General comment:

An ancillary statistic *by itself* can tell us nothing about $\theta$, but when combined with other statistics it may give information about $\theta$.

Example: $X = (X_1, \ldots, X_n)$ iid Unif$(\theta, \theta+1)$.

We know $(X_{(1)}, X_{(n)})$ is MSS.

Any 1-1 function of a MSS is also MSS.

Therefore $(X_{(1)}, X_{(n)} - X_{(1)})$ is MSS.

We can**not** drop $X_{(n)} - X_{(1)}$ without losing info about $\theta$.

But $X_{(n)} - X_{(1)}$ is ancillary!

It is ancillary because

Uniform$(\theta, \theta+1)$ is a location family, and $X_{(n)} - X_{(1)}$ is a location invariant statistic.

# Complete Statistics

Suppose $X \sim P_\theta$, $\theta \in \Theta$.

**Definition:**

A statistic $T = T(X)$ is complete if

$$E_\theta g(T) = 0 \text{ for all } \theta$$

implies

$$P_\theta(g(T)=0) = 1 \text{ for all } \theta.$$

(Note: $E_\theta$ denotes expectation computed with respect to $P_\theta$.)

**Example:** $X = (X_1, \ldots, X_n)$ iid $N(\theta, 1)$.

$T(X) = (X_1, X_2)$ is a statistic which is not complete because

$$E(\underbrace{X_1 - X_2}_{\substack{\text{function}\\ \text{of } T}}) = 0 \text{ for all } \theta$$

but $P(X_1 - X_2 = 0) \neq 1$ for all $\theta$.

More formally: $T$ is not complete because the function $g(u) = u_1 - u_2$ (where $u = (u_1, u_2) \in \mathbb{R}^2$) satisfies

$$E\,g(T) = E(X_1 - X_2) = 0 \quad \text{for all } \theta$$

but $P(g(T) = 0) \neq 1$ for all $\theta$.

<u>Example</u>: $X = (X_1, \ldots, X_n)$ iid Uniform$(\theta, \theta+1)$.

$T = T(X) = (\min X_i, \max X_i)$ is a MSS.

But $T$ is <u>not</u> complete.

We know $S(X) = \max X_i - \min X_i$ is ancillary. Thus

$$E(\max X_i - \min X_i) = \underset{\substack{\uparrow \\ \text{does not depend} \\ \text{on } \theta}}{c}$$

and therefore

$$E(\underbrace{\max X_i - \min X_i - c}_{g(T)}) = 0 \quad \text{for all } \theta$$

but clearly

$$P(\max X_i - \min X_i - c = 0) \neq 1 \quad \text{for all } \theta.$$

Example: $X=(X_1,\ldots,X_n)$ iid Unif$(0,\theta)$.

$T=T(X)=\max X_i = X_{(n)}$ is MSS.

$T$ is also complete.

Proof: Assume $\exists\, g$ such that

$Eg(T)=0$ for all $\theta>0$.

$T$ has cdf $H(t)=\left(\frac{t}{\theta}\right)^n,\ 0\le t\le\theta$

pdf $h(t)=\dfrac{nt^{n-1}}{\theta^n},\ 0\le t\le\theta.$

$$Eg(T)=\int_0^\theta g(t)\,\frac{nt^{n-1}}{\theta^n}\,dt=0$$

for all $\theta>0$

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

implies (by differentiating both sides and using the Fund. Thm. of Calculus)

$$g(\theta)\,n\theta^{n-1}=0\quad\forall\theta>0$$

implies $g(t)=0\ \ \forall t>0$

implies $P(g(T)=0)=1\ \ \forall\theta>0.$

Theorem:

Suppose $X_1, \ldots, X_n$ iid with pdf (pmf)

$$f(x|\theta) = c(\theta)h(x)\exp\left\{\sum_{j=1}^{k} w_j(\theta)t_j(x)\right\}$$

for $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

Let $X = (X_1, \ldots, X_n)$. Define

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

Then

(a) $T(X)$ is sufficient statistic for $\theta$.

