LECTURE 1

INTRODUCTION

. . . practitioners seem to prefer the language of populations; theoreticians, that of exchangeability
Lindley and Novick, 1981

1.1. Unveiling the uncertainty: An attempt to motivate Statistics in a short lecture

Random phenomena drive many aspects of this world, including life itself, the global economy, weather conditions, and more. How can we improve our chances of having a better life, prepare for an unfortunate contingency, or simply have a wealthier bank account? It all starts with the proper quantification of the uncertainties driving such phenomena. By understanding and quantifying such uncertainties, one might be able to establish a robust statistical induction process that allows us to learn, uncover, and predict important features about the phenomena under study. This is precisely the main challenge addressed by the theories of probability and statistics.

As interpreted by some authors, e.g. Goldstein (2013), the uncertainty of random phenomena can be due to aleatory and epistemic causes. ***Epistemic uncertainty*** is that related to lack of information, that can potentially be resolved by the arrival of new information, whereas ***aleatory uncertainty*** relates to the variational nature intrinsic to the phenomenon or targeted population under study. This classification is directly related to the potential scope of a statistical induction process, e.g. while an air pollution contingency day could be potentially avoided or predicted, a hurricane cannot be handled with the same degree of certainty. See Figure 1.1. To some extent, such a classification is not precise, as there is the line of thinking that supports that all uncertainty concerns with a lack of information, uncertainty that we need to quantify and can have a finite or infinite dimensional nature. We will come back to this point later.

From the mathematical perspective the theory of statistics uses probabilistic models to provide a solution to the problem of statistical induction on random phenomena. Indeed, the basic mathematical setup starts with a triplet $(\Omega, \mathcal{A}, \mathbb{P})$, termed ***probability space***, where Ω denotes the set of all possible outcomes of the phenomenon or experiment under study, here termed the ***sample space***; $\mathcal{A}$ is a collection of subsets of Ω including all events of interest which constitutes a σ-***field***[1], and $\mathbb{P} : \mathcal{A} \mapsto [0, 1]$ termed ***probability measure*** which satisfies $\mathbb{P}(A) \geq 0$ for all ***events*** $A \in \mathcal{A}$ with $\mathbb{P}(\Omega) = 1$ and

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

for disjoint events $A_1, A_2, \ldots \in \mathcal{A}$. In words, $\mathbb{P}(\cdot)$ constitutes a *coherent* way to quantify all uncertain events of interest.

Within such a framework, a feature of interest, regarding the underlying random phenomenon or experiment, can be translated into "numeric" quantities through $(\mathbb{X}, \mathcal{X})$-valued functions, $X : \Omega \mapsto \mathbb{X}$, termed ***random variables (r.v.'s)***, satisfying $X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$ for every $B \in \mathcal{X}$. This latter condition states that the event at issue is a true event, element of $\mathcal{A}$, that can be measured by $\mathbb{P}$, i.e. X is simply a ***measurable function***. Here $(\mathbb{X}, \mathcal{X})$ is also a

[1] (i) $\emptyset \in \mathcal{A}$, (ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ and (iii) $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{A}$. We can always consider $\mathcal{A}$ sufficiently large to include all events to which we are willing to compute probabilities

