The Measurement of Statistical Evidence
Lecture 3 - part 2

Michael Evans
University of Toronto
http://www.utstat.utoronto.ca/mikevans/sta4522/STA4522.html

2021

2. Frequentism and Birnbaum's Theorem

- *frequentism* in statistics means that any statistical procedure must be justified based on its properties under repeated sampling such as mean-squared error for estimates, power for tests, expected size of confidence sets, etc.

- repeated sampling means considering data sets $x_1, x_2, \ldots$ i.i.d. f_θ and the average performance of the procedure for each $\theta \in \Theta$

- so if one procedure does better with respect to a particular repeated sampling criterion than another, uniformly in θ, then it is preferred

- there is currently no frequentist theory that produces answers to **E** and **H** for many meaningful problems and, in some instances, the answers provided are somewhat questionable

- the criteria used to judge a procedure are typically loss-based and loss functions (optimality criteria) need to be chosen and are not falsifiable via the data which is contrary to the goal of objectivity

- for example, in an estimation problem should we use squared error, absolute error or something else?

- often the choice is based on mathematical convenience and convention

Birnbaum, A. (1962) On the foundations of statistical inference. JASA, 57, 298, 269-306.

- attempted to characterize what are good frequentist procedures based on commonly used, partial characterizations of statistical evidence and produced a surprising result

- there are two basic principles of frequentism which most accept as sensible: the sufficiency **S** and the conditionality **C** principles

- furthermore, there is the non-frequentist likelihood principle **L**

- Birnbaum apparently proved that, if you accept **S** and **C,** then you must accept **L**

- this is paradoxical because **S** and **C** allow for frequentism but **L** doesn't

- Bayesianism conforms to **L,** so Birnbaum's Theorem is sometimes cited as support for Bayesian inference

- we examine this result more closely

Evans, M. (2013) What does the proof of Birnbaum's theorem prove? Electronic J. of Statistics, 7, 2645-2655.

- wlog we simplify to the context where $\mathcal{X}$ is finite

- let $\mathcal{I}_\Theta =$ denote the set of all inference bases based on such $\mathcal{X}$ with fixed Θ (easily generalized to allow for reparameterizations)

- a *relation* R on a set $\mathcal{I}$ is a subset of $\mathcal{I} \times \mathcal{I}$ so, if $(I_1, I_2) \in R$, then I_1 and I_2 are related

- a relation R on $\mathcal{I}$ is an *equivalence relation* if it satisfies
(i) (reflexive) $(I, I) \in R$ for all $I \in \mathcal{I}_\Theta$
(ii) (symmetric) if $(I_1, I_2) \in R$ then $(I_2, I_1) \in R$
(iii) (transitive) if $(I_1, I_2) \in R$ and $(I_2, I_3) \in R$ then $(I_1, I_3) \in R$

- an eq. rel. on $\mathcal{I}$ partitions $\mathcal{I}$ into equivalence classes

- a *statistical principle* is a relation on $\mathcal{I}_\Theta$ such that two related inference bases contain the same amount of evidence concerning the true value of θ and so inferences should be the same

- to be a valid characterization of evidence the principle should be an equivalence relation

- if a relation R on $\mathcal{I}$ is not an eq .rel., various equivalence relations can be obtained from it

- let $\mathcal{R}_* = \{R_* : R_* \subset R, R_*$ is an eq. rel. and if $R_* \subset R_{**} \subset R$ with R_{**} an eq. rel. then $R_* = R_{**}\}$ and since the intersection of eq. rel.'s on $\mathcal{I}$ is an eq. rel. then $R_{lam} = \cap_{R_* \in \mathcal{R}} R_*$ is an eq. rel. called the *laminal eq. rel. induced by* R (the biggest eq. rel. within R consistent with all the others)

- also, let $\mathcal{R}^* = \{R^* : R \subset R^*, R^*$ is an eq. rel.$\}$ and define $\bar{R} = \cap_{R^* \in \mathcal{R}} R^*$ the smallest eq. rel. containing R

Lemma (*chaining*) If R is a reflexive relation on $\mathcal{I}$, then $\bar{R} = \{((I, I') : \exists n$ and $I_1, \ldots, I_n \in \mathcal{I}$ s.t. $I_1 = I, I_n = I'$ and $(I_i, I_{i+1}) \in R$ or $(I_{i+1}, I_i) \in R\}$.

