Lecture 20: Minimal Sufficient and Ancillary Statistics

Mathematical Statistics I, MATH 60061/70061

Thursday December 2, 2021

Reference: Casella & Berger, 6.2.2-6.2.3

There are many sufficient statistics in any problem

The complete sample, $\boldsymbol{X}$, is a sufficient statistic, since

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}) \mid \theta)h(\boldsymbol{x}),$$

where $T(\boldsymbol{x}) = \boldsymbol{x}$, $g(\boldsymbol{x} \mid \theta) = f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)$, and $h(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$.

Any one-to-one function of a sufficient statistic is a sufficient statistic. Suppose $T = T(\boldsymbol{X})$ is sufficient, and define $T^*(\boldsymbol{X}) = r(T(\boldsymbol{X}))$, where r is a one-to-one function with inverse r^{-1}. Then

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) &= g(T(\boldsymbol{x}) \mid \theta)h(\boldsymbol{x}) \\ &= g(r^{-1}(T^*(\boldsymbol{x})) \mid \theta)h(\boldsymbol{x}) \\ &= g^{-1}(T^*(\boldsymbol{x}) \mid \theta)h(\boldsymbol{x}), \end{aligned}$$

where g^{-1} is the composition of g and r^{-1}.

Normal sufficient statistic

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $0 < \mu < 1$ and σ_0^2 is known. Each of the following statistics is sufficient:

- $T_1(\boldsymbol{X}) = \bar{X}$
- $T_2(\boldsymbol{X}) = (X_1, \sum_{i=2}^{n} X_i)$
- $T_3(\boldsymbol{X}) = (X_{(1)}, \ldots, X_{(n)})$
- $T_4(\boldsymbol{X}) = \boldsymbol{X}$

How much data reduction is possible?

Minimal sufficient statistics

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

This means that if you know $T^*(\boldsymbol{x})$, you can calculate $T(\boldsymbol{x})$, and

$$T^*(\boldsymbol{x}) = T^*(\boldsymbol{y}) \implies T(\boldsymbol{x}) = T(\boldsymbol{y}).$$

A minimal sufficient statistic achieves the *greatest possible data reduction*. In terms of partition sets formed by statistics, a minimal sufficient statistic admits the coarsest possible partition.

Minimal sufficient statistic

Using the definition to find a minimal sufficient statistic is impractical. The following result by Lehmann and Scheffé gives an easier way to find a minimal sufficient statistic.

Suppose $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)$, where $\theta \in \Theta$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)}{f_{\boldsymbol{X}}(\boldsymbol{y} \mid \theta)} \text{ is free of } \theta \iff T(\boldsymbol{x}) = T(\boldsymbol{y}).$$

Then $T(\boldsymbol{X})$ is a minimal sufficient statistic.

Normal minimal sufficient statistic

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $0 < \mu < 1$ and σ_0^2 is known. The PDF of $\boldsymbol{X}$ is

$$
\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_i-\mu)^2/2\sigma_0^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \right)^n e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma_0^2}.
\end{aligned}
$$

where

$$
\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2.
$$

The ratio

$$
\frac{f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)}{f_{\boldsymbol{X}}(\boldsymbol{y} \mid \theta)} = \frac{\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)/2\sigma_0^2\right]}{\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\left(\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right)/2\sigma_0^2\right]},
$$

is free of μ if and only if $\bar{x} = \bar{y}$. Therefore, $T(\boldsymbol{X}) = \bar{X}$ is a **minimal sufficient statistic**.

Uniform minimal sufficient statistic

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(\theta, \theta+1)$, where $-\infty < \theta < \infty$. The PDF of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) = I(x_{(1)} > \theta) I(x_{(n)} < \theta + 1) \prod_{i=1}^{n} I(x_i \in \mathbb{R}).$$

The ratio

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)}{f_{\boldsymbol{X}}(\boldsymbol{y} \mid \theta)} = \frac{I(x_{(1)} > \theta) I(x_{(n)} < \theta + 1) \prod_{i=1}^{n} I(x_i \in \mathbb{R})}{I(y_{(1)} > \theta) I(y_{(n)} < \theta + 1) \prod_{i=1}^{n} I(y_i \in \mathbb{R})},$$

is free of θ if and only if $(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$. Therefore, $\boldsymbol{T}(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic.

Uniform minimal sufficient statistic

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(\theta, \theta+1)$, where $-\infty < \theta < \infty$. The PDF of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) = I(x_{(1)} > \theta) I(x_{(n)} < \theta + 1) \prod_{i=1}^{n} I(x_i \in \mathbb{R}).$$

The ratio

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)}{f_{\boldsymbol{X}}(\boldsymbol{y} \mid \theta)} = \frac{I(x_{(1)} > \theta) I(x_{(n)} < \theta + 1) \prod_{i=1}^{n} I(x_i \in \mathbb{R})}{I(y_{(1)} > \theta) I(y_{(n)} < \theta + 1) \prod_{i=1}^{n} I(y_i \in \mathbb{R})},$$

is free of θ if and only if $(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$. Therefore, $\boldsymbol{T}(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic.

- Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.
- So, $\boldsymbol{T}^*(\boldsymbol{X}) = (X_{(n)} - X_{(1)}, (X_{(1)} + X_{(n)})/2)$ is also a minimal sufficient statistic.

Ancillary statistics

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

An ancillary statistic is *unrelated* to a sufficient statistic, in a sense that sufficient statistics contain *all* the information about θ and ancillary statistics have distributions that are free of 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**.