-trivial part that does not contain any information about θ and, hence, a further simplification of T may still be needed.
- A sufficient statistic $T(X)$ appears to be most successful in reducing the data if no nonconstant function of $T(X)$ is ancillary or even first-order ancillary, which leads to the following definition.

Definition 2.6 (Completeness)

A statistic $T(X)$ is *complete* (or *boundedly complete*) for $P \in \mathscr{P}$ iff, for any Borel f (or bounded Borel f), $E[f(T)] = 0$ for all $P \in \mathscr{P}$ implies $f = 0$ a.s. $\mathscr{P}$.

Remarks

- A complete statistic is boundedly complete.
- If T is complete (or boundedly complete) and $S = \psi(T)$ for a measurable ψ, then S is complete (or boundedly complete).
- Intuitively, a complete and sufficient statistic should be minimal sufficient (Exercise 48).
- A minimal sufficient statistic is not necessarily complete; for example, the minimal sufficient statistic $(X_{(1)}, X_{(n)})$ in Example 2.13 is not complete (Exercise 47).

Proposition 2.1

If P is in an exponential family of full rank with p.d.f.'s given by

$$f_\eta(x) = \exp\{\eta^\tau T(x) - \zeta(\eta)\} h(x),$$

then $T(X)$ is complete and sufficient for $\eta \in \Xi$.

Proof

We have shown that T is sufficient.

We now show that T is complete.
Suppose that there is a function f such that $E[f(T)] = 0$ for all $\eta \in \Xi$.
By Theorem 2.1(i),

$$\int f(t) \exp\{\eta^\tau t - \zeta(\eta)\} d\lambda = 0 \quad \text{for all } \eta \in \Xi,$$

where $\lambda(A) = \int_A h(x) d\nu$ is a measure on $(\mathscr{R}^p, \mathscr{B}^p)$.
Let η_0 be an interior point of Ξ. Then

$$\int f_+(t) e^{\eta^\tau t} d\lambda = \int f_-(t) e^{\eta^\tau t} d\lambda \quad \text{for all } \eta \in N(\eta_0), \qquad (1)$$

where $N(\eta_0) = \{\eta \in \mathscr{R}^p : \|\eta - \eta_0\| < \varepsilon\}$ for some $\varepsilon > 0$.
In particular,

$$\int f_+(t) e^{\eta_0^\tau t} d\lambda = \int f_-(t) e^{\eta_0^\tau t} d\lambda = c.$$

If $c = 0$, then $f = 0$ a.e. λ.
If $c > 0$, then $c^{-1} f_+(t) e^{\eta_0^\tau t}$ and $c^{-1} f_-(t) e^{\eta_0^\tau t}$ are p.d.f.'s w.r.t. λ and result (1) implies that their m.g.f.'s are the same in a neighborhood of 0.
By Theorem 1.6(ii), $c^{-1} f_+(t) e^{\eta_0^\tau t} = c^{-1} f_-(t) e^{\eta_0^\tau t}$, i.e., $f = f_+ - f_- = 0$ a.e. λ, which implies that $f = 0$ a.s. $\mathscr{P}$.
Hence T is complete.

Example 2.15

Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, $\mu \in \mathscr{R}$, $\sigma > 0$.

From Example 2.6, the joint p.d.f. of $X_1, ..., X_n$ is

$$(2\pi)^{-n/2} \exp\{\eta_1 T_1 + \eta_2 T_2 - n\zeta(\eta)\},$$

where $T_1 = \sum_{i=1}^n X_i$, $T_2 = -\sum_{i=1}^n X_i^2$, and $\eta = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$.

Hence, the family of distributions for $X = (X_1, ..., X_n)$ is a natural exponential family of full rank ($\Xi = \mathscr{R} \times (0, \infty)$).

By Proposition 2.1, $T(X) = (T_1, T_2)$ is complete and sufficient for η.

Since there is a one-to-one correspondence between η and $\theta = (\mu, \sigma^2)$, T is also complete and sufficient for θ.

It can be shown that any one-to-one measurable function of a complete and sufficient statistic is also complete and sufficient (exercise).

Thus, $(\bar{X}, S^2)$ is complete and sufficient for θ, where $\bar{X}$ and S^2 are the sample mean and sample variance, respectively.

