The Bayesian (nonparametric) approach to Statistics via exchangeability

Alessandra Guglielmi

Dipartimento di Matematica
Politecnico di Milano, Milano (Italy)

BAYSM 2016, Firenze (ITALY)

21st June 2016

POLITECNICO
MILANO 1863

Outline

Use probabilities informally to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning it will be cloudy in Florence is ... , according to my knowledge (on the average weather in June and today's weather in Florence)

Use probabilities informally to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning it will be cloudy in Florence is ... , according to my knowledge (on the average weather in June and today's weather in Florence)

But what if a storm occurs in the evening? I will **UPDATE** the probability of that event on the basis of some **DATA** (rain this evening!)

Use probabilities informally to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning it will be cloudy in Florence is ... , according to my knowledge (on the average weather in June and today's weather in Florence)

But what if a storm occurs in the evening? I will **UPDATE** the probability of that event on the basis of some **DATA** (rain this evening!)

✓ Conditional probabilities and Bayes' Theorem provide a rational method for updating beliefs in the light of new information (DATA)

Use probabilities informally to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning it will be cloudy in Florence is ... , according to my knowledge (on the average weather in June and today's weather in Florence)

But what if a storm occurs in the evening? I will **UPDATE** the probability of that event on the basis of some **DATA** (rain this evening!)

✓ Conditional probabilities and Bayes' Theorem provide a rational method for updating beliefs in the light of new information (DATA)

✓ The process of inductive learning via Bayes' Theorem is referred to as

BAYESIAN INFERENCE

Use probabilities informally to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning it will be cloudy in Florence is ... , according to my knowledge (on the average weather in June and today's weather in Florence)

But what if a storm occurs in the evening? I will **UPDATE** the probability of that event on the basis of some **DATA** (rain this evening!)

✓ Conditional probabilities and Bayes' Theorem provide a rational method for updating beliefs in the light of new information (DATA)

✓ The process of inductive learning via Bayes' Theorem is referred to as

BAYESIAN INFERENCE

✓ It is a typical scientific approach:

Use probabilities informally to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning it will be cloudy in Florence is ... , according to my knowledge (on the average weather in June and today's weather in Florence)

But what if a storm occurs in the evening? I will **UPDATE** the probability of that event on the basis of some **DATA** (rain this evening!)

✓ Conditional probabilities and Bayes' Theorem provide a rational method for updating beliefs in the light of new information (DATA)

✓ The process of inductive learning via Bayes' Theorem is referred to as

BAYESIAN INFERENCE

✓ It is a typical scientific approach:

- the prior belief is UPDATED via observed data and yields posterior distribution

Use probabilities informally to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning it will be cloudy in Florence is ... , according to my knowledge (on the average weather in June and today's weather in Florence)

But what if a storm occurs in the evening? I will **UPDATE** the probability of that event on the basis of some **DATA** (rain this evening!)

✓ Conditional probabilities and Bayes' Theorem provide a rational method for updating beliefs in the light of new information (DATA)

✓ The process of inductive learning via Bayes' Theorem is referred to as

BAYESIAN INFERENCE

✓ It is a typical scientific approach:

- the prior belief is UPDATED via observed data and yields posterior distribution
- it suggests that scientific inference is based on 2 parts: one depends on the scientist's subjective opinion and understanding of the phenomenon under study **BEFORE an EXPERIMENT** was performed, the other depends on the observed data the scientist has obtained from the experiment.

This lesson is based mainly on Regazzini (2015), an item of the encyclopedia Wiley StatsRef.

- $\mathbb{X}$ is a separable and complete metric space
- $\mathscr{X}$ is the Borel σ–algebra of subsets of $\mathbb{X}$
- $(X_n)_{n\geq 1}$ is a sequence of random elements defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbb{X}^\infty, \mathscr{X}^\infty)$

$(X_n)_{n\geq 1}$ is the sequence of observations (DATA); X_n is the result of the random experiment at trial n

This lesson is based mainly on Regazzini (2015), an item of the encyclopedia Wiley StatsRef.

