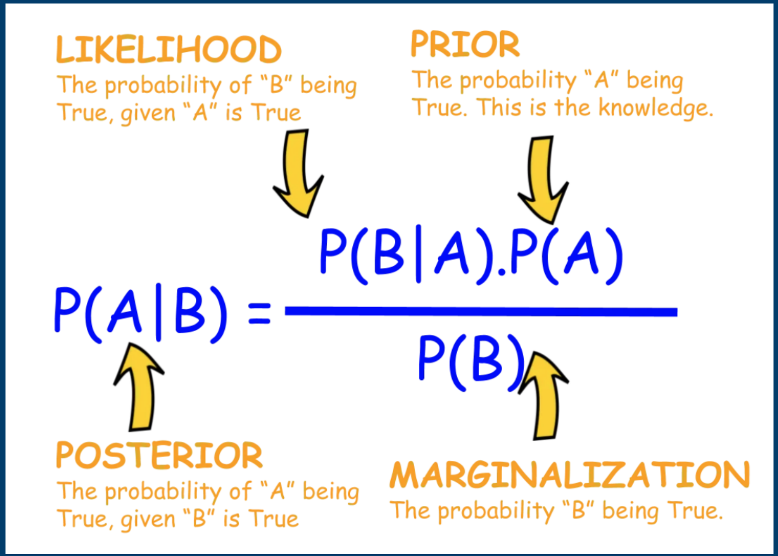
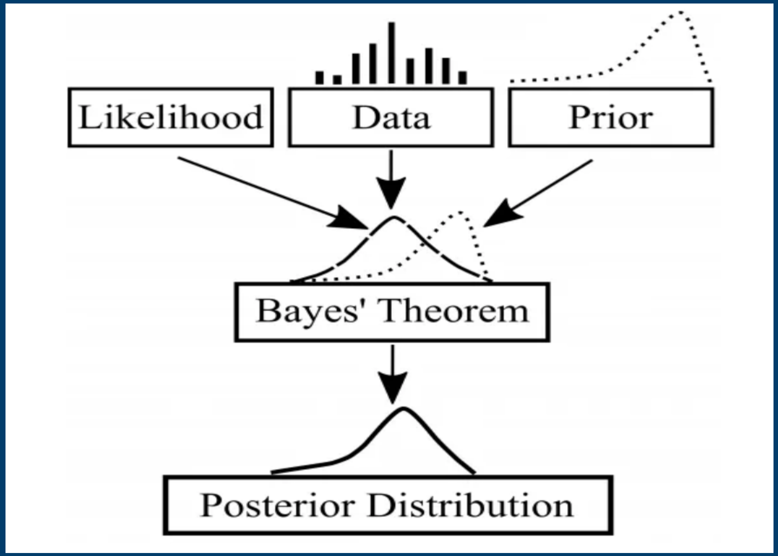
REPRESENTATION THEOREMS IN BAYESIAN STATISTICS

06.01.2025 I J. ZSOLT BERNÁD





JÜLICH Forschungszentrum
Wandel gestalten

OVERVIEW

BAYESIAN VIEW

„The only relevant thing is uncertainty-the extent of our own knowledge and ignorance.“ (B. de Finetti)

„Bayesian statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterising how an individual should act in order to avoid certain kinds of undesirable behavioural inconsistencies.“ (J. M. Bernardo and A. F. M. Smith)

Example: game of roulette; the possible outcomes are the numbers 0,1, . . . ,36; all 37 outcomes have the same probability $p = 1/37$, if we have no more information; another player obtains information about the ball’s position and velocity relative to the wheel, so more accurate predictions and her/his probability is peaked around some group of numbers.

Whose probability is the true probability? From the Bayesian viewpoint: there is no such thing as a true probability. All probability assignments are subjective assignments based specifically upon one’s prior data and beliefs.

