Theory of Statistical Inference - Lecture I
STA422 and STA2162

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

2026

I.1 The Problem

- a theory of statistical inference is (should be?) a significant component of the language of science

- why?, there are scientific questions we want to know the answer to which can't be known categorically

- consider a real world object or concept Ψ

$\Psi =$ the half-life of a neutron

$\Psi =$ the median annual income of a student at U of T

$\Psi =$ the current rate of increase in mean annual global temperature

$\Psi =$ a measure of the relationship between the consumption of alcohol and the fat content of the liver

$\Psi =$ a graph describing the influences of some variables on each other

$\Psi =$ the closing price of a particular stock on the third Friday of a month

$\vdots$

- two basic questions that a scientist is concerned with re Ψ

(1) **E** - *estimation - what value does* Ψ *take?*
(2) **H** - *hypothesis assessment - does* Ψ *take the value* ψ_0?

- how to go about answering these questions?

- a scientist conducts an experiment n independent times which produces data

$$x = (x_1, x_2, \ldots, x_n)$$

- e.g. n measurements of the half-life of a neutron

- the scientist believes that the experiment will produce data that in some way reflects the value of Ψ

- for a variety of reasons the data $x_1, x_2, \ldots, x_n$ varies and this typically means that the answers to **E** and **H** cannot be definitive

- what to do? two broad approaches based on the data x

(1) the **evidential approach** (Fisher) - x contains evidence concerning the answers to **E** or **H** and the goal is to produce answers that accurately reflect this evidence and its quality

(2) the **behavioristic (decision-theoretic) approach** (Neyman) - the goal is to minimize error where error in the answers to **E** or **H** is measured in some fashion (often counterfactually through repeated performances)

E both have the goal of producing an estimate $\psi(x)$ but the evidential approach wants an estimate together with a measure of its accuracy, while the behavioristic approach wants an optimal estimate with respect to the error criterion chosen

H the evidential approach has the goal of asserting either evidence against or in favor of ψ_0, together with an assessment of the strength of this evidence, while the behavioristic approach either optimally accepts or rejects ψ_0

- often these two approaches are somewhat confounded with no clear justification for doing so

The Problem of Statistical Inference: produce a theory (whether evidential or behavioristic) that will always produce satisfactory answers to **E** *and* **H** **for any** Ψ.

- a number of solutions have been proposed which we'll discuss here

- do any succeed?

- basic principle: a potential theory needs to be based on a consistent idea and if a theory produces clearly bad answers to reasonable problems, or perhaps even no answer at all, then that theory is not a solution to this problem

- is that a problem?

My Answer: to be a major, positive part of the scientific enterprise, the subject of Statistics needs to address this issue and offer a sound reasoning process (our real goal) to answer **E** and **H**

- the discussion here involves a degree of idealization, e.g., the data are always collected correctly (meaning later), so caveats may be in order in particular applications, but we want a solid core

I.2 Measurement and Design

- recall that the experiment is designed so that the data x in some way reflects the value of Ψ

- a poorly designed experiment may not do this well

- e.g. a sample of students at U of T is drawn but only from one class

- the **design** of the experiment is important

- what does it mean for an experiment to be well-designed?

- a subject in itself but a few things to note about design for this course

Measurements

- the data arises as the result of taking measurements, the scientist chooses what to measure and the *measurement accuracy* of each x_i

- all measurements are discrete (made to finite accuracy) and there is an upper limit on the number that can be taken

- so continuity and infinity are idealizations that may lead to convenient approximations but ....

Sample size n

- for a variety of reasons, the data values vary

- we will assume that n is under our control so we can control the *statistical accuracy* of the answers to **E** and **H**

- if n cannot be controlled, then that is a defect of the experiment, **not** the theory

- any theory needs to be clear about when a particular application doesn't measure up

I.3 Ingredients

I.3.1 The Basic Inference Base

- if we could devise a satisfactory theory of inference based only on the data, that would be ideal but this does not "seem" possible

- the primary candidates for theories of inference all contain some or all of the following ingredients which must be specified by the statistician

- a theory is then applied to the ingredients to produce answers to **E** and **H**, the inferences

- it is *assumed* that the data $x \in \mathcal{X}$ (the *sample space*) can be described as arising from a (true) probability distribution in a set, called the *model*, given by

$$\{f_\theta : \theta \in \Theta\}$$

where, for each $\theta \in \Theta$, f_θ is a probability density on $\mathcal{X}$ wrt some support measure ν so

$$P_\theta(B) = \int_B f_\theta(z)\,\nu_{\mathcal{X}}(dz) = \text{ probability unobserved } x \in B \subset \mathcal{X}$$

- θ is the *model parameter* and Θ is the *model parameter space*

- it is assumed that θ indexes, namely, $f_{\theta_1} \neq f_{\theta_2}$ whenever $\theta_1 \neq \theta_2$ (no nonidentifiability)

- interest is in inference about $\psi = \Psi(\theta) \in \Psi(\Theta) =$ set of possible values of Ψ

- **note** - ψ corresponds to something in the real world typically a characteristic of f_θ

- $\Psi^{-1}\{\psi\}$ may not be singleton for any ψ

    All models are wrong (f_θ is not the true distribution of x for any $\theta \in \Theta$) but it is required that $\psi_{true} \in \Psi(\Theta)$.