- $\mathbb{X}$ is a separable and complete metric space
- $\mathscr{X}$ is the Borel σ–algebra of subsets of $\mathbb{X}$
- $(X_n)_{n\geq 1}$ is a sequence of random elements defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbb{X}^\infty, \mathscr{X}^\infty)$

$(X_n)_{n\geq 1}$ is the sequence of observations (DATA); X_n is the result of the random experiment at trial n

- $\mathbf{P}_{\mathbb{X}}$ space of all probability measures on $(\mathbb{X}, \mathscr{X})$, with the topology of weak convergence
- $\mathscr{P}_{\mathbb{X}}$ Borel σ–algebra of subsets of $\mathbf{P}_X$
- A random element $\tilde{p}$ defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbf{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$ is a *random probability measure*.

Traditionally:

$$X_1, X_2, X_3, \ldots | \tilde{p} = p \overset{\text{i.i.d.}}{\sim} p \text{ “true” distribution of each observation}$$
$$\tilde{p} \sim Q \text{ prior}$$

$$\Rightarrow \pi(A_1 \times \cdots \times A_n \times B) := P(X_1 \in A_1, \ldots X_n \in A_n, \tilde{p} \in B) = \int_B \prod_{i=1}^n p(A_i) Q(dp)$$

π is a probability on $\mathbb{X}^\infty \times \mathbf{P}_{\mathbb{X}}$

Traditionally:

$$X_1, X_2, X_3, \ldots | \tilde{p} = p \overset{\text{i.i.d.}}{\sim} p \text{ “true” distribution of each observation}$$
$$\tilde{p} \sim Q \text{ prior}$$

$$\Rightarrow \pi(A_1 \times \cdots \times A_n \times B) := P(X_1 \in A_1, \ldots X_n \in A_n, \tilde{p} \in B) = \int_B \prod_{i=1}^{n} p(A_i) Q(dp)$$

π is a probability on $\mathbb{X}^\infty \times \mathbf{P}_{\mathbb{X}}$

There exists a function $Q_n(B; x^{(n)})$, $B \in \mathscr{P}_{\mathbb{X}}$, $x^{(n)} := (x_1, \ldots, x_n) \in \mathbb{X}^n$ such that

- ✓ $B \mapsto Q_n(B; x^{(n)})$ is a probability on $\mathscr{P}_{\mathbb{X}}$ for all $x^{(n)}$;
- ✓ $x^{(n)} \mapsto Q_n(B; x^{(n)})$ is $\mathscr{X}^n$-measurable for all B;
- ✓ $Q_n(B; x^{(n)}) \overset{a.s.}{=} P(\tilde{p} \in B | X^{(n)})$ for all B

$\Rightarrow Q_n(\cdot; X^{(n)})$ is the conditional distribution of $\tilde{p}$ (the ‘true” distribution of the data), given $X^{(n)} := (X_1, \ldots, X_n)$ ($\mathbb{X}$ Polish space)

Traditionally:

$$X_1, X_2, X_3, \ldots | \tilde{p} = p \overset{\text{i.i.d.}}{\sim} p \text{ “true” distribution of each observation}$$
$$\tilde{p} \sim Q \text{ prior}$$

$$\Rightarrow \pi(A_1 \times \cdots \times A_n \times B) := P(X_1 \in A_1, \ldots X_n \in A_n, \tilde{p} \in B) = \int_B \prod_{i=1}^n p(A_i) Q(dp)$$

π is a probability on $\mathbb{X}^\infty \times \mathbf{P}_\mathbb{X}$

There exists a function $Q_n(B; x^{(n)})$, $B \in \mathscr{P}_\mathbb{X}$, $x^{(n)} := (x_1, \ldots, x_n) \in \mathbb{X}^n$ such that

- ✓ $B \mapsto Q_n(B; x^{(n)})$ is a probability on $\mathscr{P}_\mathbb{X}$ for all $x^{(n)}$;
- ✓ $x^{(n)} \mapsto Q_n(B; x^{(n)})$ is $\mathscr{X}^n$-measurable for all B;
- ✓ $Q_n(B; x^{(n)}) \overset{a.s.}{=} P(\tilde{p} \in B | X^{(n)})$ for all B

$\Rightarrow Q_n(\cdot; X^{(n)})$ is the conditional distribution of $\tilde{p}$ (the ‘true” distribution of the data), given $X^{(n)} := (X_1, \ldots, X_n)$ ($\mathbb{X}$ Polish space)