BAYESIAN VIEW

Example: Suppose a die is thrown 10 times. Let $k \in \{1, . . . , 6\}$ represent the outcome of a single throw of the die, and suppose the results in the 10 throws were these:

k=1 appeared 1 times,
k=2 appeared 4 times
k=3 appeared 2 times
k=4 appeared 2 times
k=5 appeared 1 times
k=6 appeared 0 times

A typical inference problem is to assign a probability, p, to the outcome k = 6 in the next throw of the die, given these data. ***Impossible*** with this data alone, so we need prior probability assignments + Bayes‘ theorem. Data alone is never enough to specify a probability distribution.

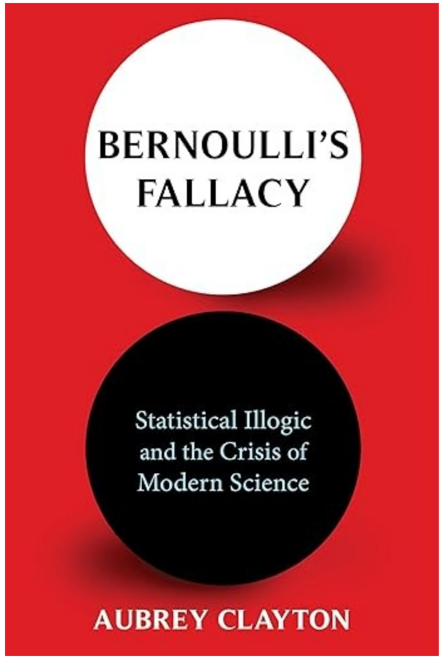
„...learning processes, whatever their particular concerns and fashions at any given point in time, are necessarily reasoning processes which take place in the minds of individuals. To be sure, the object of attention and interest may well be an assumed external, objective reality: but the actuality of the learning process consists in the evolution of individual, subjective beliefs about that reality.“ (J. M. Bernardo and A. F. M. Smith)

BAYESIAN VIEW

Bayesian methodology

- The use of random variables to model all sources of uncertainty in statistical models including uncertainty resulting from lack of information. Example: P($c=3\times10^{8}$ m/s| data)=99.99%; Planck constant, ….
- The need to determine the prior probability distribution: objective (old school) vs. subjective (new school).
- Bayes' theorem used sequentially: posterior probability becomes prior.

Frequentist approach under criticism:



The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

Pages 129-133

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

EXCHANGEABILITY

- Bernoulli trials with unknown probability (frequentist approach).
- In the Bayesian approach, it does not make sense to talk about estimating this probability.
- Choosing a prior among all possible distributions could be a daunting task.
- Bruno de Finetti found an answer for this.

- ***Exchangeability***:

A *finite* sequence of random variables $X_1, X_2, \ldots, X_n$ is *(finitely) exchangeable* with (joint) probability measure P, if, for any permutation π of indices

$$P(X_1, X_2, \ldots, X_n) = P(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$$

For example, the random variables (X_1, X_2, X_3, X_4) are exchangeable if

$$P(X_1, X_2, X_3, X_4) = P(X_2, X_4, X_1, X_3) = P(X_1, X_3, X_2, X_4) = \cdots$$

An *infinite* sequence, $X_1, X_2, \ldots$, is *infinitely exchangeable* if any finite subset of the sequence is finitely exchangeable.

EXCHANGEABILITY

- ***Non-extendible exchangeability***

three random variables x_1, x_2, and x_3 taking values 0 or 1

joint probability mass function $p(x_1=0, x_2=1, x_3=1)=p(x_1=1, x_2=0, x_3=1)=p(x_1=1, x_2=1, x_3=0)=1/3$ with all other combinations of x_1, x_2, and x_3 having probability zero

We shall now try to identify an x_4 taking only values 0 or 1 such that x_1, x_2, x_3, and x_4 are exchangeable

