Theory of Statistical Inference - Lecture II.1 STA422 and STA2162

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

2026

II. Probability

II.1 What is probability?

- probability model $(\Omega, \mathcal{A}, P)$ for unknown $\omega \in \Omega$

- what is the meaning of $P(A)$ for $A \in \mathcal{A}$?

- clearly $P(A) = 1$ means it is definitive that $\omega \in A$ and $P(A) = 0$ means it is definitive that $\omega \notin A$

- but what about when $0 < P(A) < 1$?

- two broad conceptions

  (i) $P(A)$ represents "our" degree of belief that $\omega \in A$ (subjective)
  (ii) $P(A)$ represents an aspect of reality such as a long-run relative frequency (objective)

- in either case $P(A)$ comes from the choice of $(\Omega, \mathcal{A}, P)$

- so from a formal, mathematical viewpoint we can avoid answering the question but not for applications

- consider the frequentist answer

Example II.1.1 *coin tossing*

- a coin is tossed (without control) 10^6 times and 816,743 heads are obtained

- is it not reasonable to say that our belief that the next coin toss will be a head satisfies $P(\text{"head"}) \approx 0.82$?

- the point being that, even with the frequentist interpretation, probabilities can be taken as representing beliefs ■

- so considering probabilities as degrees of belief is not in itself wrong as it arises from the frequentist formulation as well

- certainly relative frequency make some sense as a way of assigning probability but, as we will see, there are others

- perhaps best to think of probabilities, however assigned, as measuring degrees of belief

- in essence our view is that a probability assignment is primarily useful as part of a reasoning process

Example II.1.2

- does more education decrease the incidence of dementia and, if so, by how much?

- getting an informative, reasonable answer to this question is much more important than getting the "probability" assignments correct ■

- the meaning of probability has been discussed in the literature in a number of different ways

- for $(\Omega, \mathcal{A}, P)$ for response $\omega \in \Omega$ from some system that can be repeatedly performed

- imagine independent performances $\omega_1, \ldots, \omega_n \in \Omega$

- then assign

$$\begin{aligned} P(A) &= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} I_A(\omega_i) \\ &= \lim_{n \to \infty} (\text{proportion of times } \omega_i \in A \text{ in first } n \text{ performances}) \end{aligned}$$

- but more is needed

Example II.2.1

- suppose $\Omega = \{0,1\}$ and we observe data $(0,1,0,1,0,1,\ldots,0,1)$ for every even n and $(0,1,0,1,0,1,\ldots,0)$ for every odd n

- clearly $P(A) = 1/2$, but this makes no sense because when we know, n we know every subsequent outcome ■

Example II.2.2 *Champernowne's sequence*

- here $\Omega = \{0,1,2,3,4,5,6,7,8,9\}$ and we observe

0,1,2,3,4,5,6,7,8,9,1,0,1,1,1,2,1,3,1,4,1,5,1,6,1,7,1,8,1,9,2,0,2,1,. . .

- then

$$\begin{array}{rcl} P(\{i\}) & = & 1/10, \\ P(\{(i,j)\}) & = & 1/100 = P(\{i\})P(\{j\}), \\ P(\{(i,\omega_1,\ldots,\omega_n,j) & : & \omega_i \in \Omega\}) = 1/100 = P(\{i\})P(\{j\}) \end{array}$$

- sequence is iid Uniform(Ω)

- but as soon as we know the first n responses, we know the next ■

- the issue is "randomness" as we want the outcomes to be unpredictable

- if observed long enough, Champernowne's sequence will satisfy every statistical test for being iid Uniform(Ω) but we don't need probability to model this

- von Mises recognized this and tried to formalize when such a sequence could be called "probabilistic"

- he defined *kollectives:*
*(*i) convergence of relative frequencies in the sequence
(ii) if a subsequence is formed via a selection rule that selected the next element based only on the present and past, then the limiting frequencies must remain the same

- Wald proved the existence of kollectives under some restrictions (countable number of selection rules) but Ville showed there are sequences that pass the tests but are quite regular so this approach failed to characterize randomness

- Kolmogorov (1963) characterized randomness but it has nothing to with probability

- see the book Sugita, H. (2017) Probability and Random Number - A First Guide to Randomness for a basic presentation

