Theory of Statistical Inference - Lecture III.8
STA422 and STA2162

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

2026

- we have inference base $(\{f_\theta : \theta \in \Theta\}, x)$ and likelihood $L(\theta \,|\, x) = cf_\theta(x)$

- the difference from pure likelihood inference about θ is that now repeated sampling properties play a role

- for the **E** problem the natural estimate is a MLE

$$\theta(x) = \arg\sup L(\theta \,|\, x)$$

as this maximizes the probability of the observed data over all possible values of θ

- a natural assessment of the error (how much it differs from the true value of θ) in $\theta(x)$ would be to quote a relevant likelihood region say

$$C_k(x) = \{\theta : L^{rel}(\theta \,|\, x) \geq k\}$$

for some $k \in [0,1]$, as $\theta(x) \in C_k(x)$ for all such k, and then use the size of $C_k(x)$

- the question then is, which $k \in [0,1]$?

- perhaps the most obvious answer is to select a coverage probability $\gamma \in [0,1]$ and then choose the largest k so that

$$P_\theta(\theta \in C_k(x)) \geq \gamma$$

for all $\theta \in \Theta$, so $C_k(x)$ is γ-confidence region for θ

- this works fine for *some* examples

Example III.8.1

- $x = (x_1, \ldots, x_n)$ an iid sample from $\{N(\mu, \sigma_0^2) : \mu \in \mathbb{R}\}$, with $\Psi(\mu) = \mu$ then $L^{rel}(\mu \mid x) = \exp(-n(\bar{x} - \mu)^2/2\sigma_0^2)$ so the MLE is $\mu(x) = \bar{x}$ and for $k \in [0,1]$

$$\begin{aligned} C_k(x) &= \{\mu : \exp(-n(\bar{x} - \mu)^2/2\sigma_0^2) \geq k\} \\ &= \left\{\mu : (\bar{x} - \mu)^2 \leq -\frac{2\sigma_0^2}{n} \log k\right\} \\ &= \left[\bar{x} - \sqrt{-\frac{2\sigma_0^2}{n} \log k}, \bar{x} + \sqrt{-\frac{2\sigma_0^2}{n} \log k}\right] \end{aligned}$$

-this gives exact γ coverage by choosing k s.t.

$$\begin{aligned}\sqrt{-\frac{2\sigma_0^2}{n}\log k} &= \frac{\sigma_0}{\sqrt{n}} z_{(1+\gamma)/2} \text{ or} \\ k &= \exp\left(-\frac{z_{(1+\gamma)/2}^2}{2}\right)\end{aligned}$$

and so $C_k(x) = [\bar{x} - z_{(1+\gamma)/2}\sigma_0/\sqrt{n}, \bar{x} + z_{(1+\gamma)/2}\sigma_0/\sqrt{n}\}]$, which is the same as the interval using a pivotal in Example III.7.1 ■

- consider, however, the following example

Example III.8.2

- $x = (x_1, \ldots, x_n)$ an iid sample from $\{\text{Bernoulli}(\theta) : \theta \in [0,1]\}$, with $\Psi(\theta) = \theta$ then

$$L^{rel}(\theta \,|\, x) = \left(\frac{\theta}{\bar{x}}\right)^{n\bar{x}} \left(\frac{1-\theta}{1-\bar{x}}\right)^{n(1-\bar{x})}$$

so the MLE is $\theta(x) = \bar{x}$ and for $k \in [0,1]$

$$\begin{aligned} C_k(x) &= \left\{\theta : \left(\frac{\theta}{\bar{x}}\right)^{n\bar{x}} \left(\frac{1-\theta}{1-\bar{x}}\right)^{n(1-\bar{x}))} \geq k\right\} \\ &= \left\{\theta : \theta^{n\bar{x}} (1-\theta)^{n(1-\bar{x}))} - k\bar{x}^{n\bar{x}} (1-\bar{x})^{n(1-\bar{x}))} \geq 0\right\} \end{aligned}$$

and we can find the two values of $\theta \in [0,1]$ that give equality by finding the roots of the polynomial of degree n on the left

- but this does not tell us what k has to be to give coverage at least γ

- is there a solution? ■

- so the frequentist likelihood approach does not lead to a solution to one of the simplest inference problems

- rather, likelihood asymptotics are used (see below) to obtain approximate confidence intervals but these do not satisfy the likelihood ordering

- **many** different confidence intervals have been proposed for this problem, which one is right?