$p(x_1=0, x_2=1, x_3=1, x_4=0)=p(x_1=0, x_2=0, x_3=1, x_4=1)$

But $p(x_1=0, x_2=1, x_3=1, x_4=0)=p(x_1=0, x_2=1, x_3=1) - p(x_1=0, x_2=1, x_3=1, x_4=1)=1/3-p(x_1=1, x_2=1, x_3=1, x_4=0)=1/3$

However, $p(x_1=0, x_2=0, x_3=1, x_4=1)\leq p(x_1=0, x_2=0, x_3=1)=0$

DE FINETTI'S 0-1 REPRESENTATION THEOREM

If $X_1, X_2, \ldots$ is an infinitely exchangeable sequence of 0-1 variables with probability measure P, then there exists a distribution function Q such that the joint mass function of $(X_1, X_2, \ldots, X_n)$ has the form

$$p\left(X_1, X_2, \ldots, X_n\right) = \int_0^1 \left\{\prod_{i=1}^{n} \theta^{X_i} \left(1-\theta\right)^{1-X_i}\right\} dQ\left(\theta\right)$$

where

$$Q\left(t\right) = \lim_{n\to\infty} P\left[\frac{Y_n}{n} \leq t\right]$$

and $Y_n = \sum_{i=1}^{n} X_i$, *and*

$$\theta \stackrel{def}{=} \lim_{n\to\infty} Y_n/n \qquad \because \qquad Y_n/n \xrightarrow{a.s.} \theta$$

is the (strong-law) limiting relative frequency of 1s.

DE FINETTI'S 0-1 REPRESENTATION THEOREM

By exchangeability, for $0 \leq y_n \leq n$

$$P\left[Y_n = y_n\right] = \binom{n}{y_n} p\left(x_1, x_2, \ldots, x_n\right) = \binom{n}{y_n} p\left(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}\right)$$

where $X_i = x_i$ and

$$y_n = \sum_{i=1}^{n} x_i$$

$$P\left[Y_n = y_n\right] = \sum P\left[Y_n = y_n | Y_N = y_N\right] P\left[Y_N = y_N\right]$$

$$P\left[Y_n = y_n | Y_N = y_N\right] = \frac{\binom{y_N}{y_n}\binom{N - y_N}{n - y_n}}{\binom{N}{n}} \qquad 0 \leq y_n \leq n.$$

hypergeometric mass function

DE FINETTI'S 0-1 REPRESENTATION THEOREM

Falling factorial

$$P\left[Y_n = y_n\right] = \binom{n}{y_n} \sum \frac{\left(y_N\right)_{y_n} \left(N - y_N\right)_{n-y_n}}{\left(N\right)_n} P\left[Y_N = y_N\right]$$

$$(x)_r = x(x-1)(x-2)\ldots(x-r+1)$$

Define $Q_N(\theta)$ on the real line as the step function which is zero for $\theta<0$ and has steps of size $P[Y_N=y_N]$ at $\theta=y_N/N$ for $y_N=0, 1, 2, ..., N$

$$P\left[Y_n = y_n\right] = \binom{n}{y_n} \int_0^1 \frac{\left(\theta N\right)_{y_n} \left(\left(1-\theta\right) N\right)_{n-y_n}}{\left(N\right)_n} dQ_N\left(\theta\right)$$

$$\frac{\left(\theta N\right)_{y_n} \left(\left(1-\theta\right) N\right)_{n-y_n}}{\left(N\right)_n} \rightarrow \theta^{y_n} \left(1-\theta\right)^{n-y_n} = \prod_{i=1}^{n} \theta^{x_i} \left(1-\theta\right)^{1-x_i}$$