$Q_n(\cdot; X^{(n)})$ is the **posterior distribution** of $\tilde{p}$, given observations $X_1, \ldots, X_n$

$$X_i|\tilde{\theta} = \theta \overset{\text{i.i.d.}}{\sim} f_\theta(\cdot) \text{ “true” density of each observation}$$

$$\tilde{\theta} \sim \pi \text{ probability on } \Theta \text{ Euclidean space}$$

Interpretation: $\theta \mapsto \prod_{i=1}^{n} f_\theta(x_i)$ likelihood, $\pi(d\theta)$ prior

Bayes' Theorem for dominated models

$$X_i|\tilde{\theta} = \theta \overset{\text{i.i.d.}}{\sim} f_\theta(\cdot) \text{ "true" density of each observation}$$
$$\tilde{\theta} \sim \pi \text{ probability on } \Theta \text{ Euclidean space}$$

Interpretation: $\theta \mapsto \prod_{i=1}^n f_\theta(x_i)$ likelihood, $\pi(d\theta)$ prior

Then the posterior distribution of $\tilde{\theta}$, given $X_1 = x_1, \ldots, X_n = x_n$, can be computed by **Bayes' Theorem**:

$$P(\tilde{\theta} \in B | X_1 = x_1, \ldots, X_n = x_n) \overset{a.s.}{=} \frac{\int_B \prod_{i=1}^n f_\theta(x_i)\pi(d\theta)}{\int_\Theta \prod_{i=1}^n f_\theta(x_i)\pi(d\theta)}, \quad B \in \mathscr{B}(\Theta)$$

Proof: definition of conditional distribution (as the solution of an integral equation) + Radon-Nikodym Theorem

Making inference about $\tilde{\theta}$ is to learn about the unknown $\tilde{\theta}$ from the **DATA**: based on the data, explore which values of $\tilde{\theta}$ are probable, what might be plausible numbers as estimates of the different components of θ and the extent of uncertainty associated with such estimates

the distribution of $\tilde{\theta}$, **prior distribution**: it quantifies the uncertainty about $\tilde{\theta}$ *prior* to seeing data

Making inference about $\tilde{\theta}$ is to learn about the unknown $\tilde{\theta}$ from the **DATA**: based on the data, explore which values of $\tilde{\theta}$ are probable, what might be plausible numbers as estimates of the different components of θ and the extent of uncertainty associated with such estimates

the distribution of $\tilde{\theta}$, **prior distribution**: it quantifies the uncertainty about $\tilde{\theta}$ *prior* to seeing data

The prior represents the subjective belief and knowledge $\mapsto$ subjective prior,

or, a conventional prior supposed to represent small or no information $\mapsto$ noninformative/vague/objective prior

Making inference about $\tilde{\theta}$ is to learn about the unknown $\tilde{\theta}$ from the **DATA**: based on the data, explore which values of $\tilde{\theta}$ are probable, what might be plausible numbers as estimates of the different components of θ and the extent of uncertainty associated with such estimates

the distribution of $\tilde{\theta}$, **prior distribution**: it quantifies the uncertainty about $\tilde{\theta}$ *prior* to seeing data

The prior represents the subjective belief and knowledge $\mapsto$ subjective prior,

or, a conventional prior supposed to represent small or no information $\mapsto$ noninformative/vague/objective prior ϟARGH!!!! ϟ Jupiter Fulminator could hurl thunderbolts to us at any moment since now, since we follow, at least in principle,

Bruno de Finetti's approach to Bayesian Statistics, i.e.
the subjective approach!

- born on 13 June 1906 in Innsbruck (Austria)
- at 17 he enrolled at the Polytechnic of Milan, aiming at getting a degree in Engineering
- in 1925 he decided to leave Polytechnic and started studying Mathematics at the newly opened faculty at Università di Milano
- in 1927 he got a degree in Applied Mathematics, with a thesis on affine geometry
- after his graduation, he started to work at what is now ISTAT, the Italian central institute of Statistics, until 1931, when he started to work as an actuary in Trieste (Italy) at the Assicurazioni Generali, and started to teach at the university there (he had also qualified in a competition as university lecturer)
- in 1947 he obtained his chair as full professor
- in 1954 he moved to Università La Sapienza, Rome (Italy); he retired in 1976
- he died on 20 July 1985

