# *Foundations of probability: past and present*

Danial Dervovic

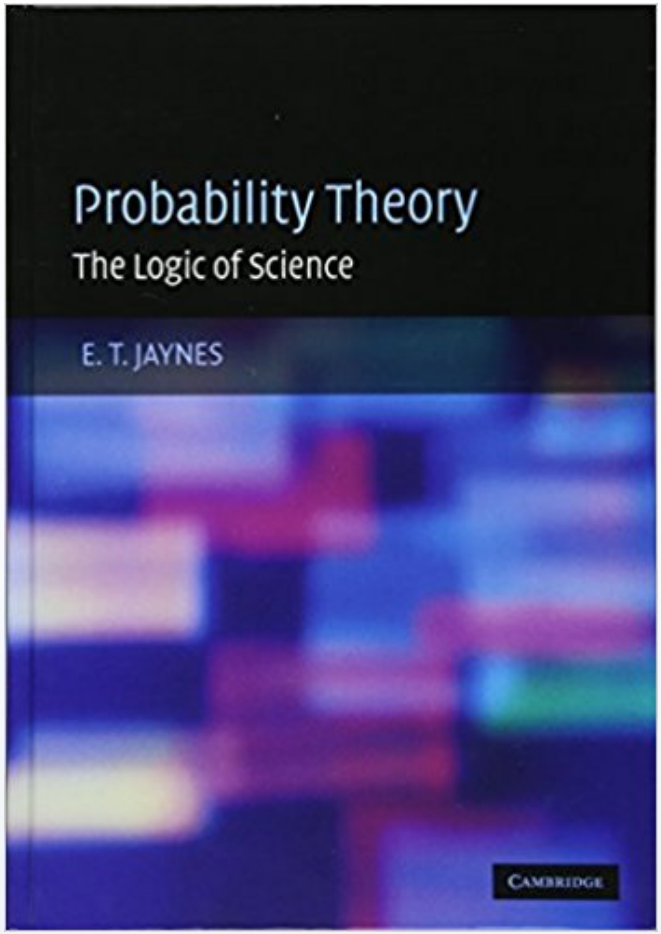
8 February 2019

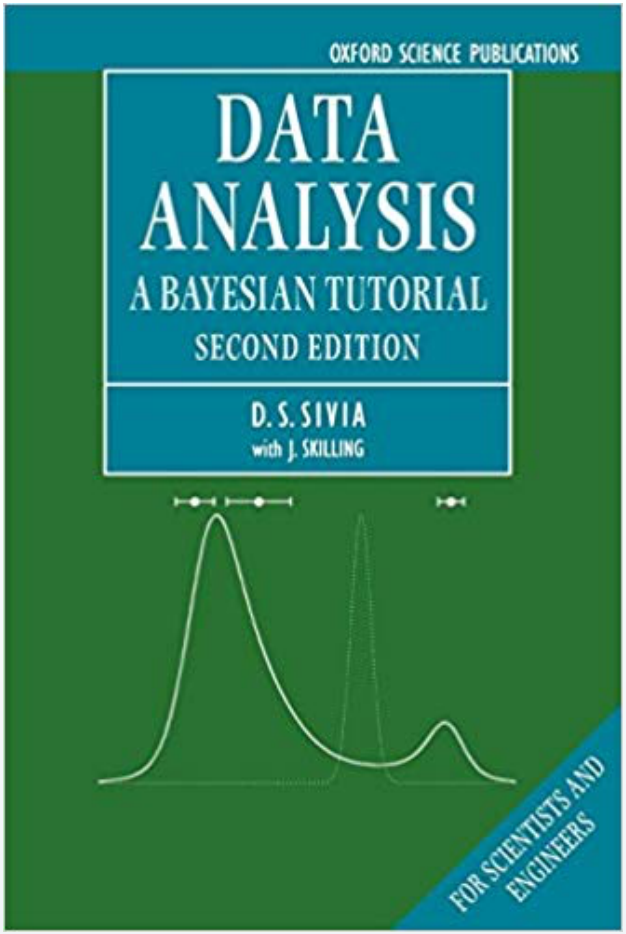
# *Summary*

- Timeline
- Frequentist and Bayesian statistics
- Cox derivation of probability rules
- Maximum Entropy principle
- Quantum Bayesianism