- Brown, Cai and Das Gupta (2002) Confidence Intervals for a binomial proportion and asymptotic expansions. Ann. Statist. 30(1): 160-201, does a comparison among these and makes recommendations but all of these are asymptotic and none are likelihood intervals

- likelihood asymptotics

- *log-likelihood* $l(\theta \,|\, x) = \log L(\theta \,|\, x)$

- *score function* $S(\theta \,|\, x) = \left(\frac{\partial l(\theta \,|\, x)}{\partial \theta_i} \right) \in \mathbb{R}^k$ assuming $\Theta \subset \mathbb{R}^k$ is open and assume all the partials are continuously differentiable

Lemma III.8.1 Under conditions (dominated derivative theorem) $E_\theta\left(S(\theta \mid x)\right) = 0$ for all $\theta \in \Theta$.
Proof:

$$\begin{aligned} E_\theta\left(\frac{\partial l(\theta \mid x)}{\partial \theta_i}\right) &= \int_{\mathcal{X}} \frac{\frac{\partial f_\theta(x)}{\partial \theta_i}}{f_\theta(x)} f_\theta(x)\,\nu(dx) = \int_{\mathcal{X}} \frac{\partial f_\theta(x)}{\partial \theta_i}\,\nu(dx) \\ &= \frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} f_\theta(x)\,\nu(dx) = \frac{\partial}{\partial \theta_i} 1 = 0. \blacksquare \end{aligned}$$

- *observed Fisher information* $\hat{\jmath}(\theta(x)) = \left(-\frac{\partial^2 l(\theta(x) \mid x)}{\partial \theta_i \partial \theta_j}\right) \in \mathbb{R}^{k \times k}$ assuming all second partials are continuous

- *Fisher information* $j(\theta) = E_\theta\left(-\frac{\partial^2 l(\theta \mid x)}{\partial \theta_i \partial \theta_j}\right) \in \mathbb{R}^{k \times k}$ assuming all expectations are finite

Lemma III.8.2 Under conditions $j(\theta) = Var_\theta(S(\theta\,|\,x))$ for all $\theta \in \Theta$.
Proof:

$$\begin{aligned}
E_\theta\left(\frac{\partial^2 l(\theta\,|\,x)}{\partial\theta_i\partial\theta_j}\right) &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left(\frac{\frac{\partial f_\theta(x)}{\partial\theta_j}}{f_\theta(x)}\right) f_\theta(x)\,\nu(dx) \\
&= \int_{\mathcal{X}} \left(\frac{\frac{\partial^2 f_\theta(x)}{\partial\theta_i\partial\theta_j}}{f_\theta(x)} - \frac{\frac{\partial f_\theta(x)}{\partial\theta_i}}{f_\theta(x)}\frac{\frac{\partial f_\theta(x)}{\partial\theta_j}}{f_\theta(x)}\right) f_\theta(x)\,\nu(dx) \\
&= \int_{\mathcal{X}} \frac{\partial^2 f_\theta(x)}{\partial\theta_i\partial\theta_j}\,\nu(dx) - \int_{\mathcal{X}} S_i(\theta\,|\,x)S_j(\theta\,|\,x)f_\theta(x)\,\nu(dx) \\
&= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int_{\mathcal{X}} f_\theta(x)\,\nu(dx) - Cov_\theta(S_i(\theta\,|\,x)S_j(\theta\,|\,x)) \\
&= -Cov_\theta(S_i(\theta\,|\,x)S_j(\theta\,|\,x)). \blacksquare
\end{aligned}$$