- born on 13 June 1906 in Innsbruck (Austria)
- at 17 he enrolled at the Polytechnic of Milan, aiming at getting a degree in Engineering
- in 1925 he decided to leave Polytechnic and started studying Mathematics at the newly opened faculty at Università di Milano
- in 1927 he got a degree in Applied Mathematics, with a thesis on affine geometry
- after his graduation, he started to work at what is now ISTAT, the Italian central institute of Statistics, until 1931, when he started to work as an actuary in Trieste (Italy) at the Assicurazioni Generali, and started to teach at the university there (he had also qualified in a competition as university lecturer)
- in 1947 he obtained his chair as full professor
- in 1954 he moved to Università La Sapienza, Rome (Italy); he retired in 1976
- he died on 20 July 1985
- 1928: de Finetti presents his first results on exchangeability at the World Meeting of Mathematicians in Bologna

- born on 13 June 1906 in Innsbruck (Austria)
- at 17 he enrolled at the Polytechnic of Milan, aiming at getting a degree in Engineering
- in 1925 he decided to leave Polytechnic and started studying Mathematics at the newly opened faculty at Università di Milano
- in 1927 he got a degree in Applied Mathematics, with a thesis on affine geometry
- after his graduation, he started to work at what is now ISTAT, the Italian central institute of Statistics, until 1931, when he started to work as an actuary in Trieste (Italy) at the Assicurazioni Generali, and started to teach at the university there (he had also qualified in a competition as university lecturer)
- in 1947 he obtained his chair as full professor
- in 1954 he moved to Università La Sapienza, Rome (Italy); he retired in 1976
- he died on 20 July 1985
- 1928: de Finetti presents his first results on exchangeability at the World Meeting of Mathematicians in Bologna
- 1935: lectures at the Institut Poincaré, Paris (France); de Finetti (1937)

Definition. *The sequence* $(X_n)_{n\geq 1}$ *is* exchangeable *if*

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{\pi(1)}, \dots, X_{\pi(n)})$$

for any $n \geq 1$ *and permutation* π *of* $(1, \dots, n)$.

Definition. *The sequence* $(X_n)_{n\geq 1}$ *is* exchangeable *if*

$$(X_1, \dots, X_n) \stackrel{\mathrm{d}}{=} (X_{\pi(1)}, \dots, X_{\pi(n)})$$

for any $n \geq 1$ *and permutation* π *of* $(1, \dots, n)$.

- According to this definition, the order with which data are recorded is irrelevant for inferential purposes

Definition. *The sequence* $(X_n)_{n\geq 1}$ *is* exchangeable *if*

$$(X_1,\dots,X_n) \stackrel{d}{=} (X_{\pi(1)},\dots,X_{\pi(n)})$$

for any $n \geq 1$ *and permutation* π *of* $(1,\dots,n)$.

- According to this definition, the order with which data are recorded is irrelevant for inferential purposes
- it is a *weak* assumption, and translates lack of (enough) information through a condition of *symmetry*

Definition. *The sequence* $(X_n)_{n\geq 1}$ *is* exchangeable *if*

$$(X_1, \dots, X_n) \stackrel{\mathrm{d}}{=} (X_{\pi(1)}, \dots, X_{\pi(n)})$$

for any $n \geq 1$ *and permutation* π *of* $(1, \dots, n)$.

- According to this definition, the order with which data are recorded is irrelevant for inferential purposes
- it is a *weak* assumption, and translates lack of (enough) information through a condition of *symmetry*
- For example, in coin-tossing sequence one would have

$$\mathbb{P}[X_1 = 0,\ X_2 = 1,\ X_3 = 1,\ X_4 = 0] = \mathbb{P}[X_1 = 1,\ X_2 = 0,\ X_3 = 0,\ X_4 = 1]$$

Critical aspects of the Bayesian paradigm according to de Finetti (Regazzini, 2015)

De Finetti gave his first results on exchangeability (he used the term equivalence) for sequences of trials on a given phenomenon, all made under analogous conditions, i.e. $(X_n)_n$ 0-1 r.v.'s, $X_n = 1$ if the n-th trial was a "success"

Critical aspects of the Bayesian paradigm according to de Finetti (Regazzini, 2015)

De Finetti gave his first results on exchangeability (he used the term equivalence) for sequences of trials on a given phenomenon, all made under analogous conditions, i.e. $(X_n)_n$ 0-1 r.v.'s, $X_n = 1$ if the n-th trial was a "success"

✓ Phenomena whose trials are independent with a fixed but unknown probability distribution (p.d.) are exchangeable, since the law of $(X_n)_n$ is a mixture of Bernoulli r.v.'s.

