Theory of Statistical Inference - Lecture III.3
STA422 and STA2162

Michael Evans
University of Toronto
https://utstat.utoronto.ca/mikevans/sta422/sta4222026.html

2026

- the discussion will be for the case where each f_θ is a discrete distribution for every θ (no statistical meaning is lost by doing this) but the examples will be more general

Definition III.3.1 The map $T : \mathcal{X} \to \mathcal{T}$ is a *sufficient statistic* for the model $\{f_\theta : \theta \in \Theta\}$ whenever the likelihood satisfies $L(\theta \,|\, x) = g(x)h(\theta, T(x))$ for all $\theta \in \Theta, x \in \mathcal{X}$ for some functions $g : \mathcal{X} \to (0, \infty), h : \Theta \times \mathcal{X} \to [0, \infty)$. ■

- so the likelihood ordering only depends on the data x through the value of $T(x)$ and if $T(x_1) = T(x_2)$, then x_1 and x_2 induce the same ordering and this suggests that the information locating x in $T^{-1}\{x\}$ is not useful for inference about θ

Theorem III.3.1 If T is a sufficient statistic for $\{f_\theta : \theta \in \Theta\}$, then the conditional distribution of the data given $T(x)$ does not depend on θ for those values of θ where $P_{\theta T}(\{T(x)\}) > 0$.

Proof: Suppose $P_{\theta T}(\{T(x)\}) = \sum_{z \in T^{-1}T(x)} f_\theta(z) > 0$. We have

$$\begin{aligned} f_\theta(x \mid T(x)) &= \frac{f_\theta(x)}{\sum_{z \in T^{-1}T(x)} f_\theta(z)} = \frac{g(x)h(\theta, T(x))}{\sum_{z \in T^{-1}T(x)} g(z)h(\theta, T(z))} \\ &= \frac{g(x)h(\theta, T(x))}{h(\theta, T(x)) \sum_{z \in T^{-1}T(x)} g(z)} = \frac{g(x)}{\sum_{z \in T^{-1}T(x)} g(z)}. \blacksquare \end{aligned}$$

Example III.3.1

- consider $x = (x_1, \ldots, x_n)$ a sample of n from the $U(0, \theta)$ where $\Theta = (0, \infty)$

- then $f_\theta(x) = \theta^{-n} I_{(0,\theta)}(x_{(n)})$ where $x_{(n)}$ is the largest order statistic and note that $L(\theta \mid x) = c\theta^{-n} I_{(0,\theta)}(x_{(n)})$ and so putting

$$g(x) = c, h(\theta, T(x)) = \theta^{-n} I_{(0,\theta)}(x_{(n)})$$

we have that $T(x) = x_{(n)}$ is sufficient

- the conditional distribution $x \mid T(x)$ is the same for all $\theta \in \Theta$ for which it is defined but this isn't defined when $\theta < x_{(n)}$

- when θ is true $T(x) = x_{(n)}$ has density $\theta^{-n} n x_{(n)}^{n-1} I_{(0,\theta)}(x_{(n)})$ when $x_{(n)} < \theta$ and then

$$f_\theta(x \mid T(x)) = \frac{\theta^{-n} I_{(0,\theta)}(x_{(n)}) J_T(x)}{\theta^{-n} n x_{(n)}^{n-1} I_{(0,\theta)}(x_{(n)})} = \frac{1}{x_{(n)}^{n-1}}$$

■

Theorem III.3.2 If for every x for which $P_{\theta T}(\{T(x)\}) > 0$, the conditional distribution $x \mid T(x)$ doesn't depend on θ, then T is a sufficient statistic for $\{f_\theta : \theta \in \Theta\}$.
Proof: We have that

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

and so T is sufficient. ■

Definition III.3.2 A sufficient statistic T for $\{f_\theta : \theta \in \Theta\}$ is a *minimal sufficient statistic (mss)* for $\{f_\theta : \theta \in \Theta\}$ if, for another sufficient statistic $U : \mathcal{X} \to \mathcal{U}$ for $\{f_\theta : \theta \in \Theta\}$, there is a function g_U such that $T(x) = g_U(U(x))$. ■

- a mss makes the maximum reduction in the data by discarding all information in the data irrelevant for inference about θ

- assume for each $x \in \mathcal{X}$ that there is a θ such that $f_\theta(x) > 0$ (otherwise discard)

- define an equivalence relation on $\mathcal{X}$ by $x_1 \equiv x_2$ whenever $L(\theta \,|\, x_1) = cL(\theta \,|\, x_2)$ for some constant $c > 0$ for every $\theta \in \Theta$

- let $[x]$ denote the equivalence class containing x

Theorem III.3.3 $[\cdot]$ is a mss for $\{f_\theta : \theta \in \Theta\}$.
Proof: Assume $P_\theta([x]) = \sum_{z\in[x]} f_\theta(z) > 0$. If $z \in [x]$,then $f_\theta(z) = c(z,x) f_\theta(x)$ and so

$$\begin{aligned} f_\theta(x\,|\,[x]) &= \frac{f_\theta(x)}{\sum_{z\in[x]} f_\theta(z)} = \frac{f_\theta(x)}{f_\theta(x)\sum_{z\in[x]} c(z,x)} \\ &= \left(\sum_{z\in[x]} c(z,x)\right)^{-1} \end{aligned}$$

which implies $[\cdot]$ is sufficient by Theorem III.3.2.
If U is a sufficient statistic for $\{f_\theta : \theta \in \Theta\}$, then $L(\theta\,|\,x) = g(x)h(\theta, U(x)) = g(x)h(\theta, U(z))$ whenever $U(z) = U(x)$. Since

$$L(\theta\,|\,z) = g(z)h(\theta, U(z)) = \frac{g(z)}{g(x)} g(x)h(\theta, U(x)) = \frac{g(z)}{g(x)} L(\theta\,|\,x)$$

and $g(z)/g(x) > 0$ (recall $g(x) > 0$) so $z \in [x]$. Then define g_U by $g_U(U(x)) = [x]$ which establishes the result. ■

note - there exist counterexamples to the existence of a mss but of necessity they involve models with continuous probability distributions and as such they are irrelevant for the development of a theory of statistical inference since the concept works fine in the case where $\mathcal{X}$ and Θ are finite

Exercise III.3.1 Prove that any 1-1 function of a mss for $\{f_\theta : \theta \in \Theta\}$ is a mss for $\{f_\theta : \theta \in \Theta\}$. So, if we pick a single $x_* \in [x]$ for each equivalence class and define $T(x) = x_*$ (a cross-section), then T is a mss.

Exercise III.3.2 For the following model determine a mss.

$\theta \setminus x$	1	2	3	4
a	1/8	1/8	1/4	1/2
b	0	1/2	1/6	1/3
c	0	5/8	1/8	1/4

Exercise III.3.3 Show that there is a 1-1 correspondence between a mss and the set of possible likelihood functions for a model (recall that the likelihood doesn't change when multiplied by a positive constant).

- sufficiency is a key component of frequentist approaches to statistical inference but it is not needed for likelihood based approaches because it is automatically satisfied, namely, inferences depend on the data only through the value of a mss

- but the fact that the conditional distribution given a mss is independent of θ **is** relevant to all theories because it can play a key role in model checking (later)