Theory of Statistical Inference - Lecture III.7
STA422 and STA2162

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

2026

- we consider frequentist inferences based on the inference base $(\{f_\theta : \theta \in \Theta\}, x)$

- principle of frequentism, as a philosophical approach to statistical problems, is something like the following

> *Inferences are justified by their behavior in a (counterfactual) sequence of independent repeats of the data collection process.*

- so we imagine a sequence of independent repeats of the data collection process that produced x, say $x_1, x_2, \ldots$

- then, say for the **E** problem, we consider how an estimator $\psi(x)$ of $\psi = \Psi(\theta)$ performs in the sequence producing estimates $\psi(x_1), \psi(x_2), \ldots$ and we do that for each possible value of $\theta \in \Theta$, i.e., the x_i are iid f_θ

- as is typical there are many different estimators and we want one that has good properties no matter what the true value of $\psi = \Psi(\theta)$ is

- there are many approaches to making such comparisons, e.g. mean-squared error $= E_\theta(\psi(x) - \Psi(\theta))^2$, but most of these belong in our discussion of decision theory

- why is the principle of frequentism relevant?

- because the frequentist properties of an inference tells us about its reliability and that seems like an essential component of any approach to inference

- for example, we want any estimator we use to be *consistent* in the sense that, as the amount of data grows the estimate converges to the true value wp1

- so, intuitively, the more data we have, the more reliable is the estimate, but the question is: how reliable is the estimate at a given sample size?

- there are three main tools that form the basis for much of frequentist inference (as opposed to decision theory which we will discuss later):

(1) p-values
(2) confidence regions
(3) the likelihood function

- the likelihood appears again but now its repeated sampling behavior is also of importance

- we have the inference base $(\{f_\theta : \theta \in \Theta\}, x)$ and we want to assess the hypothesis $H_0 : \Psi(\theta) = \psi_0$ for some specific value ψ_0

- the idea is that, if the observed data is surprising for every distribution f_θ s.t. $\Psi(\theta) = \psi_0$, then we have "evidence" that H_0 is false

- how do we measure this?

- it seems clear that any such assessment must depend on the model being true, otherwise the falsity of the model could leads us to erroneously conclude H_0 is false

- some would argue that this is not a problem, but recall, Ψ is a real-world object and its meaning is independent of the model

- so, to avoid such a confounding, we need to check the model first (more on this later) and, provided the model is okay, we assume it is true for the inference step

- as part of avoiding the confounding, we only use the relevant part of the data to make the assessment, namely, the observed value of a mss T

- now suppose U is some real-valued statistic that only depends on the data through T

- typically, a p-value for $H_0 : \Psi(\theta) = \psi_0$ would compute something like

$$\sup_{\theta \in \Psi^{-1}\{\psi_0\}} P_{\theta U}(U > U(T(x)) \,|\, x), \tag{1}$$

where $P_{\theta U}$ is the marginal probability measure induced by U when θ is true

- note - is there actually a formal, precise definition of a p-value?

- then if (1) is small it is concluded that the observed value $U(T(x))$ is surprising and we have "evidence" against H_0

- note that (1) essentially assumes that the distribution $P_{\theta U}$ is unimodal for each θ and the right-tail contains the surprising values, otherwise a different measure is probably needed, e.g.,

$$\sup_{\theta \in \Psi^{-1}\{\psi_0\}} P_{\theta U}(f_{\theta U}(U) > f_{\theta U}(U(T(x))) \,|\, x),$$

Example III.7.1

- $x = (x_1, \ldots, x_n)$ an iid sample from $\{N(\mu, \sigma_0^2) : \mu \in \mathbb{R}\}$, with $\Psi(\mu) = \mu$ and $H_0 : \Psi(\mu) = \mu_0$

- mss is $T(x) = \bar{x} \sim N(\mu, \sigma_0^2/n)$ and we take $U(T) = \left|\sqrt{n}(T - \mu_0)/\sigma_0\right|$ where $\sqrt{n}(T - \mu_0)/\sigma_0 \sim N(0, \sigma_0^2/n)$ when H_0 is true so (1) equals

$$2(1 - \Phi(\left|\sqrt{n}(\bar{x} - \mu_0)/\sigma_0\right|)) \blacksquare \tag{2}$$

- questions/problems:

(1) How small does (1) have to be to say there is "evidence" against H_0?