Theorem III.8.1 Under conditions, when $x = (x_1, \ldots, x_n)$ is iid f_θ, then as $n \longrightarrow \infty$

(i) Wald statistic

$$\hat{\jmath}^{1/2}(\theta(x))(\theta(x) - \theta) \xrightarrow{d} N_k(0, I)$$

(ii) Rao statistic

$$\hat{\jmath}^{-1/2}(\theta(x))S(\theta \,|\, x) \xrightarrow{d} N_k(0, I)$$

(iii) Wilks statistic

$$2\{l(\theta(x) \,|\, x) - l(\theta \,|\, x)\} \xrightarrow{d} \text{chi-square}(\dim \theta)$$

Example III.8.2 *(continued)*

- $\theta(x) = \bar{x}$
- $l(\theta \,|\, x) = n\bar{x}\log\theta + n(1-\bar{x})\log(1-\theta)$
- $S(\theta \,|\, x) = n\bar{x}/\theta - n(1-\bar{x})/(1-\theta) = n(\bar{x}-\theta)/\theta(1-\theta)$
- $\partial^2 l(\theta \,|\, x)/\partial\theta^2 = -n\bar{x}/\theta^2 - n(1-\bar{x})/(1-\theta)^2$
- $\hat{\jmath}(\theta(x)) = n/\bar{x} - n/(1-\bar{x}) = n/\bar{x}(1-\bar{x})$

- Wald statistic leads to approximate γ-confidence interval

$$\bar{x} \pm \sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}} z_{(1+\gamma)/2}$$

- Rao statistic leads to approximate γ-confidence interval (called the *Wilson interval*)

$$\frac{\bar{x} + \frac{z^2_{(1+\gamma)/2}}{2n} \pm z_{(1+\gamma)/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n} + \frac{z^2_{(1+\gamma)/2}}{4n^2}}}{1 + \frac{z^2_{(1+\gamma)/2}}{n}}$$

- Wilks statistic leads to an approximate γ-confidence (likelihood) interval that needs to be solved for numerically ■

- the **H** problem can also be addressed via these statistics via using them to compute p-values

- the frequentist approach does not resolve the marginal parameter problem, namely, inferences about $\psi = \Psi(\theta)$ when Ψ is many-to-one, as using the profile likelihood is again the general approach but with the asymptotic frequentist aspect added

- Brown, Cai and Das Gupta (2001) Interval estimation for a binomial proportion (with discussion). Statist. Sci. 16 101–133.

> *"Interval estimation of a binomial proportion is a very basic problem in practical statistics. The standard Wald interval is in nearly universal use. We first show that the performance of this standard interval is persistently chaotic and unacceptably poor. Indeed its coverage properties defy conventional wisdom. The performance is so erratic and the qualifications given in the influential texts are so defective that the standard interval should not be used. We provide a fairly comprehensive evaluation of many natural alternative intervals. Based on this analysis, we recommend the Wilson or the equal-tailedJeffreys prior interval for small n, $(n \leq 40)$. These two intervals are comparable in both absolute error and length for $n \leq 40$, and we believe that either could be used, depending on taste."*

- Jeffreys prior referenced here is $\theta \sim \text{beta}(1/2, 1/2)$ and the interval is obtained by discarding $(1 - \gamma)/2$ from each tail of the posterior, none of the intervals recommended are (finite sample) likelihood intervals

Personal Opinion

Is the construction of confidence regions that attain (or are close to attaining) exact coverage probabilities (confidences) an appropriate major goal of a theory of statistical inference? In my opinion, the answer is no. Confidence itself is a somewhat recalcitrant concept. It just doesn't lend itself to developments that produce satisfactory answers in general, even in fairly simple contexts like the binomial, and it has a habit of producing apparently unresolvable paradoxes. This isn't to say that coverage probabilities have no role to play, but the role is somewhat different than being presented as primary inferences. We will subsequently discuss this further.

- if interested in a deeper dive into the world of confidence see the work of Andre Plante, e. g., Plante (2020) A Gaussian alternative to using improper confidence intervals. Canadian Journal of Statistics, 48(4).