Theory of Statistical Inference - Lecture III.1
STA422 and STA2162

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

2026

- the *inference base* for likelihood inferences is $(\{f_\theta : \theta \in \Theta\}, x)$

- **note** - when discussing inference it is always assumed that the data was collected "correctly" and that the model is correct (the true distribution from which x was generated is in $\{f_\theta : \theta \in \Theta\}$)

- of course the model is wrong but what we really care about is that the inferences for $\psi = \Psi(\theta)$ are "correct" in the sense that the model isn't so wrong that we are being misled by the inference base

- checking that $\{f_\theta : \theta \in \Theta\}$ is at least approximately correct (model checking) is part of statistical reasoning but it is not part of the inference step

- we want to define the likelihood function prescribed by $(\{f_\theta : \theta \in \Theta\}, x)$

- recall that f_θ is a probability density on the sample space $\mathcal{X}$ wrt support measure $\mu_{\mathcal{X}}$ so

$$P_\theta(B) = \int_B f_\theta(z)\,\mu_{\mathcal{X}}(dz) \tag{1}$$

- when P_θ is a discrete probability measure then $\mu_{\mathcal{X}}$ can be taken as counting measure and $f_\theta(z) = P_\theta(\{z\})$, $P_\theta(B) = \sum_{z\in B} f_\theta(z)$ and there is no ambiguity

- but in the absolutely continuous case, any density f_θ can be changed on a set having $\mu_{\mathcal{X}}$ measure 0 and (1) will still hold

Example III.1.1

- suppose $\mathcal{X} = \mathbb{R}^k$ and $\mu_{\mathcal{X}} =$ volume measure and suppose $\{f_\theta : \theta \in \Theta\}$ is a family of density functions (e.g., normal distributions)

- define

$$f_\theta^{\#}(z) = \begin{cases} 0 & z \in \mathbb{Q}^k \\ f_\theta(z) & \text{otherwise} \end{cases}$$

then $f_\theta^{\#}$ satisfies (1) ■

- so, if we define some basic statistical concept in terms of density, then it would appear that we have a problem, e.g., in practice the elements of x are always rational numbers

- but is the only purpose of a density to provide a way of computing $P_\theta(B)$?

- for me all continuous distributions serve as approximations to a context that is fundamentally discrete (even finite)

- so (1) is saying that when B is a "small" neighborhood of a point x, then

$$P_\theta(B) \approx f_\theta(x)\mu_{\mathcal{X}}(B) \tag{2}$$

- to be precise, if $B_\delta(x)$ is a neighborhood of x for each $\delta > 0$ and $B_\delta(x) \downarrow \{x\}$ ("nicely" see Rudin (1974)) as $\delta \to 0$, then

$$f_\theta(x) = \lim_{\delta \to 0} \frac{P_\theta(B_\delta(x))}{\mu_{\mathcal{X}}(B_\delta(x))} \tag{3}$$

- e.g. take $B_\delta(x) =$ ball of radius δ centered at x, or $B_\delta(x) =$ rectangle centered at x where the length of the sides go to 0 at the same rate as controlled by δ, etc.

- **fact**: if a version of f_θ exists which is continuous at x, then (3) equals that value

- but the real justification for choosing $f_\theta(x)$ as given by (3) is that $f_\theta(x)$ is then the density = probability mass per unit volume at x and otherwise it isn't

- so hereafter we will assume that all densities are defined correctly

Definition III.1.1 For inference base $(\{f_\theta : \theta \in \Theta\}, x)$ a *likelihood function* is a function $L(\cdot \mid x) : \Theta \longrightarrow [0, \infty)$ given by

$$L(\theta \mid x) = cf_\theta(x)$$

for any constant $c > 0$. ■

- the likelihood function is a basic concept in statistics (although no need for it in the proper Bayesian formulation as then $f_\theta(x)$ is the conditional density of x given θ)

- the choice of c is made for convenience and doesn't matter, why?

