The Measurement of Statistical Evidence
Lecture 1 - part 1

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

2021

I.1 Introduction (Chapter 1)

- is there a theory of statistical reasoning that is recognized as the core of the subject?

  No

- does that matter?

  Yes

- Why does it matter?

  Without a core there is no subject but rather a collection of "methods" (and opinions) and this doesn't inspire confidence.

- statistical reasoning is intrinsic to many branches of science and it is our job to provide users with a logically sound approach

- if we don't ....

- various theories or approaches have been proposed and are used

- sometimes (as we will see) different theories can lead to diametrically opposed conclusions to a given statistical problem

- so one, or both, of these theories isn't correct

- basic idea that I follow ("meaningful" to be clarified)

> *If a theory of statistical reasoning can be shown to behave inappropriately in a meaningful statistical context or, if the theory fails to provide a solution to a meaningful statistical problem, then it must be modified or discarded.*

- is it possible that there is no theory that is based on simple ideas and that leads to solutions that can be broadly accepted as based on correct statistical reasoning?

> *I'm an optimist and will make a proposal as part of the course.*

- does "big data" allows us to avoid these issues?

> *Certainly you can have so much data that statistical reasoning becomes irrelevant (e.g. effectively a census has been carried out) but in general the answer is no as the same issues arise due to uncertainty.*

- so the course is about the foundations of statistical reasoning

- we will look at the various theories that have been proposed and see where the problems lie

> *The central thesis here is that any such theory must be based on a clear prescription of what we mean by statistical evidence and it is the failure of most theories to do this in a satisfactory way that leads to current problems.*

I.2 Statistical Problems

- in a scientific context there are questions concerning an object of interest Ψ and data x has been collected believed to contain **evidence** concerning the answers

E **estimation** - provide an estimate $\psi(x)$ of Ψ together with an assessment of its accuracy based on the evidence

H **hypothesis assessment** - quote the evidence *in favor of* or against some specified value ψ_0 of Ψ *together with an assessment of the strength of the evidence*

- what is the basic context where such problems arise and that are statistical in character?

Example (the Archetypal Example)

- $\Omega =$ a population (finite)

- $X : \Omega \to \mathcal{X}$ a *measurement*

- so $X(\omega) \in \mathcal{X}$ is the value of the measurement for population member ω and this measurement is always taken to some *finite* accuracy

- this gives rise to the distribution of X over Ω, namely, for $x \in \mathcal{X}$

$$f_X(x) = \frac{\#(\{\omega : X(\omega) = x\})}{\#(\Omega)} = \text{ the proportion of members of } \Omega \text{ with measurement value } x$$

and Ψ can be expressed as some aspect of f_X

- in principle f_X can be known, as well as Ψ, by counting if we do a census

- in general we can't carry out a census

- for example, suppose $\Omega =$ students at U of T and $X(\omega) =$ ht in cm, so $\mathcal{X}$ is a finite set of rational numbers, f_X gives the distribution of ht over Ω and suppose $\Psi =$ (1st quartile, median, 3rd quartile) and, even with a census, we can only know these quantities to the nearest cm

- for example, suppose, $X = (X_1, X_2) = (\text{blood type, blood pressure at rest in mm Hg})$ and we want to know if X_1 and X_2 are related and, if so, how they are related

- define the conditional distributions $X_1 \mid X_2$

$$\begin{aligned} f_{X_1|X_2}(x_1 \mid x_2) &= \frac{\#(\{\omega : X_1(\omega) = x_1, X_2(\omega) = x_2\})}{\#(\{\omega : X_2(\omega) = x_2\})} \\ &= \text{the proportion of the members of } \Omega \text{ having} \\ &\quad X_2 \text{ measurement } x_2 \text{ who have } X_1 \text{ measurement } x_1 \end{aligned}$$

- so Ψ is the collection of all these conditional distributions as then X_1 and X_2 are related if $f_{X_1|X_2}(\cdot \mid x_2)$ (meaningfully) changes as x_2 changes and the form of the relationship is how these distributions change

- the data $x = (x_1, \ldots, x_n) = (X(\omega_1), \ldots, X(\omega_n))$ for some $\{\omega_1, \ldots, \omega_n\} \subset \Omega$

- how to select $\{\omega_1, \ldots, \omega_n\}$? "randomly" and through design

- then the distributions become probability distributions and, if the predictor is controlled, relationships become causative

note - we will always assume here that the data has been collected correctly even though in applications it seems it often isn't

- when the data isn't collected properly a caveat such as "the inferences may not apply to the population of interest Ω" is required

- in building a theory we will restrict to the ideal circumstances

note - everything is finite as measurements are bounded and made to a finite accuracy

- what about infinity?