Example 2.16

Let $X_1, ..., X_n$ be i.i.d. random variables from P_θ, the uniform distribution $U(0,\theta)$, $\theta > 0$.
The largest order statistic, $X_{(n)}$, is complete and sufficient for $\theta \in (0,\infty)$.
The sufficiency of $X_{(n)}$ follows from the fact that the joint Lebesgue p.d.f. of $X_1, ..., X_n$ is $\theta^{-n} I_{(0,\theta)}(x_{(n)})$.
From Example 2.9, $X_{(n)}$ has the Lebesgue p.d.f. $(nx^{n-1}/\theta^n) I_{(0,\theta)}(x)$.
Let f be a Borel function on $[0,\infty)$ such that $E[f(X_{(n)})] = 0$ for all $\theta > 0$.
Then

$$\int_0^\theta f(x) x^{n-1} dx = 0 \quad \text{for all } \theta > 0.$$

Let $G(\theta)$ be the left-hand side of the previous equation.
Applying the result of differentiation of an integral (see, e.g., Royden (1968, §5.3)), we obtain that $G'(\theta) = f(\theta)\theta^{n-1}$ a.e. m_+, where m_+ is the Lebesgue measure on $([0,\infty), \mathscr{B}_{[0,\infty)})$.
Since $G(\theta) = 0$ for all $\theta > 0$, $f(\theta)\theta^{n-1} = 0$ a.e. m_+ and, hence, $f(x) = 0$ a.e. m_+.
Therefore, $X_{(n)}$ is complete and sufficient for $\theta \in (0,\infty)$.

Example 2.17

In Example 2.12, we showed that the order statistics $T(X) = (X_{(1)}, ..., X_{(n)})$ of i.i.d. random variables $X_1, ..., X_n$ is sufficient for $P \in \mathscr{P}$, where $\mathscr{P}$ is the family of distributions on $\mathscr{R}$ having Lebesgue p.d.f.'s.

We now show that $T(X)$ is also complete for $P \in \mathscr{P}$.

Let $\mathscr{P}_0$ be the family of Lebesgue p.d.f.'s of the form

$$f(x) = C(\theta_1, ..., \theta_n) \exp\{-x^{2n} + \theta_1 x + \theta_2 x^2 + \cdots + \theta_n x^n\},$$

where $\theta_j \in \mathscr{R}$ and $C(\theta_1, ..., \theta_n)$ is a normalizing constant such that $\int f(x)dx = 1$.

Then $\mathscr{P}_0 \subset \mathscr{P}$ and $\mathscr{P}_0$ is an exponential family of full rank.

Note that the joint distribution of $X = (X_1, ..., X_n)$ is also in an exponential family of full rank.

Thus, by Proposition 2.1, $U = (U_1, ..., U_n)$ is a complete statistic for $P \in \mathscr{P}_0$, where $U_j = \sum_{i=1}^n X_i^j$.

Since a.s. $\mathscr{P}_0$ implies a.s. $\mathscr{P}$, $U(X)$ is also complete for $P \in \mathscr{P}$.

Example 2.17 (continued)

The result follows if we can show that there is a one-to-one correspondence between $T(X)$ and $U(X)$.

Let $V_1 = \sum_{i=1}^n X_i$, $V_2 = \sum_{i<j} X_i X_j$, $V_3 = \sum_{i<j<k} X_i X_j X_k$,..., $V_n = X_1 \cdots X_n$.

From the identities

$$U_k - V_1 U_{k-1} + V_2 U_{k-2} - \cdots + (-1)^{k-1} V_{k-1} U_1 + (-1)^k k V_k = 0,$$

$k = 1, ..., n$, there is a one-to-one correspondence between $U(X)$ and $V(X) = (V_1, ..., V_n)$.

From the identity

$$(t - X_1) \cdots (t - X_n) = t^n - V_1 t^{n-1} + V_2 t^{n-2} - \cdots + (-1)^n V_n,$$

there is a one-to-one correspondence between $V(X)$ and $T(X)$.

This completes the proof and, hence, $T(X)$ is sufficient and complete for $P \in \mathscr{P}$.

In fact, both $U(X)$ and $V(X)$ are sufficient and complete for $P \in \mathscr{P}$.

The relationship between an ancillary statistic and a complete and sufficient statistic is characterized in the following result.

Theorem 2.4 (Basu's theorem)

Let V and T be two statistics of X from a population $P \in \mathscr{P}$.
If V is ancillary and T is boundedly complete and sufficient for $P \in \mathscr{P}$, then V and T are independent w.r.t. any $P \in \mathscr{P}$.

Proof

Let B be an event on the range of V.
Since V is ancillary, $P(V^{-1}(B))$ is a constant.
As T is sufficient, $E[I_B(V)|T]$ is a function of T (not dependent on P).
Because

$$E\{E[I_B(V)|T] - P(V^{-1}(B))\} = 0 \quad \text{for all } P \in \mathscr{P},$$

by the bounded completeness of T,

$$P(V^{-1}(B)|T) = E[I_B(V)|T] = P(V^{-1}(B)) \quad \text{a.s. } \mathscr{P}$$

For A being an event on the range of T,

$$\begin{aligned} P(T^{-1}(A) \cap V^{-1}(B)) &= E\{E[I_A(T)I_B(V)|T]\} = E\{I_A(T)E[I_B(V)|T]\} \\ &= E\{I_A(T)P(V^{-1}(B))\} = P(T^{-1}(A))P(V^{-1}(B)). \end{aligned}$$

Hence T and V are independent w.r.t. any $P \in \mathscr{P}$.

Basu's theorem is useful in proving the independence of two statistics.

Example 2.18

Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, with $\mu \in \mathscr{R}$ and a known $\sigma > 0$.