- recall $f_\theta(x)\mu_{\mathcal{X}}(B_\delta(x))$ can be interpreted as the probability of x (provided $B_\delta(x)$ is not too big or too small) so we can compare this probability at different values of θ and this induces an ordering (a preorder) on Θ

Definition III.1.2 For inference base $(\{f_\theta : \theta \in \Theta\}, x)$ the *likelihood ordering* is given by $\theta_1 \succcurlyeq \theta_2$ (read θ_1 is preferred to θ_2) whenever $L(\theta_1 \mid x) \geq L(\theta_2 \mid x)$. ■

- note - $\theta_1 \succcurlyeq \theta_2$ and $\theta_2 \succcurlyeq \theta_1$ iff $L(\theta_1 \mid x) = L(\theta_2 \mid x)$ whence θ_1 and θ_2 are considered as equivalent

- note that the ordering is independent of c

- the ordering can also be defined via *likelihood ratios* so, $\theta_1 \succcurlyeq \theta_2$ whenever

$$\frac{L(\theta_1 \mid x)}{L(\theta_2 \mid x)} = \frac{cf_{\theta_1}(x)}{cf_{\theta_2}(x)} = \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \geq 1$$

(provided at least one of $f_{\theta_1}(x) \neq 0$ or $f_{\theta_2}(x) \neq 0$) which is also independent of c

Example III.1.2

- the following is a model with $\mathcal{X} = \{1, 2, 3, 4\}$ and $\Theta = \{\spadesuit, \heartsuit, \diamondsuit\}$

$\theta \setminus x$	1	2	3	4
$\spadesuit$	1/4	1/4	1/4	1/4
$\heartsuit$	1/3	1/3	1/3	0
$\diamondsuit$	1/8	1/8	1/4	1/2

- there are 4 possible likelihood functions, e.g., when $x = 4$, letting $c = 1$, $L(\spadesuit \mid 4) = 1/4$, $L(\heartsuit \mid 4) = 0$, $L(\diamondsuit \mid 4) = 1/2$ which induces the ordering $\diamondsuit \succcurlyeq \spadesuit \succcurlyeq \heartsuit$ ■

- the likelihood ordering seems very natural and we could accept the following intuitive principle (axiom) of inference: any inference about θ **must** conform to the likelihood ordering

- e.g. if the inference about θ is to quote a region $C(x) \subset \Theta$ that is supposed to contain θ_{true} then, if $\theta \in C(x)$ and $\theta' \succcurlyeq \theta$, then $\theta' \in\in C(x)$, i.e., $C(x)$ must be a likelihood region

Definition III.1.3 For inference base $(\{f_\theta : \theta \in \Theta\}, x)$ a *likelihood region* for θ is a set of the form

$$C_k(x) = \{\theta : L(\theta \mid x) \geq k\}$$

for some $k \geq 0$. ■

Example III.1.2 (continued)

- we have the following possible likelihood regions when $x = 1$

$$\begin{aligned} C_k(1) &= \phi \text{ when } k > 1/3 \\ C_k(1) &= \{\heartsuit\} \text{ when } 1/4 < k \leq 1/3 \\ C_k(1) &= \{\spadesuit, \heartsuit\} \text{ when } 1/8 < k \leq 1/4 \\ C_k(1) &= \Theta \text{ when } 0 \leq k \leq 1/8 \end{aligned}$$

■

Exercise III.1.1 Determine all the likelihood regions in **Example III.1.1.2.**

- what likelihood region should be reported?

- more importantly, if our parameter of interest is $\psi = \Psi(\theta)$ what is the relevant likelihood ordering?

Example III.1.2 (continued)

- suppose interest is in $\psi = \Psi(\theta) = I_{\{\heartsuit,\diamondsuit\}}(\theta)$

- based on $(\{f_\theta : \theta \in \Theta\}, x)$ the likelihood ordering for ψ is not defined in any obvious way as when $x = 1$, $\heartsuit \succcurlyeq \spadesuit$ but $\spadesuit \succcurlyeq \diamondsuit$ ■

- we need to define a likelihood ordering for a general Ψ, how?

- the likelihood ordering for θ possesses an important property

- define a *relabelling* of a parameter $\psi = \Psi(\theta)$ as a 1-1 function $T : \Psi(\Theta) \to T(\Psi(\Theta))$ which leads to the parameter $\tau = T(\psi) = T(\Psi(\theta))$

- since ψ and τ are essentially referencing the same object it should not matter which parameterization we use

> **Invariance Principle:** *since T is 1-1, the ordering of ψ values, as in $\psi_1 \succeq \psi_2$ must also lead to $\tau_1 = T(\psi_1) \succeq \tau_2 = T(\psi_2)$ and conversely.*

Theorem III.1.1 The likelihood ordering for θ satisfies the invariance principle.