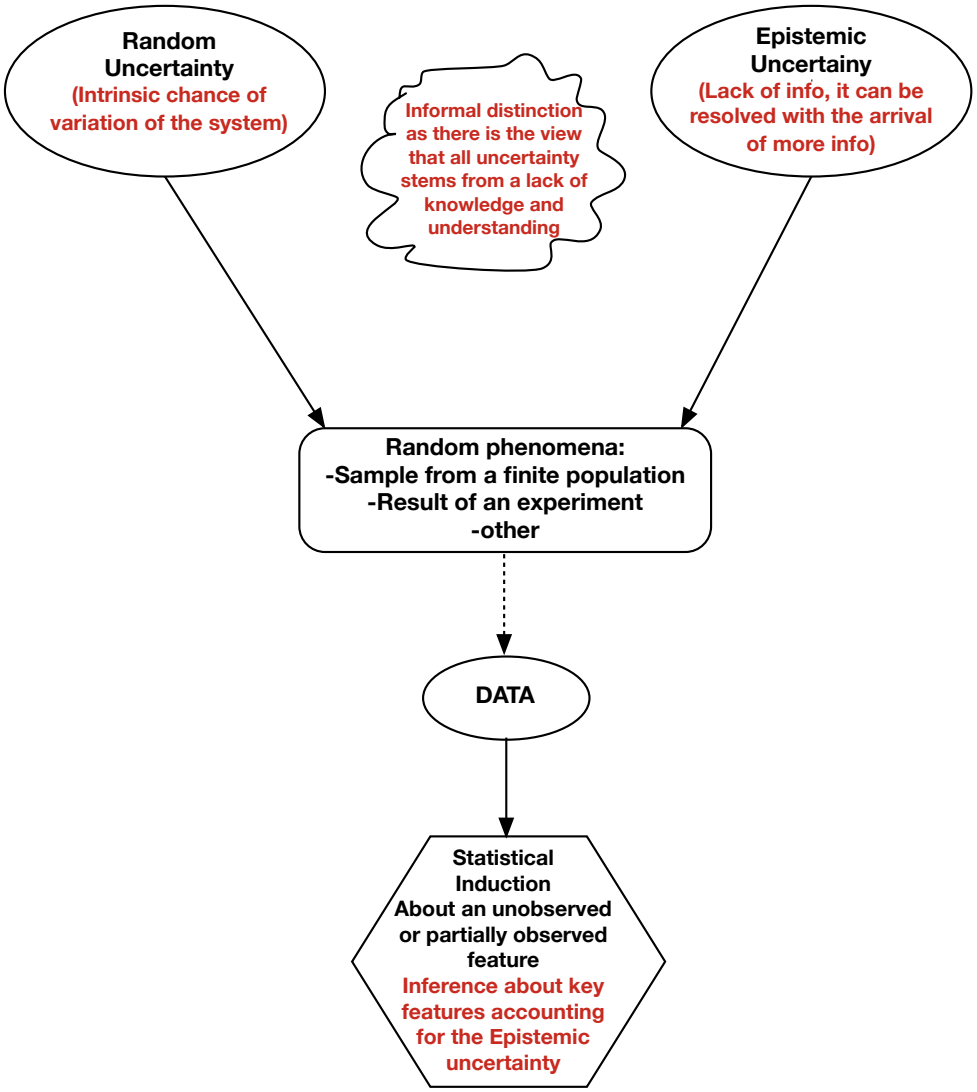

Figure 1.1: Process of statistical induction

measure space representing the possible values the random variable X can take and it is endowed with its Borel σ-field, $\mathcal{X}$. It is worth noting that the nature of $\mathbb{X}$ gives rise to a classification for the generic term of random variable, e.g. random vectors, random variables, random manifolds, random functions, etc.

Given a random variable X one can define the set function

$$\mathsf{P}_X(B) = \mathbb{P}(X^{-1}(B)), \qquad \text{for all } B \in \mathcal{X} \tag{1.1}$$

known generically as the ***distribution or law*** of the random variable X. In particular, when $\mathbb{X} = \mathbb{R}$ and $B = (-\infty, x]$ then we write $\mathsf{F}_X(x) = \mathsf{P}_X((-\infty, x]) = \mathbb{P}(X \leq x)$ and name it the ***cumulative distribution function (cdf)*** of X. In the most common settings, one assumes that (1.1) is absolutely continuous with respect to a σ-finite reference measure on $(\mathbb{X}, \mathcal{X})$, e.g. λ, typically the Lebesgue measure, or with respect to the counting measure. In such cases we will work with the corresponding ***density function*** or ***mass probabilities***, generically denoted by

$$f_X(x) = \frac{\mathrm{d}\mathsf{P}_X}{\mathrm{d}\lambda}(x)$$

Example 1. *A didactic and historical example to introduce such concepts is to consider the outcome of flipping a coin. This has two possible outcomes, "head" or "tail". Hence we can construct* $\Omega = \{head, tail\} = \{\omega_1, \omega_2\} = \{0, 1\}$ *with* $\mathcal{A} = \{\Omega, \{0\}, \{1\}, \emptyset\}$. *Let* X *the random variable that assigns* 1 *if the outcome is tail and* 0 *otherwise. Hence* $\mathsf{P}_X(\{1\}) = \mathbb{P}(X(\omega_1) = 1)$. *For this later quantity, we might assign a value, say* $\theta \in [0, 1]$, *i.e.* $\mathsf{P}_X(\{1\}) = \theta$.