It can be easily shown that the family $\{N(\mu, \sigma^2) : \mu \in \mathscr{R}\}$ is an exponential family of full rank with natural parameter $\eta = \mu/\sigma^2$.

By Proposition 2.1, the sample mean $\bar{X}$ is complete and sufficient for η (and μ).

Let $\bar{X}$ be the sample mean and S^2 be the sample variance.

Since $S^2 = (n-1)^{-1}\sum_{i=1}^n (Z_i - \bar{Z})^2$, where $Z_i = X_i - \mu$ is $N(0, \sigma^2)$ and $\bar{Z} = n^{-1}\sum_{i=1}^n Z_i$, S^2 is an ancillary statistic (σ^2 is known).

By Basu's theorem, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ with $\mu \in \mathscr{R}$.

Since σ^2 is arbitrary, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ for any $\mu \in \mathscr{R}$ and $\sigma^2 > 0$.

If a minimal sufficient statistic T is not complete, then there may be an ancillary statistic V such that V and T are not independent.

Example 2.13

In this example, $X_1,...,X_n$ is a random sample from *uniform*$(\theta,\theta+1)$, $\theta \in \mathscr{R}$, and $T=(X_{(1)},X_{(n)})$ is the minimal sufficient statistic for θ.

We now show that T is not complete.

Note that $V(T)=X_{(n)}-X_{(1)}=(X_{(n)}-\theta)-(X_{(1)}-\theta)$ is in fact ancillary.

It is easy to see that $E_\theta(V)$ exists and it does not depend on θ since V is ancillary.

Letting $c=E(V)$, we see that $E_\theta(V-c)=0$ for all θ.

Thus, we have a function $g(x,y)=x-y-c$ such that
$E_\theta[g(X_{(1)},X_{(n)})]=E_\theta(V-c)=0$ for all θ but
$P_\theta(g(X_{(1)},X_{(n)})=0)=P_\theta(V=c)\neq 0$.

This shows that T is not complete.

In this case, $\sigma(V)\subset\sigma(T)$ and $\sigma(V)$ contains no information about θ.

The relationship between minimal sufficiency and sufficiency with completeness is given by the following theorem.

Theorem

Suppose that S is a minimal sufficient statistic and T is a complete and sufficient statistic. Then T must be minimal sufficient and S must be complete.

Proof.

Since S is minimal sufficient and T is sufficient, there exists a Borel function h such that $S = h(T)$ a.s. $\mathscr{P}$.

Since h cannot be a constant function and T is complete, we conclude that S is complete.

Consider $T - E(T|S) = T - E[T|h(T)]$, which is a Borel function of T and hence can be denoted as $g(T)$.

Note that $E[g(T)] = 0$.

By the completeness of T, $g(T) = 0$ a.s. $\mathscr{P}$, i.e., $T = E(T|S)$ a.s. $\mathscr{P}$

This means that T is also a function of S and, therefore, T is minimal sufficient.

Example (ancillary precision)

Let X_1 and X_2 be iid from the discrete uniform distribution on three points $\{\theta, \theta+1, \theta+2\}$, where $\theta \in \Theta = \{0, \pm 1, \pm 2, ...\}$.

Using the same argument as in Example 2.13, we can show that the order statistics $(X_{(1)}, X_{(2)})$ is minimal sufficient for θ.

Let $M = (X_{(1)} + X_{(2)})/2$ and $R = X_{(2)} - X_{(1)}$ (mid-range and range).

Since (M, R) is a one-to-one function of $(X_{(1)}, X_{(2)})$, it is also minimal sufficient for θ.

Consider the estimation of θ using (M, R).

Note that $R = (X_{(2)} - \theta) - (X_{(1)} - \theta)$ is the range of the two order statistics from the uniform distribution on $\{0, 1, 2\}$ and, hence the distribution of R does not depend on θ, i.e., R is ancillary.

One may think R is useless in the estimation of θ and only M is useful.

Suppose we observe $(M, R) = (m, r)$ and m is an integer.

From the observation m, we know that θ can only be one of the 3 values m, $m-1$, and $m-2$; however, we are not certain which of the 3 values is θ.

We can know more if $r = 2$, which must be the case that $X_{(1)} = m-1$ and $X_{(2)} = m+1$.

With this additional information, the only possible value for θ is $m-1$.

When m is an integer, r cannot be 1. If $r = 0$, then we know that $X_1 = X_2$ and we are not certain which of the 3 values is θ.

The knowledge of the value of the ancillary statistic R increases our knowledge about θ, although R alone gives us no information about θ.

What we learn from the previous example?

- An ancillary statistic that is a function of a minimal sufficient statistic T may still be useful for our knowledge about θ. (Note that the ancillary statistic is still a function of T.)
- This cannot occur to a sufficient and complete statistic T, since, if $V(T)$ is ancillary, then by the completeness of T, V must be a constant and is useless.
- Therefore, the sufficiency and completeness together is a much desirable (and strong) property.