(b)* If $\Theta$ contains an open set in $\mathbb{R}^k$, then $T(X)$ is complete.

* More precisely, if

$$\{(w_1(\theta), w_2(\theta), \ldots, w_k(\theta)) : \theta \in \Theta\}$$

contains an open set in $\mathbb{R}^k$, then $T(X)$ is complete.

**Remarks:**

The statistic $T(X)$ in the Theorem is called the <u>natural</u> sufficient statistic.

$\eta = (\eta_1, \ldots, \eta_k) \equiv (w_1(\theta), \ldots, w_k(\theta))$ is called the <u>natural</u> parameter of the exponential family.

Condition (b) is the "open set condition" (OSC).

The OSC is easily verified by inspection.

> Let $A \subset \mathbb{R}^k$.
>
> $A$ contains an open set in $\mathbb{R}^k$ iff $A$ contains a $k$-dimensional ball. That is, $\exists\, x \in \mathbb{R}^k$ and $r > 0$ such that $B(x, r) \subset A$. Here $B(x, r)$ denotes the ball of radius $r$ about $x$.
>
> Let $A \subset \mathbb{R}$ (take $k = 1$).
>
> $A$ contains an open set in $\mathbb{R}$ iff $A$ contains an interval. That is, $\exists\, c < d$ such that $(c, d) \subset A$.

**Facts:**

1. Under weak conditions which are almost always true a complete sufficient statistic is also minimal.

   Abbreviation: CSS $\Rightarrow$ MSS.

   (But MSS $\not\Rightarrow$ CSS as we saw earlier.)

2. A one-to-one function of a CSS is also a CSS. (See later remarks.)

   (Reminder: A 1-1 function of an MSS is also an MSS.)

**Example:**

The $N(\theta, 1)$ family is a 1pef with $w(\theta) = \theta$, $t(x) = x$.

Let $X = (X_1, \ldots, X_n)$ iid $N(\theta, 1)$.

$T(X) = \sum_{i=1}^n X_i$ is the natural SS. (It is a SS for any $\Theta$.)

Is $T$ complete? That depends on $\Theta$.

- $\Theta = \mathbb{R}$: Yes. (OSC holds)
- $\Theta = [.01, .02]$: Yes. (OSC holds)
- $\Theta = (1, 2) \cup \{4, 7\}$: Yes. (OSC holds)
- $\Theta = \mathbb{Z}$ (the integers): OSC fails so Theorem says nothing. But can show it is **not** complete.
- $\Theta = \{1, 1/2, 1/3, 1/4, \ldots\}$: OSC fails so Theorem says nothing. Yes or no? Don't know.
- $\Theta =$ Cantor Set: Ditto, but would bet money it is complete.
- $\Theta =$ finite set: OSC fails so Theorem says nothing. But can show it is **not** complete.

Remark: In general, it is typically true that if $\Theta$ is finite and the support of $T = T(X)$ is infinite, then $T$ is **not** complete.

**Example:**

The $N(\mu, \sigma^2)$ family with $\theta = (\mu, \sigma^2)$ is a 2pef with

$$w(\theta) = \left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right),\ t(x) = (x, x^2).$$

Let $X = (X_1, \ldots, X_n)$ iid $N(\theta, 1)$.

$T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is the natural SS. (It is a SS for any $\Theta$.)

$T(X)$ is a one-to-one function of $U(X) = (\bar{x}, s^2)$.

So $T$ is CSS iff $U$ is CSS.

Is $T$ (or $U$) complete? That depends on $\Theta$.

- $\Theta_1 = \{(\mu, \sigma^2) : \sigma^2 > 0\}$. OSC holds. Yes, complete.

- $\Theta_2 = \{(\mu, \sigma^2) : \sigma^2 = \sigma_0^2\}$. OSC fails. Thm says nothing. No, not complete.
  Proof: $Eg(U) = E(s^2 - \sigma_0^2) = \sigma^2 - \sigma_0^2 = 0$ for all $\theta \in \Theta_2$.