In the above example, clearly the uncertainty underlying X, quantified through a probability measure, has been transferred to θ, i.e. the ***parameter*** of interest. Here the notion of statistics arises naturally, namely, how can we improve our knowledge about θ in the prescience of observations from the random phenomena, here modelled through the random variable X. Historically, a way to resolve this quantification of uncertainty problem is through the assumption of a ***parametric family*** of probability measures

$$\mathcal{P}_{\mathbb{O}} = \{\mathsf{P}_\theta : \forall B \in \mathcal{X}, \mathsf{P}_\theta(B) = \mathbb{P}(X \in B \mid \Theta = \theta), \text{ for } \theta \in \mathbb{O}.\} \tag{1.2}$$

Here, $\mathbb{O}$ denotes an arbitrary topological space, typically a finite or countable set, an Euclidean space or a Polish space.

In this way, n observations of the random phenomena under study could be assumed as distinct realizations of the random variable X, namely $X_1, \ldots, X_n$. Therefore, under the assumption that no causality alters the observations among each other one typically says that observations are independent. This latter statement has to be taken with care, having observations with no-physical relation, say ***physically independent***, does not necessarily imply that the set of random variables $X_1, \ldots, X_n$, which model such observations, are ***stochastically independent***. Indeed, statistical learning among observations demands stochastic dependence among the r.v.'s modelling replications of the random phenomena. In other words, statistical learning can only take place if $X_1, \ldots, X_n$ are stochastically connected. All that the physical independence among observations implies is a symmetry among $X_1, \ldots, X_n$, namely the law of these random variables is not affected by the order in which they are sampled. See Walker (2013) for a similar way of reasoning.

Hence, when one tries to apply a probabilistic model to the uncertainties mentioned at the beginning of this section, cf. Figure 1.1, the aleatory uncertainty can be interpreted as the parametric family, $\mathcal{P}_{\mathbb{O}}$, assumed as the universe of all available models. In other words, within such a setting, one can only learn about the only remaining source of variation, i.e. the parameter, $\theta \in \mathbb{O}$. The parametric form driving X, (1.2), has been taken for granted. On the other hand, the epistemic uncertainty, represented by θ, could be resolved with the arrival of more information.

In ***classical statistical inference*** one typically assumes a parametric family, as in (1.2), to model the random uncertainty, and transfer all epistemic uncertainty to the value for the parameter θ, which is assumed to be fixed but unknown. Therefore, data-based inference about θ, resolves the controversy about which $\mathsf{P} \in \mathcal{P}_{\mathbb{O}}$ fits best to the data but provides no learning beyond $\mathcal{P}_{\mathbb{O}}$. As it will be clarified later, this also relates to the fact that, within this approach,

$$\mathbb{P}(X_{n+1} \in B \mid X_1, \ldots, X_n) = \mathsf{P}_X(X_{n+1} \in B),$$

that is, there is no statistical learning in the model. It is worth stressing that stochastic dependence does not necessarily compromises the potential physical-independence inherent to the actual observations.

In principle, one could widen the space $\mathcal{P}_{\mathbb{O}}$ to the extreme case of considering the ***space of all probability measures on*** $(\mathbb{X}, \mathcal{X})$, here denoted by $\mathcal{P}_{\mathbb{X}}$, namely one could consider $X_1, \ldots, X_n$ ***independently and identically distributed (iid)*** from an unknown distribution P, however parametrising such a space via a finite dimensional parameter is not possible. In particular, this latter case would result in the lack of likelihood function and thus, estimators such as the ***maximum likelihood estimator (MLE)***, would not be available.

Depending of the nature of the random phenomena under study, observations or measurements might be physically independent or might have some clear dependence, e.g. via and ordered index set such as time series observations collected from a phenomenon evolving in time. This poses a natural tradeoff between the assumed stochastic dependence among the r.v.'s, modelling such phenomena, and the choice and/or generality of (1.2). As mentioned above, when no clear physical dependence is present, the natural stochastic connection is that which is invariant under different sampling order.

