The Measurement of Statistical Evidence
Lecture 1 - part 2

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

2021

I.3 The Ingredients

- we have data x from measurement X and want to answer **E** and/or **H** concerning Ψ
- how?
- we need a theory that allows us to reason from the data to answers, now called inferences, to **E** and/or **H**
- if we could do this simply from x that would be ideal but ....
- so it would appear that we need more ingredients
- these ingredients are typically not determined by the application but rather are chosen by the statistician and so are subjective
- of course, we want the minimal number of ingredients
- what criteria should these ingredients satisfy?

bias

- as noted, the chosen ingredients are subjective and this is good, as it allows for informed choices, but it also seems contrary to the ideals of science
- can these ingredients be chosen, before seeing the data, in such a way, that a desired answer can be produced (a foregone conclusion)? Yes (at least with high probability)
- this will be called *bias* here and we will subsequently discuss how to measure and control this
- so this concern about subjectivity can be dealt with

falsifiability

- what establishes the relevance, or lack of relevance, of the chosen ingredients to the application
- if the data x is chosen correctly, then the data is considered *objective*
- the following principle is then invoked
- *Principle of Empirical Criticism*: each chosen ingredient must be checked against the observed x as to its relevance
- this is basically an application of Popper's idea that a theory is only scientific if it can be falsified by empirical data
- this rules out some ingredients, e.g., how do you falsify a loss/utility function?
- losses/utilities may play a role in decisions but one can still state the inferences that are evidence based only and, if the decision contradicts this inference, one need only justify why
- lots of paradoxes involving utility so good to keep these out of the story

Example *The Allais Paradox*

- *Sure Thing Principle*: if you prefer option A to B and you are presented with the choice of $\{A \text{ and } C\}$ or $\{B \text{ and } C\}$ you will choose $\{A \text{ and } C\}$

- suppose presented with two contexts where you choose (a) or (b)

> *Context 1 (a) You receive* $\$10^6$ *with probability* 1.00. *(b) You receive* $\$10^6$ *with probability* 0.89, *receive* $\$2 \times 10^6$ *with probability* 0.10 *and receive nothing with probability* 0.01.

- Allais claims that most people would choose (a) to avoid the small chance of getting nothing

> *Context 2 (a) You receive* $\$10^6$ *with probability* 0.11 *and nothing with probability* 0.89. *(b) You receive* $\$2 \times 10^6$ *with probability* 0.10 *and nothing with probability* 0.90.

- Allais claims that most people would choose (b) because the difference in the chance of receiving nothing is small and the payoff is greater with (b)

- the paradox: by the principle you should prefer (b) over (a) in Context 1 (C = You receive $\$10^6$ with probability 0.89)

Ingredient 1: the statistical model

- $\{f_\theta : \theta \in \Theta\}$ a collection of prob. distributions and it is believed (assumed) that $f_X \in \{f_\theta : \theta \in \Theta\}$

- here we will also require that the model parameter θ indexes the possible distributions

- then there is $\theta_{true} \in \Theta$ such that $f_X = f_{\theta_{true}}$

- also for the object of interest $\Psi : \Theta \overset{onto}{\rightarrow} \Psi$ (overload the notation) so each possible distribution gives a potentially different value $\psi = \Psi(\theta)$ and the true value is $\psi_{true} = \Psi(\theta_{true})$

- note - Ψ is a real-world object not just some mathematical construct

- the statistical model is falsifiable via model checking where we are asking: is the observed x a reasonable value from at least one of the distributions in $\{f_\theta : \theta \in \Theta\}$?

- **note** - the statistical model is generally obviously wrong because it is unrealistic to assume $f_X \in \{f_\theta : \theta \in \Theta\}$ and when checking the model we can never say the model is correct only that we have not obtained any indication that it is substantially incorrect
- this brings up an important point about any of the ingredients specified:

> *The purpose of the ingredients is to allow us to build a sound theory that allows us to reason to answer* **E** *and/or* **H**. *They are devices for this purpose and, while we care that their incorrectness may lead us to make incorrect inferences, we don't want to obsess about their correctness to the extent that we can't build an appropriate theory.*
>
> **The theory of statistical reasoning is more important than the ingredients.**