# *Disclaimer*

I will espouse the viewpoint given in the following two (excellent) books



# *Timeline*

Fermat, Pascal 1654

Laplace 1814


Laplace's program attacked


Gerolamo Cardano 1564

Rev. Thomas Bayes 1748

Year (CE)

1500 1600 1700 1800 1900 2000

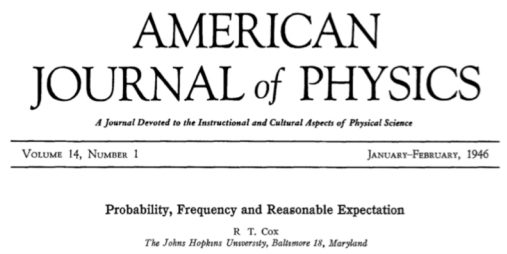
Bayesian revolution

Gamblers' era

Physicists' era

Orthodox Statistics

# Timeline


AMERICAN JOURNAL of PHYSICS

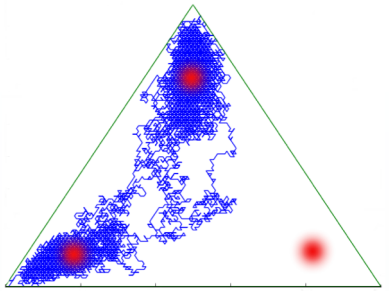
A Journal Devoted to the Instructional and Cultural Aspects of Physical Science

VOLUME 14, NUMBER 1 JANUARY–FEBRUARY, 1946

Probability, Frequency and Reasonable Expectation

R. T. Cox

The Johns Hopkins University, Baltimore 18, Maryland



Kolmogorov 1933

de Finetti 1937

Cox 1946

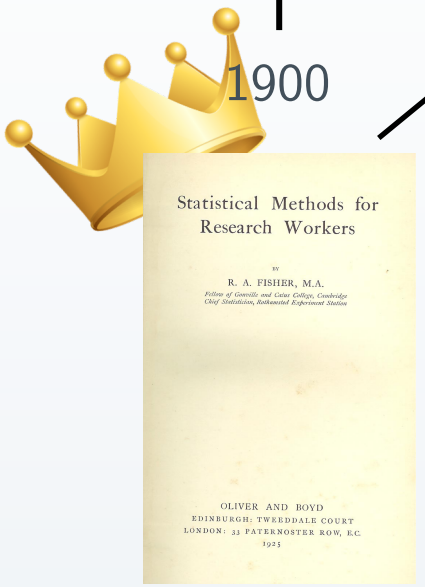
Jaynes introduces MaxEnt 1959

Markov chain Monte Carlo 1970's onwards

Year (CE)

1900

2000


Statistical Methods for Research Workers

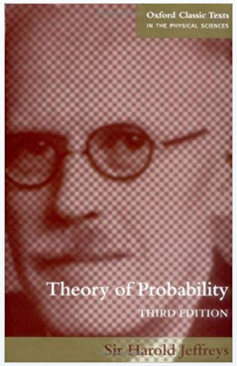
BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

Fisher 1925




Jeffreys 1939

Criticism of Laplace and Jeffreys through-out this period

# *Frequentism and Bayesianism*

Before you have seen the data, what data do you expect to get?

vs

After we have seen the data, do we have any reason to be surprised by them?

# *Frequentism and Bayesianism*

If the as yet unknown data are used to estimate parameters by some known algorithm, how accurate do you expect the estimates to be?

vs

After we have seen the data, what parameter estimates can we now make, and what accuracy are we entitled to claim?

# *Frequentism and Bayesianism*

If the hypothesis being tested is in fact true, what is the probability that we shall get data indicating that it is true?

vs

What is the probability conditional on the data, that the hypothesis is true?

# *Frequentism and Bayesianism*

How much would the estimate of a parameter $\alpha$ vary over the class of all data sets that we might conceivably get?

vs

How accurately is the value of $\alpha$ determined by the one data set $D$ that we actually have?