- no clear answer

(2) If we agree on a cut-off, say (1) $< \alpha$, ($\alpha = 0.05$?) for evidence against then is (1) $\geq \alpha$ evidence in favor?

- no, why? when H_0 is true (1) can have a $U(0,1)$ distribution when $x \sim P_\theta$ for $\theta \in \Psi^{-1}\{\psi_0\}$, so large values are as likely as small values and note, as $n \to \infty$, while (1) typically converges to 0 when H_0 is false, it does not converge to 1 when H_0 is true

Exercise III.7.1 Establish the above two statements for (2).

(3) How do we accommodate the fact that a hypothesis like that in Example III.7.1. is always false but perhaps not meaningfully false, and, as the sample size increases, evidence against will inevitably be found against H_0?

- probably we need to change $H_0 : \Psi(\theta) = \psi_0$ to $H_0 : \Psi(\theta) \in (\psi_0 - \delta, \psi_0 + \delta)$ where $\delta =$ "the difference that matters", namely, if $\psi_{true} \notin (\psi_0 - \delta, \psi_0 + \delta)$, then the hypothesized value ψ_0 is meaningfully false

- recall that in power calculations such a δ is used, namely, one specifies an α and then computes

$$\sup_{\theta' \in \Psi^{-1}\{\psi_0 \pm \delta\}} P_{\theta'} \left(\sup_{\theta \in \Psi^{-1}\{\psi_0\}} P_{\theta U}(U > U(T(x)) \,|\, x) \leq \alpha \right),$$

and then chooses a sample size to make this probability large

Exercise III.7.2 Show how to do the power calculations for Example III.7.1.

(4) is there a theory for constructing a p-value for an arbitrary problem?

- see Royall (1997) for an excellent discussion of p-values and rejection trials (when you specify an α to determine when there is evidence against)

> *Personal Opinion*
> *There are many problems, but the bottom line is that p-values do not measure statistical evidence properly and this can be seen from the fact that a p-value does not tell us unambiguously when we have evidence for or against* H_0.

Definition III.7.1 For confidence level $\gamma \in [0,1]$, a *γ-confidence region* for Ψ is a map $C : \mathcal{X} \to 2^{\Psi(\Theta)}$ s.t. $\inf_{\theta\in\Theta} P_\theta(\Psi(\theta) \in C(x)) \geq \gamma$. ■

- $C(x) = \Psi(\Theta)$ is a γ-CR for Ψ for every γ so we need something more to specify C and typically this means get the coverage probabilities $P_\theta(\Psi(\theta) \in C(x))$ as close as possible to γ, while remaining conservative

- so a preference for γ-CR's where the coverage is exactly γ for every $\theta \in \Theta$

- what is the point of quoting γ-CR $C(x)$?

- sometimes it is said that $C(x) \subset \Psi(\Theta)$ provides an indication of where the true value of ψ lies with high confidence when γ is close to 1

- more importantly, typically C can be associated with an estimate $\psi(x) \in C(x)$ and then the "size" of $C(x)$ is taken as a measure of the accuracy of $\psi(x)$

- so how to obtain such a C?

- (strangely) this is commonly approached via hypothesis assessment

- for a r.v. X with cdf F define the γ-th quantile by

$$x_\gamma = \inf\{x : F(x) \geq \gamma\}$$

so $F(x_\gamma) \geq \gamma$

- suppose for $\gamma = 1 - \alpha$ and for each $H_0 : \Psi(\theta) = \psi_0$ with test statistic U_{ψ_0} and let $u_{\psi_0,\gamma}$ = the supremum of all the γ-th quantiles of U_{ψ_0} for $\theta \in \Psi^{-1}\{\psi_0\}$, so

$$\sup_{\theta \in \Psi^{-1}\{\psi_0\}} P_{\theta U_{\psi_0}}(U_{\psi_0} > u_{\psi_0,\gamma}) \leq 1 - \gamma = \alpha$$