- consider the set $\Omega_N = \{0,1\}^N$ where N is large and consider what sequences in Ω_N can be considered as random

- consider all programs that can print out a sequence in Ω_N and note that such a program can also be represented as a sequence of 0's and 1's

- for $\omega_N \in \Omega_N$, let q_{ω_N} denote the shortest length $(L(q_{\omega_N}))$ program that produces ω_N

- e.g. $L(\text{"print } \omega_N\text{"}) = n + N$ for some small n so $L(q_{\omega_N}) \leq n + N$

- if $L(q_{\omega_N}) << n + N$, then ω_N is not random and otherwise it is random (vague but intuitively clear)

- $\#\{\omega_N : L(q_{\omega_N}) = k\} \leq 2^k$ (at most 2^k sequences with $L(q_{\omega_N}) = k$ to print a ω_N) so

$$\#\{\omega_N : L(q_{\omega_N}) \leq M\} \leq 2 + 2^2 + \cdots + 2^M = 2^{M+1} - 2$$

- letting $p_{\text{random}}(N, M) =$ proportion of random sequences in Ω_N then

$$p_{\text{random}}(N, M) \geq \frac{2^N - 2^{M+1} + 2}{2^N} = 1 - \left(\frac{1}{2}\right)^{N-M-1} + \left(\frac{1}{2}\right)^{N-1}$$

N	M	$p_{\text{random}}(N, M) \geq$
10^2	$10^2 - 5$	0.93750
10^3	$10^3 - 10$	0.99805
10^4	$10^4 - 15$	0.99994
10^6	$10^6 - 20$	1.00000

- so most sequences are random

- note this "definition" of randomness does not involve probability

- but if our "belief" is that all elements of Ω_N are equally likely, then our belief that we will observe a "random" sequence is virtual certainty for reasonably long sequences

- **fact** - there is no test for randomness (see books on Kolmogorov complexity, etc.)

- my take-way

> *Assigning probabilities using relative frequencies makes good sense when these are available, but there is no reason to restrict the usage of probability to such contexts.*

- there are other ways to think about assigning probabilities

- $\omega \in \Omega$ is unknown but will be revealed

- a *gamble* is a map $X : \Omega \rightarrow \mathbb{R}$

- a bettor purchases X for $P(X)$ (the price in "utiles") from a bookie who pays out $X(\omega)$ when ω is revealed

- $L(\Omega) =$ set of all gambles is a linear space

- think of yourself as a bettor/bookie (gambler) who will post a price to buy and a price to sell each $X \in L(\Omega)$ and as a rational gambler you subscribe to

> *Principle of Avoiding Sure Loss (Accepting Sure Gains): A rational gambler will never choose prices such that a sure loss will be incurred by some combination of bets and will never choose prices such that alternative choices would produce a sure gain.*

Lemma II.3.1 For a rational gambler, the selling price of $X =$ the buying price of X.

Proof: Let p_1, p_2 be the buying and selling prices for a gamble X. Suppose $p_1 - p_2 > 0$. Then if you both buy and sell X the payoff is

$$(X(\omega) - p_1) + (p_2 - X(\omega)) = p_2 - p_1 < 0$$

is a sure loss. Now suppose $p_1 - p_2 < 0$. So you will not buy X for $p_1 + \epsilon$ for any $\epsilon > 0$ but payout is

$$(X(\omega) - p_1 - \epsilon) + (p_2 - X(\omega)) = p_2 - p_1 - \epsilon > 0$$

which is a sure gain for all ϵ small enough. Therefore, $p_1 = p_2$. ■

- each gambler has a *prevision* $P : L(\Omega) \to \mathbb{R}$ such that $P(X)$ is the buying/selling price of X

Definition II.3.1 The prevision P is *coherent* if for every $m, n \in \mathbb{N}$ and $X_1, \ldots, X_m, Y_1, \ldots, Y_n \in L(\Omega)$

$$\sup_{\omega \in \Omega} \left\{ \sum_{i=1}^{m} (X_i(\omega) - P(X_i)) + \sum_{i=1}^{n} (P(Y_i) - Y_i(\omega)) \right\} \geq 0.$$

- so a coherent prevision guarantees no sure losses through a finite combination of bets

Theorem II.3.1 The prevision P is coherent iff
(i) $P(X+Y)=P(X)+P(Y)$ for every $X,Y\in L(\Omega)$
(ii) $P(\lambda X)=\lambda P(X)$ for every $X\in L(\Omega),\lambda\in\mathbb{R}$
(iii) $P(X)\geq 0$ whenever $X\in L(\Omega)$ satisfies $X\geq 0$
(iv) $P(1)=1$
Proof: See Evans (2015), Theorem 2.3.3.