Critical aspects of the Bayesian paradigm according to de Finetti (Regazzini, 2015)

De Finetti gave his first results on exchangeability (he used the term equivalence) for sequences of trials on a given phenomenon, all made under analogous conditions, i.e. $(X_n)_n$ 0-1 r.v.'s, $X_n = 1$ if the n-th trial was a "success"

✓ Phenomena whose trials are independent with a fixed but unknown probability distribution (p.d.) are exchangeable, since the law of $(X_n)_n$ is a mixture of Bernoulli r.v.'s.

✓ This statement is controversial according to de Finetti's subjectivist conception of probability; the reference to an unknown probability is devoid of sense and, in any case, obscure and specious. In addition, it requires the specification of a law for the fixed, though unknown, p.d., and, under the subjective point of view, such request is of an ambiguous content.

In more modern terms: if $X_i|\theta \stackrel{\text{i.i.d.}}{\sim} Be(\theta)$, $\theta \sim F$ and the approach taken is subjective, then

$$\theta \sim F \text{ is ambiguous}$$

The role of exchangeability according to de Finetti (Regazzini, 2015)

✓ F, the unknown p.d. of θ, cannot represent the subjective belief unless θ has an objective (i.e. "physical") meaning, that is, unless θ represents a well-specified characteristic of the members of a statistical population.

The role of exchangeability according to de Finetti (Regazzini, 2015)

✓ F, the unknown p.d. of θ, cannot represent the subjective belief unless θ has an objective (i.e. “physical") meaning, that is, unless θ represents a well-specified characteristic of the members of a statistical population.

Two examples:

- the sequence of drawings with replacement from an urn containing white and nonwhite balls according to an unknown composition: $X_n = 1$ if the n-th drawn ball is white, θ is the unknown proportion of white ball.
- the sequence of tossing of the same coin: $X_n = 1$ if the n-th toss is H, θ is the unknown probability of H in each toss.

✓ F, the unknown p.d. of θ, cannot represent the subjective belief unless θ has an objective (i.e. “physical") meaning, that is, unless θ represents a well-specified characteristic of the members of a statistical population.

Two examples:

- the sequence of drawings with replacement from an urn containing white and nonwhite balls according to an unknown composition: $X_n = 1$ if the n-th drawn ball is white, θ is the unknown proportion of white ball.
  The composition is empirically verifiable!
- the sequence of tossing of the same coin: $X_n = 1$ if the n-th toss is H, θ is the unknown probability of H in each toss. The probability of H cannot be verified!

The role of exchangeability according to de Finetti (Regazzini, 2015)

✓ F, the unknown p.d. of θ, cannot represent the subjective belief unless θ has an objective (i.e. “physical") meaning, that is, unless θ represents a well-specified characteristic of the members of a statistical population.

Two examples:

- the sequence of drawings with replacement from an urn containing white and nonwhite balls according to an unknown composition: $X_n = 1$ if the n-th drawn ball is white, θ is the unknown proportion of white ball.
  The composition is empirically verifiable!
- the sequence of tossing of the same coin: $X_n = 1$ if the n-th toss is H, θ is the unknown probability of H in each toss. The probability of H cannot be verified!

✓Unlike these formulations, which are unfortunate and unclear in some respect, the condition of exchangeability they met is more general, always meaningful and sensible. De Finetti selected it to indicate a sufficiently general setting in which he might prove the validity of the principle of induction:

in a sequence of homogeneous trials, the frequency distribution of the results of n past trials can represent a good approximation, if n is “large”, to the conditional distribution of the outcome of a future trial, given the observed frequency distribution.

Theorem. *The sequence $X^{(\infty)} = (X_n)_{n\geq 1}$ of 0-1 r.v.s's is exchangeable if and only if there exists a probability measure F on $([0,1], \mathscr{B}([0,1]))$ such that*

$$P\left[X_1 = x_1, \ldots, X_n = x_n\right] = \int_{[0,1]} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i} F(\mathrm{d}\theta)$$

for any $n \geq 1$ and $(x_1, \ldots, x_n)$ in $\{0,1\}^n$.