- recall the goal is inference about Ψ and not necessarily identifying the true distribution

- the model $\{f_\theta : \theta \in \Theta\}$ is a device to further inference

- a valid question, however, is whether or not the model is so wrong that our inferences about Ψ are badly affected by this

- partly this can be answered through *model checking* (later)

- the model is a *subjective* choice but model checking involves seeing if our choice is contradicted by the *objective*, if collected correctly, data

- with enough data, it will be concluded that the model is wrong, so the real goal is to see if our choice renders inferences about Ψ substantially in error

- in general, checking any choices made against the data is at least a partial response to the criticism of subjectivity

Example *neutrino mass*

- $\Psi =$ mass of a certain kind of neutrino $\in [0, \infty)$

- since mass measurements are nonnegative, physicist assumes single measurement is coming from a distribution in

$$\{\text{gamma}_{rate}(\alpha, \beta) : \theta = (\alpha, \beta) \in \Theta = [1, \infty) \times [0, \infty)\} \text{ where}$$

$$f_\theta(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z} \text{ for } z > 0$$

and $\psi = \Psi(\theta) = \Psi(\alpha, \beta) = (\alpha - 1)/\beta =$ the mode of the distribution (does this representation make sense?, why not use the mean $\psi = \Psi(\theta) = \Psi(\alpha, \beta) = \alpha/\beta$?)

- multiple measurements are treated as an iid sample from this model

- interest is in estimating ψ and assessing whether or not $H_0 : \psi_{true} = 0$ is true or false

- note - $\psi = 0$ iff $\alpha = 1$ iff $x \overset{iid}{\sim} \text{exponential}_{rate}(\beta)$

- so indeed the true mass is captured by the model (through ψ) as well as the accuracy of the measurement process (through $\sigma^2 = \alpha/\beta^2 \in [0, \infty)$)

- but is there any reason to assume a gamma distribution for the measurement process?

- there will also be a $\Delta =$ *the difference that matters*

- since we are measuring the mass of each observed neutrino to finite accuracy we will not get exact 0's for the measurements but rather, if the true mass is in $[0, \Delta)$, then we can conclude that there is evidence in favor of the mass being 0

- so really want to assess $H_0 : \psi_{true} \in [0, \Delta]$ ■

- we call $I = (\{f_\theta : \theta \in \Theta\}, x) =$ the *basic inference base*

- probably should consider Δ as part of this too as it is part of the design in that it bears on what is a suitable n

- to the basic inference base add a *loss function*
$Loss : \Theta \times \Psi(\Theta) \to [0, \infty)$ where $Loss(\theta, \psi) = 0$ iff $\psi = \Psi(\theta)$

- e.g. $Loss(\theta, \psi) = (\Psi(\theta) - \psi)^2$ (squared-error loss) or
$Loss(\theta, \psi) = |\Psi(\theta) - \psi|$ (absolute error loss)

- then a statistical procedure $d(x) \in \Psi(\Theta)$, called a *decision function* here, is considered wrt the expected losses it leads to

$$\begin{aligned} R(\theta, d) &= E_\theta(Loss(\theta, d)) = \int_{\mathcal{X}} Loss(\theta, d(x)) f_\theta(x) \nu_{\mathcal{X}}(dx) \\ &= \textit{risk function} \text{ of } d \text{ (fix } d \text{ and vary } \theta) \end{aligned}$$

- d_1 is preferred to d_2 whenever $R(\theta, d_1) \leq R(\theta, d_2)$ for every $\theta \in \Theta$ and the inequality is strict for at least one θ

- is there an optimal d?, why expected loss?

- *Loss* is a subjective choice (often for convenience, as in squared-error loss) and, in general, there is no methodology for checking it against the data

- call $I_{Loss} = (\{f_\theta : \theta \in \Theta\}, Loss, x)$ the *decision-theoretic inference base*

- to the basic inference base we add a *prior* $\pi =$ a probability density on Θ wrt support measure ν_Θ that reflects beliefs concerning the true value of θ

$$\Pi(A) = \int_A \pi(\theta)\,\nu_\Theta(d\theta) = \text{ probability true value of } \theta \in A$$

measures our initial belief that the true value of θ is in A

- so the *Bayesian inference base* $I_{Bayes} = (\pi, \{f_\theta : \theta \in \Theta\}, x)$

- here f_θ is the conditional density of x given θ

- the prior and the model imply a joint distribution $(\theta, x) \sim \pi(\theta) f_\theta(x)$ so, before seeing x,

$$P((\theta, x) \in A \times B) = \int_A \int_B \pi(\theta) f_\theta(x)\,\nu_{\mathcal{X}}(dx)\,\nu_\Theta(d\theta)$$

- once x is observed we invoke the

Principle of Conditional Probability: if $P(A)$ is the initial probability assigned to event A and event C is observed to be true where $P(C) > 0$, then our belief in the truth of A is now given by $P(A \mid C) = \frac{P(A \cap C)}{P(C)}$, the conditional probability of A given that C is true.