$(x)_r \rightarrow x^r$ if $x \rightarrow \infty$ with r fixed

Helly's selection theorem: $\left\{Q_N\left(\theta\right); N = 1, 2, \ldots\right\}$ has a convergent subsequence $\left\{Q_{N_j}\left(\theta\right)\right\}$

for some distribution function Q,

$$\lim_{j\rightarrow\infty} Q_{N_j}\left(\theta\right) = Q\left(\theta\right)$$

DE FINETTI'S GENERAL REPRESENTATION THEOREM

If $X_1, X_2, \ldots$ is an infinitely exchangeable sequence of variables with probability measure P, then there exists a distribution function Q on $\mathcal{F}$, the set of all distribution functions on $\mathbb{R}$, such that the joint distribution of $(X_1, X_2, \ldots, X_n)$ has the form

$$p\left(X_1, X_2, \ldots, X_n\right) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(X_i)\, dQ\left(F\right)$$

where F is an unknown/unobservable distribution function

$$Q\left(F\right) = \lim_{n \to \infty} P\left(\widehat{F}_n\right)$$

is a probability measure on the space of functions $\mathcal{F}$, defined as a limiting measure as $n \longrightarrow \infty$ on the ***empirical distribution function*** $\widehat{F}_n$.

$$\widehat{F}_n(t) = \frac{\text{number of elements in the sample} \leq t}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq t},$$

INTERPRETATION

- the X_i are judged to be independent, Bernoulli random quantities conditional on a random quantity θ;
- θ is itself assigned a probability distribution Q;
- by the strong law of large numbers Q may be interpreted as “beliefs about the limiting relative frequency of 1 ’s;
- likelihood function of the Bernoulli trials multiplied by the prior probability distribution function $Q(\theta)$;
- Thus, from a very simple and natural assumption (exchangeability) about observable random quantities, we have a theoretical justification for using Bayesian methods, and a natural interpretation of parameters as limiting quantities.
- Comment: $Q_N \rightarrow Q$ looks like Glivenko–Cantelli theorem, but it is not the case, because X_i-s are not independent and identically distributed

REPRESENTATION THEOREMS VIA INVARIANCE

Spherical symmetry
A sequence of random variables $x_1, \ldots, x_n$ is said to have spherical symmetry under a predictive probability model P if the latter defines the distributions of $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{Ax}$ to be identical, for any orthogonal $n \times n$ matrix $\mathbf{A}$.

Centered spherical symmetry
A sequence of random variables $x_1, \ldots, x_n$ is said to have centered spherical symmetry if the random variables $x_1 - \bar{x}_n, \ldots, x_n - \bar{x}_n$ have spherical symmetry, where $\bar{x}_n = n^{-1}(x_1 + \cdots + x_n)$.

- Predictive probability model: mathematically specifies the form of the joint distribution

REPRESENTATION THEOREMS VIA INVARIANCE

If x_1, x_2, ... is an infinitely exchangeable sequence of real-valued random variables with probability measure P, and if, for any n, $\{x_1, \dots, x_n\}$ have centered spherical symmetry, then there exists a distribution function Q on $\mathbb{R} \times \mathbb{R}^+$ such that the joint distribution of $x_1, \dots, x_n$ has the form

$$p(x_1, \dots, x_n) = \int_{\mathbb{R}\times\mathbb{R}^+} \prod_{i=1}^{n} \Phi\left[\lambda^{1/2}(x_i - \mu)\right] dQ(\mu, \lambda),$$