Theorem. *The sequence* $X^{(\infty)} = (X_n)_{n\geq 1}$ *of 0-1 r.v.s's is exchangeable if and only if there exists a probability measure* F *on* $([0,1], \mathscr{B}([0,1]))$ *such that*

$$P\left[X_1 = x_1, \ldots, X_n = x_n\right] = \int_{[0,1]} \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} F(\mathrm{d}\theta)$$

for any $n \geq 1$ *and* $(x_1, \ldots, x_n)$ *in* $\{0,1\}^n$.

Moreover, when $(X_n)_{n\geq 1}$ is exchangeable (since P is σ-additive):

- $\dfrac{\sum_1^n X_i}{n} \overset{a.s.}{\to} \tilde{\theta} \sim F$ as $n \to +\infty$.
- Conditionally on $\tilde{\theta}$, $\quad X_1, \ldots, X_n | \tilde{\theta} \overset{\text{i.i.d.}}{\sim} Be(\tilde{\theta})$ for all n
  $\tilde{\theta} \sim F$.

✓ Therefore, there is a formal equivalence between exchangeable trials of the same phenomenon (i.e. X_n's are binary) and those trials that are designated as "independent, with a fixed, but unknown, probability".

✓ Therefore, there is a formal equivalence between exchangeable trials of the same phenomenon (i.e. X_n's are binary) and those trials that are designated as "independent, with a fixed, but unknown, probability".

✓ The representation theorem may be used to justify the principle of induction with reference to an excheangeable sequence of events:

If $(X_n)_n$ are exchangeable 0-1 r.v.'s, and $\varphi_n := \sum_1^n X_i/n$, then

$$\left| P\left(X_{n+1} = e_1, \ldots, , X_{n+k} = e_k | X_1, \ldots, X_n\right) - \varphi_n^{\sum_1^k e_j} (1-\varphi_n)^{k-\sum_1^k e_j} \right| \overset{\text{a.s.}}{\to} 0, \; n \to +\infty,$$

for all $(e_1, \ldots, e_k) \in \{0,1\}^k$, $k = 1, 2, 3, \ldots$.

✓ Therefore, there is a formal equivalence between exchangeable trials of the same phenomenon (i.e. X_n's are binary) and those trials that are designated as "independent, with a fixed, but unknown, probability".

✓ The representation theorem may be used to justify the principle of induction with reference to an excheangeable sequence of events:

If $(X_n)_n$ *are exchangeable 0-1 r.v.'s, and* $\varphi_n := \sum_1^n X_i/n$, *then*

$$\left| P\left(X_{n+1} = e_1, \ldots, , X_{n+k} = e_k | X_1, \ldots, X_n\right) - \varphi_n^{\sum_1^k e_j} (1-\varphi_n)^{k-\sum_1^k e_j} \right| \stackrel{\text{a.s.}}{\rightarrow} 0, \; n \to +\infty,$$

for all $(e_1, \ldots, e_k) \in \{0,1\}^k$, $k = 1, 2, 3, \ldots$.

✓ Roughly speaking its says that:

the frequency of success in n past trials can be used to evaluate the probability distribution of "future" trials of the same phenomenon, if n is large.

- ✓ De Finetti's re-interpretation of the Bayesian approach is strictly linked to the subjective notion of probability (i.e. definition of probability of an event through coherence): the probability p of an event E is the personal degree of belief in the event; p has to ensure that there exists no real value c such that any bet on E with gain $c(p - \mathbf{1}_E)$ has realizations of the gain all strictly positive or all strictly negative
- ✓ De Finetti's original approach to exchangeability was different from what I have introduced here; see Cifarelli and Regazzini (1996) for an historical picture of de Finetti's contributions
- ✓ exchangeability is a term introduced by Pólya; de Finetti used *eventi equivalenti* (equivalent events), other authors used symmetric
- ✓ the infinite exchangeability is a keypoint here: finite exchangeable sequences have different representations.