Example

- suppose $\#(\Omega)$ is big and X is real-valued taking many distinct values

- let f be a probability density function that satisfies, for each $A \subset \mathcal{X}$,

$$\int_A f(x)\,dx \approx \sum_{x \in A} f_X(x)$$

- so we are using infinite objects to approximate something that is finite

- so it is fine to take $\mathcal{X}$ to be an infinite set and use continuous probability distributions just don't forget that you are approximating something that is finite and don't treat the infinite object as the truth

Example *Fisher's counterexample to Bayesian inference*
- suppose $X(\omega) = 1$ if ω is male and is 0 otherwise

- let θ denote the proportion of males in Ω, so $\theta \in \{0, 1/\#(\Omega), 2/\#(\Omega), \ldots, 1\}$

- suppose we know nothing more about θ and so a uniform prior is placed on θ so $1/(\#(\Omega)+1) =$ prior prob. of θ being the true value for each possible value

- if $\#(\Omega)$ is large it makes sense to approximate this by a continuous uniform density $\pi(\theta) = 1$ for $\theta \in [0, 1]$ (the points are equispaced across $[0, 1]$)

- but now suppose instead we want to make inference about $\psi = \Psi(\theta) = \theta^2 \in [0,1]$ a 1-1 function of θ

- ψ has prior density (change of variable) $\pi_\Psi(\psi) = 1/2\psi^{1/2}$ a beta$(1/2,1)$ density and this is far from uniform

- but if our beliefs were uniform about θ shouldn't these also be uniform on ψ?

- Fisher's conclusion was that because π_Ψ is not uniform, then this implies that priors are "rubbish"

- **but** remember that $\theta \in \{0, 1/\#(\Omega), 2/\#(\Omega), \ldots, 1\}$ which implies $\psi \in \{0, (1/\#(\Omega))^2, (2/\#(\Omega))^2, \ldots, 1\}$ and these points are not equispaced across $[0,1]$ (pts to the left of $1/2$ are closer together and points to the right are further apart and all are compressed towards 0)

- what continuous distribution best approximates a discrete uniform distribution on $\{0, (1/\#(\Omega))^2, (2/\#(\Omega))^2, \ldots, 1\}$? π_Ψ

- note too, that if we randomly sample (without replacement) $\{\omega_1, \ldots, \omega_n\}$, then it is reasonable to "approximate" the distribution of $\sum_{i=1}^{n} X(\omega_i)$ as a binomial(n, θ) at least when $n << \#(\Omega)$ and θ isn't too small or too large

Example *Measurement models*

- a "continuous" measurement X with true value Ψ

- often this will be presented as $X = \Psi + e$ where e is an error term following a continuous prob. distribution f and repeated measurements $x_1, \ldots, x_n$ are made

- recall, X is measured to finite accuracy and it is bounded, so Ψ can only be known to that accuracy and the possible values for Ψ are finite

- further to be statistically meaningful, constraints must be *assumed* on f, e.g., if Ψ is the mean of X then f has mean 0

- mathematically, the idea is that, with infinitely many measurements, the distribution of X has density $f_X(x) = f(x - \Psi)$

- there is always an upper bound N on the number of observations that could possibly be taken

- so we can *imagine* a population Ω of N measurements and if $n << N$ it is reasonable to assume the measurements taken are independent and approximate the distribution of X by a continuous one provided X has many distinct values

- this situation isn't as realistic (is it real? where is the "randomness" coming from?) as the archetypal example where everything can be known via counting and where there are simple mechanisms to ensure "randomness" (to be discussed)