# *Deductive Reasoning*

$A$ and $B$ are Boolean propositions, they can **only** be true or false.

Example: If $A \Rightarrow B$ and $A$ is true, then $B$ is true.

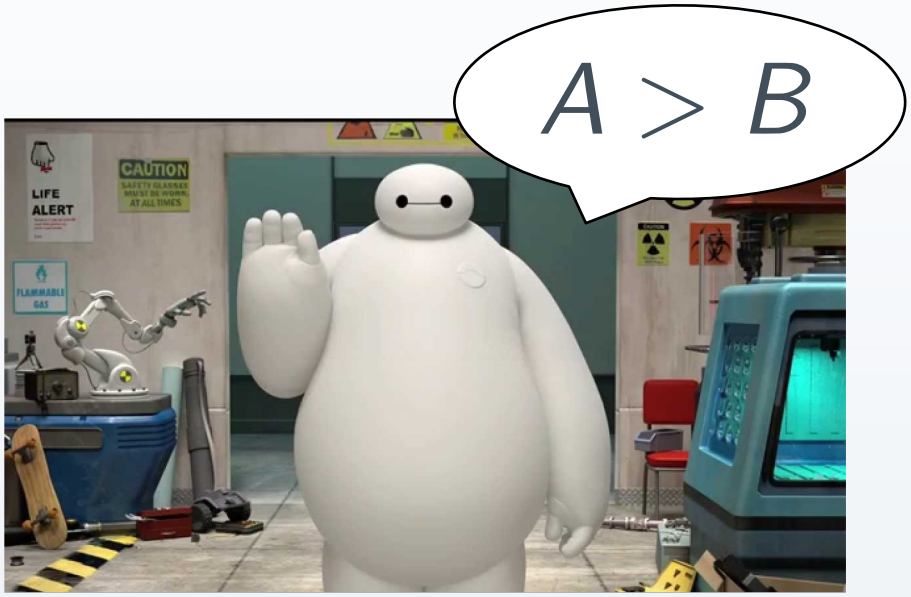
If we discover $A$ is true, then we know with absolute certainty that $B$ is true. This is the accepted standard of mathematical proofs.

This is not like in real life!

# *Plausible Reasoning*

We are going to construct a machine that reasons about the plausibility of Boolean propositions, based on evidence that has been given to it.

We shall impose some constraints that the machine needs to satisfy.



# *Plausible Reasoning - Constraints*

- (I) Degrees of plausibility are represented by real numbers.
- (II) If we assert how much we believe a proposition to be true, we also assert how much we believe it to be false.
- (III) If we specify the plausibility of $Y$ and the plausibility of $X|Y$, then we have specified the plausibility of $X \wedge Y$.
- (IV) Using the same information cannot lead to mutually contradictory plausibility assignments.

# *Plausible Reasoning – Cox's Theorem*

**Theorem. (Cox 1946).** The only plausibility rules that satisfy (I)-(IV) transform in the following way:

$$\text{prob}(X) + \text{prob}(\neg X) = 1$$

and

$$\text{prob}(X \wedge Y) = \text{prob}(X|Y) \times \text{prob}(Y),$$

with $0 \leq \text{prob}(\,\cdot\,) \leq 1$.

# *Plausible Reasoning – Cox Proof*

Defines a system where any plausibility rules not given by the previous slide give a contradiction.

Since rules must work for any possible system, that leaves the usual rules as the only possible.

We slowly whittle down the acceptable plausibility rules until only the familiar ones are left.

# *Plausible Reasoning – Cox Proof*

We consider bitstrings of length 3, so 000, 001, 010, 011, 100, 101, 110, 111, where each bit is the truth of some proposition.

Notation: $A, B, C$ are values for first, second and third bits, i.e. $A \equiv$ first bit has value 1.

We call our plausibility function $\pi : \{0,1\}^* \to \mathbb{R}$, from (I).

# *Plausible Reasoning – Cox Proof*

Consider sequential learning of plausibility (III).

Let $a := \pi(A) =$ plausibility of $A = 1$ and let $b := \pi(B|A)$. Then, $\pi(AB) = F(a, b)$ for some function $F$.