Theorem. *The sequence* $X^{(\infty)} = (X_n)_{n\geq 1}$ *is exchangeable if and only if there exists a probability measure on* $(\mathbf{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$ *such that*

$$P\left[X_1 \in A_1, \dots, X_n \in A_n\right] = \int_{\mathbf{P}_{\mathbb{X}}} \prod_{i=1}^{n} p(A_i)\, Q(\mathrm{d}p)$$

for ay $n \geq 1$ *and* $A_1, \dots, A_n$ *in* $\mathscr{X}$, *where the probability* Q *is uniquely determined.*

Theorem. *The sequence $X^{(\infty)} = (X_n)_{n \geq 1}$ is exchangeable if and only if there exists a probability measure on $(\mathbf{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$ such that*

$$P\left[X_1 \in A_1, \ldots, X_n \in A_n\right] = \int_{\mathbf{P}_{\mathbb{X}}} \prod_{i=1}^{n} p(A_i)\, Q(\mathrm{d}p)$$

for ay $n \geq 1$ and $A_1, \ldots, A_n$ in $\mathscr{X}$, where the probability Q is uniquely determined.

✏ $(X_n)_{n \geq 1}$ is exchangeable if and only if there exists a random probability measure $\tilde{p}$ on $(\mathbb{X}, \mathscr{X})$ such that $\tilde{p} \sim Q$ and

$$\mathbb{P}\left[X_1 \in A_1, \ldots, , X_n \in A_n \,|\, \tilde{p}\right] = \prod_{i=1}^{n} \tilde{p}(A_i)$$

for any $n \geq 1$ and $A_1, \ldots, A_n$ in $\mathscr{X}$.

Q is a probability measure on $\mathbf{P}_{\mathbb{X}}$ $\longrightarrow$ *de Finetti measure* of $(X_n)_{n\geq 1}$

✏ If $(X_n)_{n\geq 1}$ is exchangeable, then its empirical distribution is such that

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \Rightarrow \tilde{p} \qquad \text{a.s.--}\mathbb{P}$$

where $\Rightarrow$ denotes *weak convergence*.

Q is a probability measure on $\mathbf{P}_{\mathbb{X}}$ $\longrightarrow$ *de Finetti measure* of $(X_n)_{n\geq 1}$

✏ If $(X_n)_{n\geq 1}$ is exchangeable, then its empirical distribution is such that

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \Rightarrow \tilde{p} \qquad \text{a.s.--}\mathbb{P}$$

where $\Rightarrow$ denotes *weak convergence*.

Hierarchical representation: $(X_n)_{n\geq 1}$ exchangeable is equivalent to

$$\begin{aligned} X_i \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim Q \\ Q &= \text{prior distribution} \end{aligned}$$

de Finetti's representation theorem (general case) (1933) - continued

Q is a probability measure on $\mathbf{P}_{\mathbb{X}}$ $\longrightarrow$ *de Finetti measure* of $(X_n)_{n\geq 1}$

✏ If $(X_n)_{n\geq 1}$ is exchangeable, then its empirical distribution is such that

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \Rightarrow \tilde{p} \qquad \text{a.s.–}\mathbb{P}$$

where $\Rightarrow$ denotes *weak convergence*.

Hierarchical representation: $(X_n)_{n\geq 1}$ exchangeable is equivalent to

$$\begin{aligned} X_i \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim Q \\ Q &= \text{prior distribution} \end{aligned}$$

The Bayesian nonparametric framework is equivalent to exchangeability of $(X_n)_n$

Parametric model: Q degenerate on a finite–dimensional subset $\mathbf{P}^*_{\mathbb{X}}$ of $\mathbf{P}_{\mathbb{X}}$, such that

$$Q(\{\mathbf{P}^*_{\mathbb{X}}\}) = Q(\{p \in \mathbf{P}_{\mathbb{X}} :\ p = p_\theta, \theta \in \Theta\}) = 1$$

and and there exists a function

$$\tilde{\theta} : \mathbf{P}^*_{\mathbb{X}} \to \Theta \text{ bijective.}$$

$\Theta \subset \mathbb{R}^p$ is called parametric space. The prior Q induces a probability on Θ:

$$\pi(B) = Q(\tilde{\theta}^{-1}(B)), B \in \mathscr{B}(\Theta).$$

Parametric model: Q degenerate on a finite–dimensional subset $\mathbf{P}^*_{\mathbb{X}}$ of $\mathbf{P}_{\mathbb{X}}$, such that