- do we have to accept the elements of $\bar{R}$ as equivalent?

Example

- $\mathcal{I} = \{2, 3, 4, \ldots\}$ and $(i, j) \in R$ when i and j have a common factor bigger than 1 so reflexive and symmetric but $(6, 3) \in R$ and $(2, 6) \in R$ yet $(2, 3) \notin R$ so not transitive

- and $\bar{R} = \mathcal{I} \times \mathcal{I}$ since for any (i, j), then $(i, ij) \in R$ and $(ij, j) \in R$ and $\bar{R}$ expresses nothing meaningful

likelihood principle

> *Likelihood Principle* (**L**)
> $(I_1, I_2) \in \mathbf{L}$ *whenever the likelihood function based on* I_1 *equals the likelihood function based on* I_2.

- the likelihood function is any positive multiple of the density at the observed data considered as a function of θ, immediately gives

Lemma L is an eq. rel. on $\mathcal{I}_\Theta$

- so **L** is a potentially valid characterization of statistical evidence but

Example *Irrelevancy of stopping rules.*

- $x \sim$ binomial$(n, \theta), \theta \in (0, 1]$ observe $x = k$, gives
$L(\theta \,|\, x) = \theta^k (1 - \theta)^{n-k}$ (sample for n tosses)

- $y \sim$ negative-binomial$(k, \theta), \theta \in (0, 1]$ and observe $y = n - k$ so
$L(\theta \,|\, y) = \theta^k (1 - \theta)^{n-k}$ (sample until k heads)

- should inferences be the same?

sufficiency principle

- recall that, for model $\{f_\theta : \theta \in \Theta\}$, a statistic T (any function defined on $\mathcal{X}$) is sufficient if the conditional distribution of the data x given the value $T(x)$ is independent of θ, T is minimal sufficient if for any sufficient statistic T' there is a function $h_{T,T'}$ such that $T(x) = h_{T,T'}(T'(x))$ and obviously a 1-1 function of a mss is a mss

- let $[x] = \{z \in \mathcal{X} : f_\theta(x) = cf_\theta(z)$ for some $c > 0$ and every $\theta \in \Theta\}$ so $[x]$ is the eq. class containing x induced by the eq. rel. on $\mathcal{X}$ that says two data sets are equivalent if they give rise to the same likelihood function

Lemma $[\cdot]$ is a minimal sufficient statistic for $\{f_\theta : \theta \in \Theta\}$.

> *Sufficiency Principle (**S**)*
> *If T_i is a mss for the model of $I_i = (\{f_{i\theta} : \theta \in \Theta\}, x_i)$ for $i = 1, 2$ and there is a 1-1 function h such that $T_1 = h(T_2)$ with $T_1(x_1) = h(T_2(x_2))$, then $(I_1, I_2) \in$ **S**.*

- the underlying idea is that, because the conditional distribution given a sufficient statistic does not involve θ, reducing the data to the value of the sufficient statistic, so the information locating x within

$$T^{-1}\{T(x)\} = \{z : T(z) = T(x)\}$$

is discarded, does not lose any evidence concerning the true value of θ and we want to make the maximum reduction in the data to the value of a mss

Lemma S is an eq. rel. on $\mathcal{I}_\Theta$ and **S** $\subset$ **L**.
Proof: The eq. rel. part is obvious. If $(I_1, I_2) \in \mathbf{S}$, then by the factorization theorem $f_{i\theta}(x_i) = k(x_i)g_{T_i\theta}(T_i(x_i))$ where $g_{T_i\theta}$ is the density of the mss T_i for $\{f_{i\theta} : \theta \in \Theta\}$. Also, $g_{T_1\theta}(T_1(x_1)) = g_{T_2\theta}(h(T_2(x_2)))$ so $f_{1\theta}(x_1) = cg_{T_2\theta}(h(T_2(x_2))) = c'f_{2\theta}(x_2)$ which implies $(I_1, I_2) \in \mathbf{L}$.