- $\Theta_3 = \{(\mu, \sigma^2) : \mu = \mu_0,\ \sigma^2 > 0\}$. Ditto.
  Proof: $Eg(U) = E(\bar{x} - \mu_0) = \mu - \mu_0 = 0$ for all $\theta \in \Theta_3$.

- $\Theta_4 = \{(\mu, \sigma^2) : \mu = \sigma^2, \sigma^2 > 0\}$. Ditto.
  Proof: $Eg(U) = E(\bar{x} - s^2) = \mu - \sigma^2 = 0$ for all $\theta \in \Theta_4$.

(**Note:** It is more natural to describe the families $\Theta_2$, $\Theta_3$, $\Theta_4$ as 1pef's. If you do this, you get **different** natural sufficient statistics, which turn out to be complete.)

- $\Theta_5 = \{(\mu, \sigma^2) : \mu^2 = \sigma^2, \sigma^2 > 0\}$. Ditto.

  Proof: homework.

- $\Theta_6 = [1,3] \times [4,6]$. OSC holds. Yes, complete.

- $\Theta_7 = \Theta_6 \cup \{(5,1),(4,2)\}$. OSC holds. Yes, complete.

- $\Theta_8$ = complicated wavy curve. OSC fails. Thm says nothing. (Probably complete, but hard to say.)

Corollary:

Suppose $X \in \mathbb{R}^m$ has joint pdf (pmf)

$$f(x|\theta) = c(\theta)h(x)\exp\left\{\sum_{j=1}^{k} w_j(\theta)t_j(x)\right\}$$

for all $x \in \mathbb{R}^m$

where $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

Define

$T(X) = (t_1(X), t_2(X), \ldots, t_k(X))$.

Then

(a) $T(X)$ is sufficient stat. for $\theta$.

(b)* If $\Theta$ contains an open set in $\mathbb{R}^k$, then $T(X)$ is complete.

* More precisely, ...

## Notation:

$X \sim P_\theta$, $\theta \in \Theta$.

$S(X) = \psi(T(X))$ for some $\psi$.

$\Theta_1 \subset \Theta_2 \subset \Theta$.

## Sufficiency

① If $S(X)$ is sufficient, then $T(X)$ is suff.

② If $T(X)$ is sufficient for $\Theta_2$,
then $T(X)$ is sufficient for $\Theta_1$.

## Completeness

① If $T(X)$ is complete, then $S(X)$ is complete.

② If $T(X)$ is complete for $\Theta_1$,
then $T(X)$ is complete for $\Theta_2$.
(under mild regularity conditions)

## Ancillarity

① If $T(X)$ is ancillary, then $S(X)$ is ancillary.

② If $T(X)$ is ancillary for $\Theta_2$,
then $T(X)$ is ancillary for $\Theta_1$.

## Completeness

Proof of ①:

$$E_\theta\, g(S(X)) = 0 \quad \text{for all } \theta \in \Theta$$

$$\Rightarrow E_\theta\, g(\psi(T(X))) = 0 \quad \text{for all } \theta \in \Theta$$

$$\Rightarrow P_\theta\{g(\psi(T(X))) = 0\} = 1 \quad \forall \theta$$

(by completeness of $T(X)$)

$$\Rightarrow P_\theta\{g(S(X)) = 0\} = 1 \quad \forall \theta.$$

Proof of ②:

$$E_\theta\, g(T(X)) = 0 \quad \text{for all } \theta \in \Theta_2$$

$$\Rightarrow E_\theta\, g(T(X)) = 0 \quad \text{for all } \theta \in \Theta_1$$

$$\Rightarrow P_\theta(g(T(X)) = 0) \quad \text{for all } \theta \in \Theta_1$$

(since $T(X)$ is complete for $\Theta_1$)

$$\Rightarrow P_\theta(g(T(X)) = 0) \quad \text{for all } \theta \in \Theta_2$$

(under mild assumptions)

## Ancillarity

① uses: $Y \stackrel{d}{=} Z \Rightarrow \psi(Y) \stackrel{d}{=} \psi(Z)$.

② obvious.