- **note** - if $A\subset\Omega$ and $X=I_A$ then $P(X)=P(A)$ (notation) can be thought of as the probability of A and this probability is determined for every $A\in 2^\Omega$ and if $A,B\in 2^\Omega$ are disjoint, then

$$P(A\cup B)=P(I_{A\cup B})=P(I_A+I_B)=P(I_A)+P(I_B)=P(A)+P(B)$$

so P is a finitely additive probability measure

- there is nothing in this formulation that demands that P be countably additive when Ω is infinite so continuity of P is lost

- finitely but not countably additive probability measures do arise

Example II.3.1 *Uniform on* $\Omega = \mathbb{N}$

- for $A \in 2^{\mathbb{N}}$ define

$$P(A) = \lim_{n\to\infty} \frac{\#(A \cap \{1, 2, \ldots, n\})}{n}$$

- so $P(\{i\}) = 0$ for every i while $P(\text{”even natural numbers”}) = 1/2$ but

$$P(\text{”even natural numbers”}) \neq \sum_{i=1}^{\infty} P(\{2i\}) = 0$$

so not countably additive ■

- one can define *conditional prevision* $P(\cdot \,|\, B) : L(\Omega) \to \mathbb{R}$ where we are told that the unknown $\omega \in B$ and want $P(X \,|\, B)$ for gamble X

- *contingent gamble*: for $B \subset \Omega$, a gambler buys or sells X for $P(X \,|\, B)$ with payoff $I_B(\omega)(X(\omega) - P(X \,|\, B))$ so the bet is called off if $\omega \notin B$

Theorem II.3.2 (*multiplication rule***)** $P(A \cap B) = P(A \,|\, B)P(B)$ for coherent P.
Proof: Suppose $P(A \cap B) > P(A \,|\, B)P(B)$ and gambler sells A conditional on B, buys unconditional $A \cap B$ and sells $P(A \,|\, B)I_B$. This produces payoff

$$\begin{aligned} & I_B(\omega)(P(A \,|\, B) - I_A(\omega)) + (I_{A \cap B}(\omega) - P(A \cap B)) + \\ & (P(A \,|\, B)P(B) - P(A \,|\, B)I_B(\omega)) \\ = \; & P(A \,|\, B)P(B) - P(A \cap B) < 0 \end{aligned}$$

a sure loss. If $P(A \cap B) < P(A \,|\, B)P(B)$ then reversing the above gambles (multipy above by -1) also gives a sure loss so the result follows. ■

- **note** - for coherent P

$$\begin{aligned} & I_B(\omega)(X - P(X \mid B)) + I_{B^c}(\omega)(X - P(X \mid B^c)) \\ = \; & X - (I_B(\omega)P(X \mid B) + I_{B^c}(\omega)P(X \mid B^c)) \end{aligned}$$

so

$$P(X) = P(B)P(X \mid B) + P(B^c)P(X \mid B^c)$$

and this can be extended to finite partitions $\{B_1, \ldots, B_k\}$

$$P(X) = \sum_{i=1}^{k} P(B_i)P(X \mid B_i)$$

- but this can't in general be extended to infinite partitions so TTE can't be counted on with probability measures that are only finitely additive

- my take-aways

1. An interesting and different perspective on probability.
2. Dropping countable additivity **substantially** *complicates the theory and we lose some key properties of P - the gain isn't worth the cost.*
3. The betting formulation involves the concept of utility in a fundamental way and I don't think this has anything to do with our goals for a theory of inference at least as it applies to science.

- there are other approaches to probability (projects?)
- for me

  Probability measures belief and the concept of randomness is best left out of the story. Probability theory itself presents a sound context for a theory of statistical inference that allows us to reason to answer the **E** *and* **H** *questions in a logical and consistent way. Given that probability assignments are in general not strictly determined by the context, any theory must allow for checking these based on the objective data and for modifying these if necessary.*