$$Q(\{\mathbf{P}^*_{\mathbb{X}}\}) = Q(\{p \in \mathbf{P}_{\mathbb{X}} : p = p_\theta, \theta \in \Theta\}) = 1$$

and and there exists a function

$$\tilde{\theta} : \mathbf{P}^*_{\mathbb{X}} \to \Theta \text{ bijective.}$$

$\Theta \subset \mathbb{R}^p$ is called parametric space. The prior Q induces a probability on Θ:

$$\pi(B) = Q(\tilde{\theta}^{-1}(B)), B \in \mathscr{B}(\Theta).$$

In these cases:

$$\begin{aligned} X_i \mid \tilde{\theta} = \theta &\overset{\text{i.i.d.}}{\sim} p_\theta \\ \tilde{\theta} &\sim \pi \text{ prior distribution} \end{aligned}$$

For instance:

$$Q(\{p \in \mathbf{P}_{\mathbb{X}} : p(\mathrm{d}x) = \varphi((x-\mu)/\sigma)\,\mathrm{d}x,\ (\mu,\sigma) \in \mathbb{R} \times \mathbb{R}^+\}) = 1$$

with φ being the density function of a $\mathrm{N}(0,1)$ distribution.

When can we assume $Q(\{\mathbf{P}^*_{\mathbb{X}}\}) = 1$, where $\mathbf{P}^*_{\mathbb{X}}$ is finite-dimensional? More clearly, when could we assume the model is parametric?

When can we assume $Q(\{\mathbf{P}^*_{\mathbb{X}}\}) = 1$, where $\mathbf{P}^*_{\mathbb{X}}$ is finite-dimensional? More clearly, when could we assume the model is parametric?

- if, from past experience in cases similar to the one analized, we believe that the parametric family approximates well the “true” distribution

When can we assume $Q(\{\mathbf{P}^*_{\mathbb{X}}\}) = 1$, where $\mathbf{P}^*_{\mathbb{X}}$ is finite-dimensional? More clearly, when could we assume the model is parametric?

- if, from past experience in cases similar to the one analized, we believe that the parametric family approximates well the "true" distribution
- if, in addition to exchangeability, we assume different conditions for the sequence of observations. For example, if $(X_n)_{n\geq 1}$ is also spherically symmetric ($\mathcal{L}(X_1, \ldots, X_n)^T = \mathcal{L}(A(X_1, \ldots, X_n)^T)$ for any orthogonal matrix A), then $\mathbf{P}^*_{\mathbb{X}}$ is the family of Gaussian distributions with 0-mean.

When can we assume $Q(\{\mathbf{P}_{\mathbb{X}}^*\}) = 1$, where $\mathbf{P}_{\mathbb{X}}^*$ is finite-dimensional? More clearly, when could we assume the model is parametric?

- if, from past experience in cases similar to the one analized, we believe that the parametric family approximates well the "true" distribution
- if, in addition to exchangeability, we assume different conditions for the sequence of observations. For example, if $(X_n)_{n\geq 1}$ is also spherically symmetric ($\mathcal{L}(X_1, \ldots, X_n)^T = \mathcal{L}(A(X_1, \ldots, X_n)^T)$ for any orthogonal matrix A), then $\mathbf{P}_{\mathbb{X}}^*$ is the family of Gaussian distributions with 0-mean.

Otherwise: nonparametric model
$\longrightarrow$ greater flexibility when Q has large support, possibly $\mathrm{supp}(Q) = \mathbf{P}_{\mathbb{X}}$.

Acknowledgments

Many thanks to:

✓ Eugenio Regazzini, who suggested the material I used

✓ Antonio Lijoi for providing the slides with the notation and representation theorems

Many thanks to:

✓ Eugenio Regazzini, who suggested th

✓ Antonio Lijoi for providing the slides theorems

Grazie!

✏ Cifarelli, Regazzini (1996). *De Finetti's contribution to probability and statistics*. Statistical Science.

✏ de Finetti (1937). *La prevision: ses lois logiques, ses sources subjectives*. Ann. Inst. H. Poincaré.

✏ Regazzini (1996). *Impostazione non parametrica di problemi di inferenza statistica bayesiana*. TR IAMI 96.21.

✏ Regazzini (2015). *De Finetti's representation theorem*. Wiley StatsRef: Statistics Reference Online.