Lecture 03: Sufficient, Ancillary, and Complete Statistics

Mathematical Statistics II, MATH 60062/70062

Thursday January 27, 2022

Reference: Casella & Berger, 6.2.3-6.2.4

Recap

Suppose $X_1, \ldots, X_n$ is an iid sample from $f_X(x \mid \theta)$, where $\theta \in \Theta$.

- A **statistic**, $T = T(\boldsymbol{X}) = T(X_1, \ldots, X_n)$, is a function of the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. T cannot depend on θ.
- A statistic $T = T(\boldsymbol{X})$ is a **sufficient statistic** for θ if the conditional distribution of $\boldsymbol{X}$ given T does not depend on θ; i.e., the ratio

  $$f_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)}{f_T(t \mid \theta)}$$

  is free of θ, for all $\boldsymbol{x} \in \mathcal{X}$.
- A statistic $T = T(\boldsymbol{X})$ is a **minimal sufficient statistic** for θ if, for any other sufficient statistic $T^*(\boldsymbol{X})$, $T(\boldsymbol{x})$ is a function of $T^*(\boldsymbol{x})$.

Ancillary statistics

A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called an **ancillary statistic**.

A sufficient statistic $T(\boldsymbol{X})$ contain *all* the information about θ and the distribution of an ancillary statistic $S(\boldsymbol{X})$ is free of θ.

- Are $T(\boldsymbol{X})$ and $S(\boldsymbol{X})$ independent?
- Can $S(\boldsymbol{X})$ be useful for inferences about θ?

Normal ancillary statistic

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 > 0$.

- The sample mean $\bar{X} \sim \mathcal{N}(0, \sigma^2/n)$ is *not* ancillary, as its distribution depends on σ^2.
- The statistic
$$S(\boldsymbol{X}) = \frac{\bar{X}}{S/\sqrt{n}} \sim t_{n-1}$$
is ancillary, because its distribution, t_{n-1}, does not depend on σ^2.

Location-invariant statistic

A statistic $S(\boldsymbol{X})$ is called a **location-invariant statistic** if for any $c \in \mathbb{R}$,

$$S(x_1 + c, \ldots, x_n + c) = S(x_1, \ldots, x_n)$$

for all $\boldsymbol{x} \in \mathcal{X}$.

Each of the following is a location-invariant statistic:

- $S(\boldsymbol{X}) = X_{(n)} - X_{(1)}$
- $S(\boldsymbol{X}) = \sum_{i=1}^{n} |X_i - \bar{X}|/n$
- $S(\boldsymbol{X}) = S^2$

Ancillary statistic for location family

Suppose $X_1, \ldots, X_n$ are iid from a **location family** with standard PDF f_Z and location parameter $-\infty < \mu < \infty$,

$$f_X(x \mid \mu) = f_Z(x - \mu).$$

If $S(\boldsymbol{X})$ is **location invariant**, then it is **ancillary**.

Ancillary statistic for location family

Suppose $X_1, \ldots, X_n$ are iid from a **location family** with standard PDF f_Z and location parameter $-\infty < \mu < \infty$,

$$f_X(x \mid \mu) = f_Z(x - \mu).$$

If $S(\boldsymbol{X})$ is **location invariant**, then it is **ancillary**.

Let $W_i = X_i - \mu$, for $i = 1, \ldots, n$. The distribution of $\boldsymbol{W} = (W_1, \ldots, W_n)$ is given by

$$\begin{aligned} f_{\boldsymbol{W}}(\boldsymbol{w}) &= f_{\boldsymbol{X}}(w_1 + \mu, \ldots, w_n + \mu) \\ &= \prod_{i=1}^{n} f_X(w_i + \mu) \\ &= \prod_{i=1}^{n} f_Z(w_i + \mu - \mu) = \prod_{i=1}^{n} f_Z(w_i), \end{aligned}$$

which does depends on μ.

Because $S(\boldsymbol{X})$ is location invariant,

$$\begin{aligned} S(\boldsymbol{X}) &= S(X_1, \ldots, X_n) \\ &= S(W_1 + \mu, \ldots, W_n + \mu) \\ &= S(W_1, \ldots, W_n) \\ &= S(\boldsymbol{W}). \end{aligned}$$

The distribution of $\boldsymbol{W}$ does not depend on μ, so $S(\boldsymbol{X}) = S(\boldsymbol{W})$ does not depend on μ either. Therefore, $S(\boldsymbol{X})$ is ancillary.

Scale-invariant and ancillary statistic

A statistic $S(\boldsymbol{X})$ is called a **scale-invariant statistic** if for any $c > 0$,

$$S(cx_1, \ldots, cx_n) = S(x_1, \ldots, x_n)$$

for all $\boldsymbol{x} \in \mathcal{X}$.

Each of the following is a scale-invariant statistic:

- $S(\boldsymbol{X}) = X_{(n)}/X_{(1)}$
- $S(\boldsymbol{X}) = S/\bar{X}$

Suppose $X_1, \ldots, X_n$ are iid from a **scale family** with standard PDF f_Z and scale parameter $\sigma > 0$,

$$f_X(x \mid \sigma) = \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right).$$

If $S(\boldsymbol{X})$ is **scale invariant**, then it is **ancillary**.

Independence between sufficient and ancillary statistics?

A sufficient statistic and an ancillary statistic are *not* necessarily independent.

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(\theta, \theta + 1)$, where $-\infty < \theta < \infty$.

Independence between sufficient and ancillary statistics?

A sufficient statistic and an ancillary statistic are *not* necessarily independent.

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(\theta, \theta + 1)$, where $-\infty < \theta < \infty$.

- From Lecture 2, we know $(X_{(n)} - X_{(1)}, (X_{(1)} + X_{(n)})/2)$ is a **minimal sufficient statistic**.

Independence between sufficient and ancillary statistics?

A sufficient statistic and an ancillary statistic are *not* necessarily independent.

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(\theta, \theta+1)$, where $-\infty < \theta < \infty$.

- From Lecture 2, we know $(X_{(n)} - X_{(1)}, (X_{(1)} + X_{(n)})/2)$ is a **minimal sufficient statistic**.
- $\mathrm{Unif}(\theta, \theta+1)$ is a **location family**, and $S(\boldsymbol{X}) = X_{(n)} - X_{(1)}$ is location-invariant. Therefore, $S(\boldsymbol{X})$ is an **ancillary statistic**.
- In this case, the ancillary statistic is an important component of the minimal sufficient statistic.

Can ancillary statistics be useful for inferences?

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We are interested in inference on μ.

- The sample variance S^2 is ancillary for μ, because

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$$

  does not depend on μ.
- We know $\bar{X}$ and S^2 are independent and $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.
- The statistic

$$T(\boldsymbol{X}) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

  is used for inference on μ, where the ancillary statistic S^2 plays an essential role.

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 $\theta \in (0,1)$.

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.

Ancillary, complete and sufficient statistics

- **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.
- If a minimal sufficient statistic exists, then any **complete statistic** is also a **minimal sufficient statistic**.
- The converse is not true - a minimal sufficient statistic is not necessarily complete.
  E.g., 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 an ancillary 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 parameter space Θ 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.

An easy way to show the independence between $\bar{X}$ and S^2 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.