Ingredient 2: the prior

- it will be argued that if we want a theory of inference that is based on measuring evidence appropriately, then we need to specify a prior probability distribution Π on Θ
- so for $A \subset \Theta$ the probability that $\theta_{true} \in \Theta$ is given by $\Pi(A) = \int_A \pi(\theta)\, d\theta$
- what does this mean? answer: next lecture
- where does Π come from? it is a choice made by the statistician but it must be based on an "elicitation" (to be discussed)
- Π induces a prior on Ψ, namely, for $B \subset \Psi$ then $\Pi_\Psi(B) = \Pi(\Psi^{-1}B) = \int_B \pi_\Psi(\psi)\, d\psi$
- note - the conditional prior $\Pi(\cdot \mid \psi)$ is a probability distribution concentrated on $\Psi^{-1}\{\psi\}$ and characteristics that identify $\theta \in \Psi^{-1}\{\psi\}$ are known as *nuisance parameters*
- by the multiplication rule $\pi(\theta) = \pi(\theta \mid \psi)\pi_\Psi(\psi)$

- we now have a joint (prior) probability model $(\theta, x) \sim \pi(\theta) f_\theta(x)$
- then when we observe x we invoke the first principle (or axiom) the principle of conditional probability to replace the prior π by the *posterior* of θ (the conditional distribution of θ given x)

$$\pi(\theta \,|\, x) = \frac{\pi(\theta) f_\theta(x)}{m(x)}$$

where

$$m(x) = \int_\Theta \pi(\theta) f_\theta(x)\, d\theta$$

the marginal of x, also called the *prior predictive* of x (the dist. used to make probability statements about x before it is observed)
- suppose $T(x)$ is a (minimal) sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ so $f_\theta(x) = g_\theta(T(x))h(x)$, then $m(x) = h(x) \int_\Theta \pi(\theta) g_\theta(T(x))\, d\theta$ and so

$$\pi(\theta \,|\, x) = \frac{\pi(\theta) f_\theta(x)}{m(x)} = \frac{\pi(\theta) g_\theta(T(x)) h(x)}{h(x) \int_\Theta \pi(\theta) g_\theta(T(x))\, d\theta} = \frac{\pi(\theta) g_\theta(T(x))}{\int_\Theta \pi(\theta) g_\theta(T(x))\, d\theta}$$

so the posterior depends on the data only through the value $T(x)$

- g_θ can be taken to the density of T when θ is the true value and so

$$\pi(\theta \mid x) = \frac{\pi(\theta) g_\theta(T(x))}{m_T(T(x))}$$

where

$$m_T(t) = \int_\Theta \pi(\theta) g_\theta(t)\, d\theta$$

is the prior predictive of T
- can a prior be falsified? Yes
- Evans, M. and Moshonov, H. (2006) Checking for prior-data conflict. Bayesian Analysis, Volume 1, Number 4, 893-914, compute

$$M_T(m_T(t) \leq m_T(T(x))$$

and if this is small $T(x)$ lies in the tails of M_T and is then an indication that there is a problem with the prior
- this check presumes the model is correct, so check the model first and then check the prior

- **the difference** δ that matters
- suppose ψ is a continuous parameter and we wish to assess the hypothesis $H_0 : \Psi(\theta_{true}) = \psi_0$
- we know that we can only detect (absolute or relative) differences of a certain size as expressed via $\delta > 0$ and distance measure d
- so the hypothesis we want to assess is $H_0 : d(\Psi(\theta_{true}), \psi_0) \leq \delta$
- δ is primarily a product of the measurement process and presumably the process is designed in such a way that achieving the accuracy desired is possible

Statistical Reasoning

- here is a sequence of steps to statistical reasoning concerning **E** and/or **H**

1. choose a model $\{f_\theta : \theta \in \Theta\}$
2. choose a prior π
3. measure bias and select the amount of data to collect to avos bias
4. collect the data
5. check the model (modify if necessary)
6. check the prior (modify if necessary)
7. derive the inferences (based on principles of inference to be discussed)

Example *binomial*

- suppose we are interested in making inference about θ the proportion of people infected with COVID-19 in Toronto and in particular want to assess the hypothesis $H_0 : \theta \in [0.02, 0.05]$