- this leads to the *posterior belief* that the true value of θ is in A given by

$$\Pi(A \mid x) = \int_A \pi(\theta \mid x)\, \nu_\Theta(d\theta)$$

where

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

is the *posterior density* of θ (the conditional density of θ given x) and

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

is the *prior density* of x called the *prior predictive density* of x

note - the "Principle of Conditional Probability" is an axiom of statistical inference not probability theory

- when interest is in $\psi = \Psi(\theta)$ we have the marginal prior

$$\pi_\Psi(\psi) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta)$$

and the marginal posterior

$$\pi_\Psi(\psi \mid x) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta \mid x)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta)$$

Exercise 1. (Assume all measures are discrete) Show that $I_{Bayes} = (\pi, \{f_\theta : \theta \in \Theta\}, x)$ leads to the same posterior for ψ as

$$I_{\Psi,Bayes} = (\pi_\Psi, \{f_\psi : \psi \in \Psi(\Theta)\}, x)$$

where

$$f_\psi(x) = \int_{\Psi^{-1}\{\psi\}} f_\theta(x)\pi(\theta \mid \psi)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta).$$

- this is a nice consistency property and it suggests that "integrating out the nuisance parameters" to obtain f_ψ is well-justified

- how do we choose π? elicitation (later) and note the same concern arises with the choice of the model $\{f_\theta : \theta \in \Theta\}$

- can the prior π be checked against the data as to its suitability? checking for prior-data conflict (later)

Improper Priors and Empirical Bayes

- it is common to see a prior π used that is *improper*, namely, $\pi(\theta) \geq 0$ for all θ, but $\int_\Theta \pi(\theta)\,\nu_\Theta(d\theta) = \infty$, e.g. $\pi(\theta) \propto 1$ on $\Theta = \mathbb{R}^1$

- so, in such a case π is not a probability density and so does not represents beliefs but then quite often $\pi(\theta \,|\, x)$, as defined above, satisfies

$$\int_\Theta \pi(\theta \,|\, x)\,\nu_\Theta(d\theta) = 1$$

(namely, $m(x) = \int_\Theta \pi(\theta) f_\theta(x)\,\nu_\Theta(d\theta) < \infty$ is a valid normalizing constant, not a probability density) so **formally** $\pi(\theta \,|\, x)$ is a probability density

- what then justifies the use of the formal posterior $\pi(\theta \,|\, x)$ to describe beliefs as it isn't by the Principle of Conditional Probability?

- similarly, the theory of *empirical Bayes*, which chooses the prior from a family $\{\pi_\tau : \tau \in \Upsilon\}$ using the data x, does not satisfy the Principle of Conditional Probability

- this takes the decision theory inference base and adds a prior and we have the *Bayesian decision theory inference base*

$$I_{Bayes,Loss} = (\pi, \{f_\theta : \theta \in \Theta\}, Loss, x)$$

- this leads to the *prior risk* for decision function d given by

$$\begin{aligned} r(d) &= \int_\Theta R(\theta, d)\pi(\theta)\,\nu_\Theta(d\theta) \\ &= \int_\Theta \int_\mathcal{X} Loss(\theta, d(x))f_\theta(x)\pi(\theta)\,\nu_\mathcal{X}(dx)\,\nu_\Theta(d\theta) \\ &= \int_\mathcal{X} \int_\Theta Loss(\theta, d(x))\pi(\theta \mid x)\,\nu_\Theta(d\theta)\,m(x)\nu_\mathcal{X}(dx) \\ &= \int_\mathcal{X} r(d \mid x)\,m(x)\nu_\mathcal{X}(dx) \end{aligned}$$

where $r(d \mid x) = \int_\Theta Loss(\theta, d(x))\pi(\theta \mid x)\,\nu_\Theta(d\theta)$ is the *posterior risk*

- if $r(d) \leq r(d')$ for all decision functions d', then d is called a *Bayes rule*

- if $d(x)$ minimizes $r(d' \mid x)$ for each x, then clearly d is a Bayes rule

- again the loss function *Loss* cannot generally be checked against the data as to its suitability

- a basic scientific principle

> *All ingredients to a statistical analysis need to be checked against the data as to their suitability.*

- so when an analysis contains ingredients that can't be checked against the data it is not considered as appropriate for an objective analysis