$F$ is the function linking conditional probability.

We then also demand from consistency (IV) that

$$\pi(ABC) = F\left(F(a, b), c\right) = F\left(a, F(b, c)\right),$$

where $c := \pi(C|AB)$. This says that the order we do the conditioning should not change the result.

# *Plausible Reasoning – Cox Proof*

**Lemma 1.** The only functions $F$ satisfying

$$F\left(F(a,b),c\right) = F\left(a,F(b,c)\right)$$

are of the form

$$F(a,b) = w^{-1}\left(w(a)+w(b)\right),$$

where $w$ is some invertible (and therefore monotonic) function of one variable.

# *Plausible Reasoning – Cox Proof*

Let us now rescale our plausibility function, so we can learn additively.

We can do this as plausibility can be any real and $w$ is invertible. We can "pull back" statements about $\phi$ to statements about $\pi$. Define

$$\phi(\,\cdot\,) := w(\pi(\,\cdot\,))$$

$$\begin{aligned}\phi(AB) &= w(\pi(AB)) = w(F(a,b)) = w(a) + w(b) \\ &= \phi(A) + \phi(B|A).\end{aligned}$$

NOTE: we still have the freedom to multiply by a constant.

# *Plausible Reasoning – Cox Proof*

From (II), there exists a function a function $f$ asserting our belief about the converse in the following way:

$$\phi(\neg A) = f(\phi(A)),$$

Negating twice gives identity, so

$$f(f(x)) = x.$$

# *Plausible Reasoning – Cox Proof*

We now restrict to the particular context in which $AB \in \{01, 10, 11\}$.

$$
\begin{aligned}
\phi(AB) &= \phi(A) + \phi(B|A) && \text{sequential learning} \\
&= \phi(A) + f(\phi(\neg B|A)) && \text{definition of } f \\
&= \phi(A) + f(\phi(\neg B, A) - \phi(A)) && \text{sequential de-learning} \\
&= \phi(A) + f(\phi(\neg B) - \phi(A)) && \text{if } B = 0 \text{ we know } A = 1 \\
&= \phi(A) + f(f(\phi(B)) - \phi(A)) && \text{definition of } f \\
&= x + f(f(y) - x) && \text{define } x := \phi(A),\ y := \phi(B)
\end{aligned}
$$

Symmetry $A \wedge B = B \wedge A$ gives

$$
\begin{aligned}
x + f(f(y) - x) &= y + f(f(x) - y) \\
f(f(x)) &= x
\end{aligned}
$$

# *Plausible Reasoning – Cox Proof*

**Lemma 2.** The only solution to the system of functional equations

$$x + f(f(y) - x) = y + f(f(x) - y)$$
$$f(f(x)) = x$$

is

$$f(\xi) = \gamma^{-1} \log(1 - e^{\gamma\xi}),$$

for an arbitrary constant $\gamma \in \mathbb{R}$.

# *Plausible Reasoning – Cox Proof*

Thus, recalling that $\xi = \phi(A) \Longrightarrow f(\xi) = \phi(\neg A)$

$$\phi(\neg A) = \gamma^{-1} \log(1 - \mathrm{e}^{\gamma\phi(A)})$$
$$\Longleftrightarrow \exp\left[\gamma\phi(\neg A)\right] = 1 - \exp\left[\gamma\phi(A)\right]$$

Let us now again redefine our plausibility function $\mathrm{prob}(\,\cdot\,) := \exp\left[\gamma\phi(\,\cdot\,)\right]$

Note now that we have to have $0 \leq \mathrm{prob}(A) \leq 1$ so that neither exponential is negative. This also fixes a scale.

We thus have the sum rule

$$\mathrm{prob}(A) + \mathrm{prob}(\neg A) = 1$$

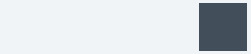
# *Plausible Reasoning – Cox Proof*

The sequential learning equations become

$$\begin{aligned} \phi(A, B) &= \phi(A) + \phi(B|A) \\ \iff \exp\left[\gamma\phi(A, B)\right] &= \exp\left[\gamma(\phi(A) + \phi(B|A))\right] \\ &= \exp\left[\gamma\phi(A)\right] \exp\left[\gamma\phi(B|A)\right] \\ \iff \text{prob}(A, B) &= \text{prob}(B|A)\,\text{prob}(A) \end{aligned}$$

the product rule.