- put

$$C(x) = \{\psi_0 : U_{\psi_0}(T(x)) \leq u_{\psi_0,\gamma}\}$$

the values of ψ for which no evidence against at size α is found

- then

$$\begin{aligned} P_\theta(\Psi(\theta) \in C(x)) &= P_\theta\left(U_{\Psi(\theta)}(T(x)) \leq u_{\Psi(\theta),\gamma}\right) \\ &= 1 - P_\theta\left(U_{\Psi(\theta)}(T(x)) > u_{\Psi(\theta),\gamma}\right) \\ &\geq 1 - (1-\gamma) = \gamma \end{aligned}$$

so C is a γ-CR for $\psi = \Psi(\theta)$

- many CR's are formed using *pivotals*: $U_\psi(T)$ is pivotal if it has a fixed distribution, as given by r.v. U, for every $\psi \in \Psi(\Theta)$ and let $C(x) = \{\psi : U_\psi(T(x)) \leq u_\gamma\}$

Example III.7.1 (continued)

- $x = (x_1, \ldots, x_n)$ an iid sample from $\{N(\mu, \sigma_0^2) : \mu \in \mathbb{R}\}$, with $\Psi(\mu) = \mu$ and $H_0 : \Psi(\mu) = \mu_0$

- then $U_\mu(T) = \left|\sqrt{n}(T - \mu)/\sigma_0\right| = |U|$ where $U \sim N(0,1)$ is pivotal

$$\begin{aligned} \gamma &= P(|U| \leq u_\gamma) = P(-u_\gamma \leq U \leq u_\gamma) \\ &= \Phi(u_\gamma) - \Phi(-u_\gamma) = 2\Phi(u_\gamma) - 1 \end{aligned}$$

so $u_\gamma = z_{(1+\gamma)/2}$ = the $(1+\gamma)/2$-th quantile of the $N(0,1)$ and

$$\begin{aligned} C(x) &= \{\mu : -z_{(1+\gamma)/2} \leq \sqrt{n}(T(x) - \mu)/\sigma_0 \leq z_{(1+\gamma)/2}\} \\ &= [\bar{x} - z_{(1+\gamma)/2}\sigma_0/\sqrt{n}, \bar{x} + z_{(1+\gamma)/2}\sigma_0/\sqrt{n}\}] \end{aligned}$$

is a γ-confidence interval for μ ■

- pivotals don't always exist, e.g. binomial(n, θ) and we want CI's for θ

- questions/problems

(1) - such regions should be associated with an estimator

(2) - as with p-values there doesn't appear to be a general theory

(3) - even when a pivotal exists arise problems can arise

Example III.7.1 (continued)

- a physicist is estimating the mass of a neutrino and uses the assumptions of this example for some σ_0

- we want to incorporate the restriction $\mu \geq 0$ but it is necessary to avoid the CR containing negative values which

$$C(x) = [\bar{x} - z_{(1+\gamma)/2}\sigma_0/\sqrt{n}, \bar{x} + z_{(1+\gamma)/2}\sigma_0/\sqrt{n}\}]$$

may do (mass of neutrino is small if positive)

- taking $C(x) \cap [0, \infty)$ will have the right coverage probability, but its length is not an appropriate assessment of the accuracy of $\bar{x}$ as an estimate of the mass and moreover there will be a positive probability that $C(x) \cap [0, \infty) = \phi$

- physicists developed the Feldman-Cousins (1997) intervals which resolve the problem in this case, but they are not a perfect solution as other problems arise ■

Example III.7.2 *Fieller's problem*

- two samples $x = (x_1, \ldots, x_{n_x})$ iid $N(\mu, \sigma_0^2)$ independent of $y = (y_1, \ldots, y_{n_y})$ iid $N(\nu, \sigma_0^2)$ where $(\mu, \nu) \in \mathbb{R}^2$ is unknown

- $\psi = \Psi(\mu, \nu) = \mu / \nu$

- pivotal exists

$$U(\bar{x}, \bar{y}) = \frac{(\bar{x} - \bar{y}\psi)}{\sigma_0 \sqrt{1/n_x + \psi^2/n_y}} \sim N(0, 1)$$

- but this can produce 0.95-CI $= \mathbb{R}$ ■

Personal Opinion

The theory of confidence regions does not resolve any of the issues associated with characterizing statistical evidence. If anything, the concept of confidence simply compounds them. For example, is it the case that the values in $C(x)$ have evidence in their favor. The lack of a complete theory, and many anomalous examples, suggests that the confidence concept is not a foundational concept to build a theory of inference on.