1.2. Exchangeability as a statistical learning model

If the point of departure is the probabilistic modeling of observations physically independent through a sequence of r.v.'s $(X_1, \ldots, X_n)$, without making further assumptions about the model then the proper probabilistic framework is the one provided by the distributional symmetry.

Definition 1. A finite set $\{X_i\}_{i=1}^n$ of random variables is said to be ***finite exchangeable*** if

$$\{X_1, \ldots, X_n\} \overset{\mathrm{d}}{=} \{X_{\tau(1)}, \ldots, X_{\tau(n)}\}$$

for any permutation τ of $\{1, \ldots, n\}$. An infinite collection $\{X_i\}_{i=1}^\infty$ is said to be ***exchangeable*** if every subcollection is exchangeable.

The above definition only says that exchangeability equals to distributional invariance under label permutations. The following result constitutes an historically didactic example

Theorem 1. *(de Finetti's Representation Theorem for binary variables)*
An infinite sequence of $\mathbb{X} := \{0, 1\}$*-valued random variables,* $\{X_i\}_{i=1}^\infty$*, is said to be exchangeable if and only if there exist a distribution* q *on* $[0, 1]$ *such that for all* $n \geq 1$

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_{[0,1]} \theta^{s_n} (1 - \theta)^{n - s_n} \, q(\mathrm{d}\theta)$$

where $s_n := \sum_{i=1}^n x_i$ *denotes the number of sucesses. Furthermore,* q *is such that its cdf is*

$$q(t) = \lim_{n \to \infty} \mathbb{P}\left(\frac{S_n}{n} \leq t \right).$$

Namely,

$$\frac{S_n}{n} \overset{a.s.}{\to} \Theta \qquad \textit{with} \qquad \Theta := \lim_{n \to \infty} \frac{S_n}{n},$$

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

Proof. (Heath and Sudderth (1976)) By exchangeability, for $0 \leq s_n \leq n$

$$\mathbb{P}(S_n = s_n) = \binom{n}{s_n} \mathbb{P}(x_1, \ldots, x_n) = \binom{n}{s_n} \mathbb{P}(x_{\tau(1)}, \ldots, x_{\tau(n)})$$

for any permutation τ of $\{1, \ldots, n\}$. Namely, for the $\binom{n}{s_n}$ vectors of n trials with exactly s_n successes (1s) and $n - s_n$ failures (0s) we have the same probability. This clearly simplifies the specification of the 2^n potential probabilities, i.e. that would be potentially different without the assumption of exchangeability.

Also, due to exchangeability, for any N such that $N \geq n \geq s_n \geq 0$

$$\mathbb{P}[S_n = s_n] = \sum_{(*)} \mathbb{P}[S_n = s_n \mid S_N = s_N]\mathbb{P}[S_N = s_N]$$

where $\mathbb{P}[S_N = s_N]$ stands for the probability that the total number of successes, S_N, is equal to s_N, and the range of the summation $(*)$ extends over $(s_n, \ldots, N-(n-s_n))$. Now

$$\begin{aligned}\mathbb{P}[S_n = s_n \mid S_N = s_N] &= \frac{\binom{s_N}{s_n}\binom{N-s_N}{n-s_n}}{\binom{N}{n}}, \qquad 0 \leq s_n \leq n \\ &= \frac{(s_N)_{s_n}(N-s_N)_{n-s_n}}{(N)_n},\end{aligned}$$

that is the ***hypergeometric*** mass function, $\mathsf{Hyp}(N, s_N, n)$. In the second equality above we have used the notation with the ***descending factorial*** (or ***Pochhammer symbol***) $(x)_r = x(x-1)\cdots(x-r+1)$. Hence

$$\mathbb{P}[S_n = s_n] = \binom{n}{s_n} \sum \frac{(s_N)_{s_n}(N-s_N)_{n-s_n}}{(N)_n} \mathbb{P}[S_N = s_n]$$

Define the function $q_N(\theta)$ on $\mathbb{R}$ as the step function which is zero for $\theta < 0$, with height $\mathsf{P}[S_N = s_n]$ at $\theta = s_n/N$. Hence, using Lebesgue integral notation we have

$$\mathbb{P}[S_n = s_n] = \binom{n}{s_n} \int_0^1 \frac{(\theta N)_{s_n}((1-\theta)N)_{n-s_n}}{(N)_n} \mathrm{d}q_N(\theta)$$