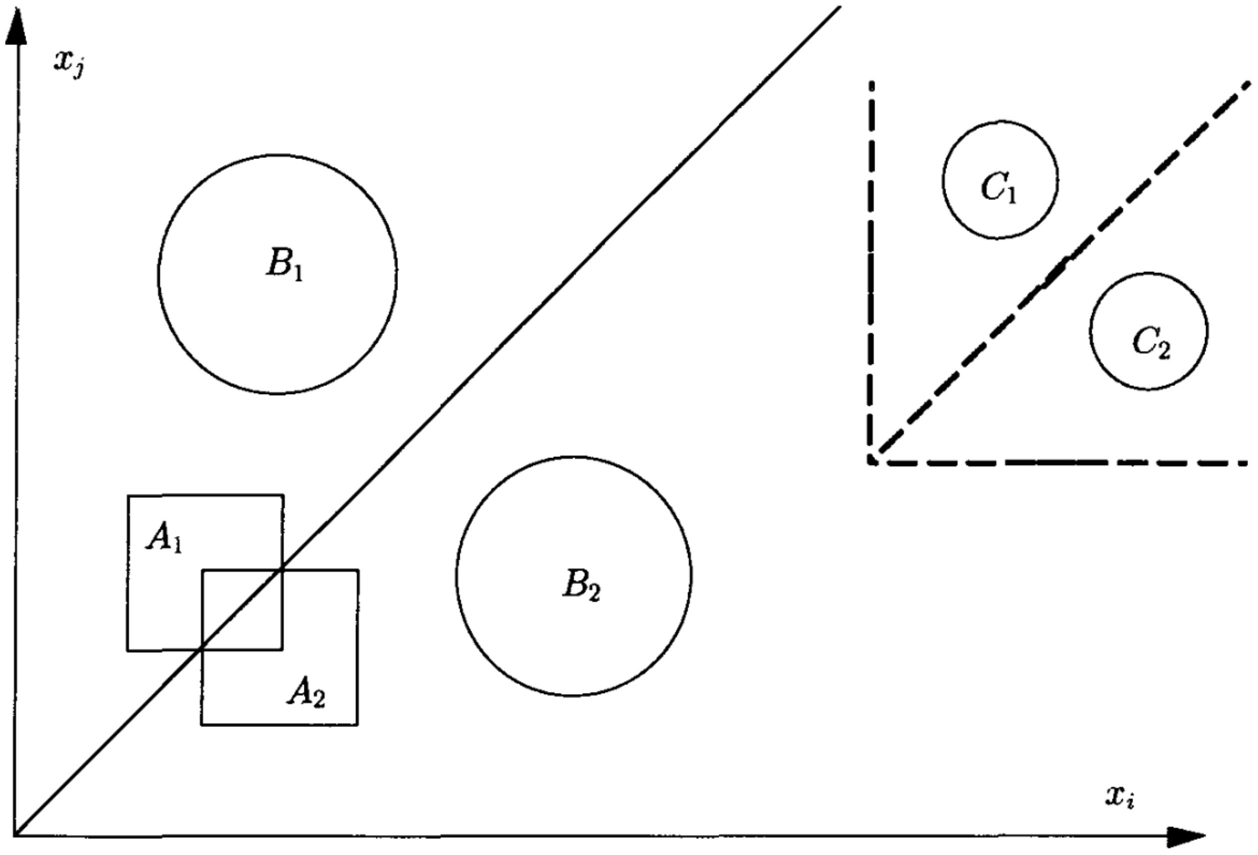
where Φ is the standard normal distribution function and

$$Q(\mu, \lambda) = \lim_{n\to\infty} P\left[(\bar{x}_n \le \mu) \cap (s_n^{-2} \le \lambda)\right],$$

with $\bar{x}_n = n^{-1}(x_1 + \dots + x_n)$, $s_n^2 = n^{-1}\left[(x_1 - \bar{x}_n)^2 + \dots + (x_n - \bar{x}_n)^2\right]$, $\mu = \lim_{n\to\infty} \bar{x}_n$, and $\lambda^{-1} = \lim_{n\to\infty} s_n^2$.

- Proof: A. F. M. Smith, J. Roy. Statist. Soc. B **43**, 208-209 (1981).