Since we could arbitrarily scale the function $\pi$, we just use prob as our starting point instead. Any calculus of plausibility that does not transform according to the (appropriately scaled) sum and product rule will violate one of the logical constraints of the theory.



# *Bayes Theorem*

$$
\begin{aligned}
& \text{prob}(A, B) = \text{prob}(B, A) \\
\iff & \text{prob}(B|A)\,\text{prob}(A) = \text{prob}(A|B)\,\text{prob}(B) \\
\iff & \boxed{\text{prob}(A|B) = \frac{\text{prob}(B|A)\,\text{prob}(A)}{\text{prob}(B)}}
\end{aligned}
$$

# *Bayes Theorem*

$$\text{prob}(A|B) = \frac{\text{prob}(B|A)\,\text{prob}(A)}{\text{prob}(B)}$$

- $\text{prob}(A|B)$ is the *posterior probability*, the machine's decision for how probable $A$ is, given $B$.
- $\text{prob}(A)$ is the *prior probability*, the probability that something will happen, having gathered no evidence.
- $\text{prob}(B|A)$ is the *likelihood*, i.e. the probability the machine assigns to $B$ assuming $A$.
- $\text{prob}(B)$ is the *evidence*, how likely the machine thinks $B$ is.

# *Principle of maximum entropy (MaxEnt)*

Big question: where do the priors come from? This question has set back the field of statistical inference by centuries!

If we think about it as physicists, there is always some prior information, even just ridiculously large and small upper and lower bounds. Also, previous experiments, experience etc.

Information-theoretically, we are led to the probability distribution over hypotheses with maximum entropy.

Way of mathematically formalising ignorance about hypothesis.

# *MaxEnt – Wallis derivation (1962)*

Suppose we have $M$ distinct possibilities $\{X_i\}$ to be considered; we want to ascribe truth values given the testable information $I$ : prob$(X_i) := p_i$. $I$ can be constraints on the size of the variable $X$, its mean, median etc.

Imagine a game: each $X_i$ is represented by a box, all of the same size, into which a large number $N \gg M$ of pennies are thrown at random.

We give the pennies to a group of monkeys, who are unbiased. The fraction of pennies ending up in a given box gives the estimate $p_i$; if the testable information $I$ is violated, we reject the given distribution over $\{X_i\}$.

# *MaxEnt – Wallis derivation (1962)*

$$\sum_{i=1}^{M} n_i = N$$

$N \gg M$ pennies


$n_1$


$n_2$

$M$ buckets

$p_i = \frac{n_i}{N}$


$n_M$

**Our choice of distribution is the choice that comes up most often in this game**

It turns out the choice that will come up most often is that which maximises the entropy $-\sum_i p_i \log p_i$.

# *MaxEnt – Wallis derivation (1962)*

Since every penny can land in any box by design, there are $M^N$ ways of scattering the coins amongst them.

We have that the frequency with which a distribution $\{p_i\}$ occurs, $F(\{p_i\})$, is given by

$$F(\{p_i\}) = \frac{\text{number of ways of obtaining } \{n_i\}}{M^N}.$$

# *MaxEnt – Wallis derivation (1962)*

$$F(\{p_i\}) = \frac{\text{number of ways of obtaining } \{n_i\}}{M^N}.$$

For the numerator, we can answer the question in the following way

- Box 1: how many ways of putting $n_1$ coins from $N$? ${}^{N}C_{n_1}$
- Box 2: how many ways of putting $n_2$ coins from $N - n_1$? ${}^{N-n_1}C_{n_2}$
- and so on