- the **model** $x = (x_1, \ldots, x_n) \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta)$ with $\theta \in [0, 1]$ so $T(x) = n\bar{x} = \sum_{i=1}^{n} x_i \sim \text{binomial}(n, \theta)$ is a mss

- the **prior** $\theta \sim \text{beta}(\alpha_0, \beta_0)$ distribution where α_0 and β_0 are specified *hyperparameters* that need to be elicited

- one possible elicitation algorithm: specify an interval (a, b) such that it is known that $\theta \in (a, b)$ with *virtual certainty*, e.g., $\Pi((a, b)) = 0.99$ and pick a point for the mode in (a, b) such as $(a + b)/2$

- this determines α_0 and β_0

- so if $a = 0$ and $b = 0.25$, then $\alpha_0 = 8.86$ and $\beta_0 = 56.05$

- the next step is to measure bias and for this we need to discuss how evidence will be measured

- the posterior of θ is $\text{beta}(\alpha_0 + n\bar{x}, \beta_0 + n(1 - \bar{x}))$

- **bias**
- we need to specify how to measure evidence first
- *principle of evidence*: evidence in favor of H_0 is found if $\Pi([0.02, 0.05] \mid n\bar{x}) > \Pi([0.02, 0.05])$ which occurs iff

$$\text{Beta}([0.02, 0.05], \alpha_0 + n\bar{x}, \beta_0 + n(1 - \bar{x})) > \text{Beta}([0.02, 0.05], \alpha_0, \beta_0)$$

where Beta is the probability measure and evidence against is found if $\Pi([0.02, 0.05] \mid n\bar{x}) < \Pi([0.02, 0.05])$
- there is bias against H_0 when the prior probability of getting evidence not in favor of H_0, when it is true, is large
- so we need to compute

$$\sup_{\theta \in [0.02, 0.05]} M\left(\begin{array}{l} \text{Beta}([0.02, 0.15], \alpha_0 + n\bar{x}, \beta_0 + n(1 - \bar{x})) \leq \\ \text{Beta}([0.02, 0.15], \alpha_0, \beta_0) \end{array} \middle| \, \theta \right)$$

and and note $M(\cdot \mid \theta)$ is the binomial(n, θ) measure

- there is bias in favor of H_0 when the prior probability of getting evidence not against H_0, when it is false, is large
- so we need to compute

$$\sup_{\theta \notin [0.02,0.05]} M\left(\begin{array}{l} \text{Beta}([0.02, 0.15], \alpha_0 + n\bar{x}, \beta_0 + n(1-\bar{x})) \geq \\ \text{Beta}([0.02, 0.15], \alpha_0, \beta_0) \end{array} \middle| \, \theta \right)$$

- determine n so that both biases are small as biases $\longrightarrow 0$ as $n \longrightarrow \infty$
- **model checking**
- compute a test statistic $S(x_1, \ldots, x_n)$ and compare the observed value with its conditional distribution given $T(x_1, \ldots, x_n)$
- runs tests

- **checking for prior-data conflict**

$$m_T(t) = \frac{\Gamma(n+1)}{\Gamma(t+1)\Gamma(n-t+1)} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(t+\alpha_0)\Gamma(n-t+\beta_0)}{\Gamma(n+\alpha_0+\beta_0)}$$

so compute

$$M_T\left(\frac{\Gamma(t+\alpha_0)\Gamma(n-t+\beta_0)}{\Gamma(t+1)\Gamma(n-t+1)} \leq \frac{\Gamma(T(x)+\alpha_0)\Gamma(n-T(x)+\beta_0)}{\Gamma(T(x)+1)\Gamma(n-T(x)+1)}\right)$$

and this is easily computed by generating $\theta \sim \text{beta}(\alpha_0, \beta_0)$ and then $t \sim \text{binomial}(n, \theta)$ many times

- **inference**

- there is evidence in favor H_0 if $\Pi([0.02, 0.05] \mid n\bar{x}) > \Pi([0.02, 0.05])$

- and evidence against H_0 if $\Pi([0.02, 0.05] \mid n\bar{x}) < \Pi([0.02, 0.05])$

- what about the strength of this evidence? record the posterior probability $\Pi([0.02, 0.05] \mid n\bar{x})$

- if there is evidence in favor and this posterior prob. is large, there is strong evidence in favor while if there is evidence against and this posterior prob. is small, there is strong evidence against