Proof: Let $T : \Theta \to T(\Theta)$ be a relabelling of the model parameter. Then $\{f_\theta : \theta \in \Theta\} = \{f_\tau : \tau \in T(\Theta)\}$ where $f_\tau = f_{T(\theta)}$ and $L(\theta_1 \mid x) \geq L(\theta_2 \mid x)$ iff $L(\tau_1 \mid x) \geq L(\tau_2 \mid x)$. ■

- there are some (many) approaches to inference that do not satisfy the invariance principle

- it is common to modify the likelihood function to the relative likelihood function

Definition III.1.3 For inference base $(\{f_\theta : \theta \in \Theta\}, x)$ the *relative likelihood function* is the function $L(\cdot \mid x) : \Theta \longrightarrow [0, \infty)$ given by

$$L^{rel}(\theta \mid x) = \frac{f_\theta(x)}{\sup\{f_\theta(x) : \theta \in \Theta\}}.$$

- note - the relative likelihood does not change the likelihood ordering and takes values in $[0, 1]$

Definition III.1.4 For inference base $(\{f_\theta : \theta \in \Theta\}, x)$ a *maximum likelihood estimate (MLE)* of θ is a value $\theta(x)$ that satisfies $\theta(x) \succcurlyeq \theta$ for every $\theta \in \Theta$. ■

- so the MLE satisfies $\theta(x) \in \{\theta : L^{rel}(\theta \mid x) = 1\}$ and is typically unique

- the MLE is the natural estimate of θ because it is preferred relative to all other θ values according to the likelihood ordering

- note $\theta(x) \in C_k(x)$ for any nonempty likelihood region

- so we can quote $\theta(x)$ together with a likelihood region C_k as our answer to the **E** problem for θ, but which likelihood region should we use?

Example III.1.3 *binomial*

- suppose $x = (x_1, \ldots, x_n) \overset{iid}{\sim}$ Bernoulli(θ) where $\theta \in [0, 1]$ is unknown specifies the model

- likelihood

$$L(\theta \mid x) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$$

where $n\bar{x} = \#$ of 1's in the sample

- now

$$l(\theta \mid x) = \log L(\theta \mid x) = n\bar{x} \log \theta + (n - n\bar{x}) \log(1-\theta)$$

is the *log-likelihood function* and since log is a strictly increasing function the MLE $\theta(x)$ also maximizes $l(\theta \mid x)$

- when, as in this case, $l(\cdot \mid x)$ is differentiable, then

$$S(\theta \mid x) = \frac{\partial l(\theta \mid x)}{\partial \theta} = \frac{n\bar{x}}{\theta} - \frac{n - n\bar{x}}{1-\theta} = \frac{n(\bar{x} - \theta)}{\theta(1-\theta)}$$

is the *score function* and so the MLE satisfies $S(\theta(x) \mid x) = 0$ which implies $\theta(x) = \bar{x}$ as $l(\cdot \mid x)$ increases to the left of $\bar{x}$ and decreases to the right (and so does $L(\theta \mid x)$)

- so the relative likelihood is

$$L^{rel}(\theta \mid x) = \frac{\theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}}{\bar{x}^{n\bar{x}}(1-\bar{x})^{n-n\bar{x}}}$$

- for $k \in [0,1]$ the likelihood region is

$$C_k(x) = \left\{\theta : \frac{\theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}}{\bar{x}^{n\bar{x}}(1-\bar{x})^{n-n\bar{x}}} \geq k\right\}$$

- since the likelihood function is unimodal in θ with mode at $\theta = \bar{x}$, then $C_k(x) = [l_k(\bar{x}), u_k(\bar{x})]$ is an interval with $\bar{x} \in C_k(x)$

- we are left with the question: how do we choose $k \in [0,1]$ so that $[l_k(\bar{x}), u_k(\bar{x})]$ provides an appropriate assessment of the accuracy of the estimate $\bar{x}$? ■