# *MaxEnt – Wallis derivation (1962)*

$$F(\{p_i\}) = \frac{\text{number of ways of obtaining } \{n_i\}}{M^N}.$$

numerator

$$
\begin{aligned}
&= {}^{N}C_{n_1} \times {}^{N-n_1}C_{n_2} \times {}^{N-n_1-n_2}C_{n_3} \times \cdots \times {}^{n_M}C_{n_M} \\
&= \frac{N!}{n_1!(N-n_1)!} \times \frac{(N-n_1)!}{n_2!(N-n_1-n_2)!} \times \frac{(N-n_1-n_2)!}{n_3!(N-n_1-n_2-n_3)!}] \times \cdots \times \frac{n_M!}{n_M!} \\
&= \frac{N!}{n_1!n_2!\cdots n_M!}
\end{aligned}
$$

$$\log\left(F(\{p_i\})\right) = -N\log M + \log(N!) - \sum_{i=1}^{M} \log(n_i!)$$

# *MaxEnt – Wallis derivation (1962)*

$$\log F \approx -N \log M + N \log N - N - \sum_{i=1}^{M} (n_i \log n_i - n_i)$$

$$= -N \log M + N \log N - N + \sum_{i=1}^{M} (N p_i \log(N p_i)) + N$$

$$= -N \log M + N \log N - N \sum_{i=1}^{M} (p_i \log p_i + p_i \log N)$$

$$= -N \log M + N \log N - N \sum_{i=1}^{M} (p_i \log p_i) - N \log N$$

$$= -N \log M - N \sum_{i=1}^{M} p_i \log p_i$$

From Stirling,
$\log n! \approx n \log n - n.$

# *MaxEnt – Wallis derivation (1962)*

Since logarithm is a monotone function, $F$ is maximised by maximising $-\sum_{i=1}^{M} p_i \log p_i$.

More rigorous derivation in [Shore, Johnson, 1980]

# *MaxEnt – Useful examples*

Assume $I : \langle x \rangle = \mu, x \geq 0$. Then, MaxEnt gives

$$\text{prob}(x|\mu) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right).$$

Assume further that $I : \left\langle (x-\mu)^2 \right\rangle = \sigma^2, x \in \mathbb{R}$. Then, MaxEnt gives

$$\text{prob}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Gaussian distribution without appeals to Central Limit theorem!

# *Quantum Bayesianism*

Field of thought started in earnest circa 2002, with paper by Caves, Fuchs, Schack.

Main thrust of paper: Gleason's theorem + dutch-book argument to show any "subjective" assignment of probabilities must follow quantum rules.

More recently, Fuchs published QBist manifesto in 2010.

Quantum states as an "experimenter's view" of situation.

# *Quantum Bayesianism*

Einstein argument for quantum states as states of knowledge.

$|\psi^{AB}\rangle$

Performs measurement $\mathcal{M}_\phi$ or $\mathcal{M}_\eta$

Immediately knows state drawn from $\{\phi_i^B\}$, or $\{\eta_i^B\}$

Since this holds when $A$ and $B$ can be spacelike-separated, Einstein concludes that quantum states cannot be "real state of affairs".

# *Quantum Bayesianism – SIC POVMs*

Set of $d^2$ rank-one projection operators $\Pi_i = |\psi_i\rangle\langle\psi_i|$ on finite $d$-dimensional Hilbert space such that

$$|\langle\psi_i|\psi_j\rangle| = \frac{1}{d+1} \quad \text{whenever} \quad i \neq j.$$

1. Operators are linearly independent and span the space of Hermitian operators.

2. They are as close as can be to a basis, when every matrix is positive semidefinite.

3. After rescaling, they resolve to identity.

Indeed, since they are psd and resolve to identity, the $\{\Pi_i\}$ form a POVM.

# *Quantum Bayesianism – SIC POVMs*

A measurement with outcome $H_i$ occurs with probability $\mathrm{Tr}(H_i)$ (abuse of notation here!). Can expand any state

$$\rho = \sum_{i=1}^{d^2} \left( (d+1)P(H_i) - \frac{1}{d} \right) \Pi_i .$$

BIG CAVEAT: SIC POVMs not shown to exist in all dimensions!

# *Summary*

- Can formulate probability theory as (unique method of) plausible reasoning.
- This forms the foundation of the "Bayesian" school.
- MaxEnt gives a well-justified priors, allowing Bayesian statistical inference.
- Quantum Bayesian perspective well-developed, but by no means complete.