- so **S** is a potentially valid characterization of statistical evidence

conditionality principle

Example *Two measuring instruments.*

- a physicist wants to measure a voltage and picks up a voltmeter

- there are two voltmeters available and, based on experience, it is known that a measurement from voltmeter 1 gives values distributed $N(\mu, \sigma_1^2)$ and voltmeter 2 gives values distributed $N(\mu, \sigma_2^2)$ where μ is the unknown voltage and $\sigma_1^2 >> \sigma_2^2$ are both known

- the stores manager tosses a fair coin giving the physicist voltmeter 1 if heads is obtained and voltmeter 2 otherwise and suppose voltmeter 2 is provided with the physicist knowing this

- voltages $x = (x_1, \ldots, x_n)$ were obtained and $\bar{x}$ is the estimate but how to quantify the accuracy of this estimate, namely, the conditional, given the voltmeter used, 0.95-CI $\bar{x} \pm (\sigma_2/\sqrt{n})z_{0.025}$ or the longer unconditional (approx.) 0.95-CI $\bar{x} \pm (\sqrt{(\sigma_1^2 + \sigma_2^2)/2n})z_{0.025}$

- most would say the conditional interval is the right one

- note - the distribution of the choice of the voltmeter does not involve the unknown μ

- a statistic U is *ancillary* for the model $\{f_\theta : \theta \in \Theta\}$ if the distribution of $U(x)$ is independent of θ

> *Conditionality Principle* (**C**) *If U is an ancillary for the model in $I = (\{f_\theta : \theta \in \Theta\}, x)$, then $(I, I_U) \in \mathbf{C}$ and $(I_U, I) \in \mathbf{C}$ where $I_U = (\{f_\theta(\cdot \mid U(x)) : \theta \in \Theta\}, x)$ and $f_\theta(\cdot \mid U(x))$ is the conditional density of the data given $U(x)$.*

- the basic idea is that we want to remove all variation that does not depend on θ so appropriate accuracy assessments can be made

Lemma C is reflexive and symmetric but not transitive and $\mathbf{C} \subset \mathbf{L}$.

- so **C** is not a proper characterization of statistical evidence

- the basic idea to the proof is that there can be many ancillaries for a model but if U_1 and U_2 are ancillaries it is not the case in general that (U_1, U_2) is ancillary

- in particular there is no *maximal ancillary* U (every other ancillary can be written as a function of U)

Birnbaum's Theorem If you accept **S** and **C** as proper characterizations of statistical evidence, then you must accept **L** as a proper characterization of statistical evidence and frequentism is not relevant.

Proof: Suppose that $(I_1, I_2) \in \mathbf{L}$. Construct a new inference base $I = (M, y)$ from I_1 and I_2 as follows. Let M be given by $\mathcal{X}_M = (\{1\} \times \mathcal{X}_{M_1}) \cup (\{2\} \times \mathcal{X}_{M_2})$,

$$f_{M,\theta}(1,x) = \begin{cases} (1/2) f_{M_1,\theta}(x) & \text{when } x \in \mathcal{X}_{M_1} \\ 0 & \text{otherwise,} \end{cases}$$

$$f_{M,\theta}(2,x) = \begin{cases} (1/2) f_{M_2,\theta}(x) & \text{when } x \in \mathcal{X}_{M_2} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$T(i,x) = \begin{cases} (i,x) & \text{when } x \notin \{x_1, x_2\} \\ \{x_1, x_2\} & \text{otherwise} \end{cases}$$

is sufficient for M and so $((M,(1,x_1)),(M,(2,x_2))) \in \mathbf{S}$. Also, $U(i,x) = i$ is ancillary for M and thus

$$((M,(1,x_1)),(M_1,x_1)) \in C, ((M,(2,x_2)),(M_2,x_2)) \in C.$$

This completes the "proof".

- but what this actually proves, using the chaining argument, is the following

Lemma $\overline{\mathbf{S} \cup \mathbf{C}} = \mathbf{L}$

- namely, the smallest eq. rel. containing $\mathbf{S} \cup \mathbf{C}$ is $\mathbf{L}$ (and note $\mathbf{S} \cup \mathbf{C} \subset \mathbf{L}$ is not an eq. rel.)

- so we do not have to accept the additional equivalences induced in $\mathbf{S} \cup \mathbf{C}$

- Evans, Fraser and Monette (1986) prove

Lemma $\overline{\mathbf{C}} = \mathbf{L}$.

- $\mathbf{C}$ is a significant problem for frequentism, can it be resolved? mostly just ignored

- note $\mathbf{C}$ is not a problem for Bayes because in that formulation we condition on all the data, not just ancillaries

- also ancillary statistics have a role to play in model checking and checking for prior-data conflict