REPRESENTATION THEOREMS VIA INVARIANCE



- equal probabilities + assumption of exchangeability = origins of the x_i and x_j axes are irrelevant

REPRESENTATION THEOREMS VIA INVARIANCE

If x_1, x_2, … *is an infinitely exchangeable sequence of positive real-valued random variables with probability measure* P, *such that, for all* n, *and any event* A *in* $\mathbb{R}^+ \times \cdots \times \mathbb{R}^+$,

$$P\left[(x_1, \ldots, x_n) \in A\right] = P\left[(x_1, \ldots, x_n) \in A + a\right]$$

for all $a \in \mathbb{R} \times \cdots \times \mathbb{R}$ *such that* $a^t\mathcal{I} = 0$ *and* $A + a$ *is an event in* $\mathbb{R}^+ \times \cdots \times \mathbb{R}^+$, *then the joint density for* $x_1, \ldots, x_n$ *has the form*

$$p(x_1, \ldots, x_n) = \int_0^\infty \prod_{i=1}^n \theta e^{-\theta x_i} dQ(\theta),$$

where

$$Q(\theta) = \lim_{n\to\infty} P\left[(\bar{x}_n^{-1}) \le \theta\right], \quad \bar{x}_n = n^{-1}(x_1 + \cdots + x_n).$$

- Proof: P. Diaconis, and D. Ylvisaker, Bayesian Statistics **2** (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), 133-156 (1985).

PARTIAL EXCHANGEABILITY

- we have restricted our attention to the case of a single sequence of random variables;
- x_i for the index set I: but we can have clinical responses for drug type I_1 or I_2 ;
- I different irrigation system, J different tree pruning techniques, K different trees;
- each sequence is infinitely exchangeable;
- unrestricted exchangeability: academic performance 9th graders from different schools; x_{ij} for i-th students from j-th school; no completely exchangeable, because that would imply that schools have no effect on academic performance; assume that students are exchangeable within each school, and that schools are exchangeable among themselves.
- separate exchangeability, joint exchangeability...

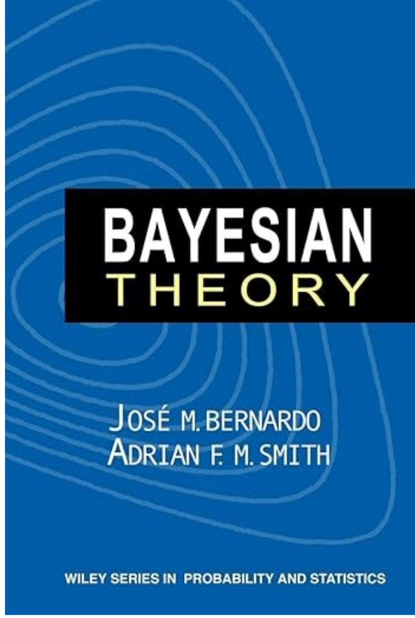
OBJECTIVE AND SUBJECTIVE PROBABILITIES

- model: defines ”objective” probabilities for outcomes defined in terms of observables;
- these probabilities being determined by the values of the “unknown parameters“;
- if the “true” parameter were known, the corresponding parametric form would be the “true” model for the observables;
- the “prior”, on the other hand, is seen as a ”subjective” optional extra, a potential contaminant of the objective statements

- Bayesians: in the absence of any general agreement over assumptions of symmetry or invariance the individuals are each simply left with their own subjective assessments
- Bayesians + common assumptions: the parametric forms adopted will be the same (representation theorems) ⟶ **intersubjective agreement**, the priors remain the subjective components;
- intersubjective agreement clearly facilitates communication
- as more data are acquired, a group of Bayesians may eventually come to share very similar beliefs;

- objectivity seems to be equal to convenient or intersubjective communality of beliefs.

OUTLOOK



- role and nature of models in statistical analysis: a focused framework to serve as a basis for subsequent identification of areas of agreement and disagreement; reservoir of models; one must necessarily work with simplified representations;
- scientific and technological approaches to models;
- scientific model: aim at providing insight into and understanding of the "true" mechanisms of the phenomenon under study;
- technological model: providing a reliable basis for practical action in predicting and controlling phenomena of interest;
- identifiability.

THANK YOU FOR YOUR ATTENTION!