So, letting $N \to \infty$ we get

$$\frac{(\theta N)_{s_n}((1-\theta)N)_{n-s_n}}{(N)_n} \to \theta^{s_n}(1-\theta)^{n-s_n} = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

since $(x)_r \to x^r$ when $x \to \infty$ for r fixed. Noticing that $q_N(t)$ is a step function with N steps of varying sizes particular values of t. Applying Helly Theorem , there is a sequence $\{q_N(\theta); N = 1, 2, \ldots\}$ that has a convergent subsequence $\{q_{N_j}(\theta)\}$ such that, for some distribution q

$$\lim_{j\to\infty} q_{N_j}(\theta) = q_\theta$$

Then the result follows. □

The representation theorem by de Finetti, changes the basic idea that probabilities are just frequencies of an infinite number of observations. Indeed, Theorem (1) says that the quantity that allows us to see the probabilities of distinct replicates of our random phenomena as independent is indeed random. More specifically, the limit of frequencies, i.e. Θ, is only a conditional probability of information not yet available. This supports the idea that such "probabilities of success", under the exchangeability assumption, can be computed ***subjectively***. In order to set ideas, let us borrow the following example from Schervish (1995), from where we have also stolen part of this discussion.

Example 2. *Let $\{X_i\}_{i=1}^{\infty}$ be Bernoulli r.v.'s and consider the two different exchangeable laws, corresponding to two different persons, given by*

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{12}{s_n+2}\frac{1}{\binom{n+4}{s_n+2}} \quad \textit{and} \quad \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{1}{[n+1]\binom{n}{s_n}},$$

respectively, with $s_n = \sum_{j=1}^{n} x_j$, for all $n \geq 1$. Hence, the first person believes that $\mathbb{P}(X_1 = 1) = 0.4$ and the second that $\mathbb{P}(X_1 = 1) = 0.5$. Now, due to exchangeability, both persons believe that $\Theta := \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i$ exists almost surely

and that $\mathbb{P}(X_1 = 1 \mid \Theta = \theta) = \theta$. *But, as we will see later* $\mathbb{P}(X_1 = 1) = \mathbb{E}(\Theta)$ *and so these persons must have different values for* $\mathbb{E}(\Theta)$.

Now assume that both persons observe the result of $n = 20$ *given by 14 " 1s " and 6 " 0s ". Then*

$$\mathbb{P}[X_{21} \mid x_1, \ldots, x_{20}] = 0.64 \quad \textit{and} \quad \mathbb{P}[X_{21} \mid x_1, \ldots, x_{20}] = 0.68.$$

Notice, how the difference shortens after using some observations and at the same time gets closer to the proportion of successes, regardless of the prior mean of Θ. *This tells us that we should modify our opinion, about the proportion of successes, after observations are performed. Such a consequence is merely due to exchangeability, regardless of frequencies being interpreted as probabilities.*

We have stated de Finetti's representation theorem in its simpler form, namely when $\mathbb{X} = \{0, 1\}$, before further discussing the consequences of such as celebrated result, let us state it in its more general form, i.e. when $\mathbb{X}$ is Polish. As before, denote by $\mathcal{P}_{\mathbb{X}}$ the space of all probability measures on $(\mathbb{X}, \mathcal{X})$, endowed with the σ-algebra $\mathscr{P}_{\mathbb{X}}$, generated by the topology of weak convergence.

Theorem 2. *Let* $\mathbb{X}$ *be a Polish space endowed with is Borel* σ*-field* $\mathcal{X}$ *and* $\mathcal{P}_{\mathbb{X}}$ *the space of all probability measures on* $(\mathbb{X}, \mathcal{X})$. *An infinite sequence of* $\mathbb{X}$*-valued r.v.'s,* $\{X\}_{i=0}^{\infty}$ *is exchangeable if and only if there exist* Q *on* $(\mathcal{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$ *such that*

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n] = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} \mathsf{P}(A_i) \mathsf{Q}(\mathrm{d}\mathsf{P}) \tag{1.3}$$

for all $n \geq 1$ *and* $A_i \in \mathcal{X}$.

For a proof of this general version of de Finetti's representation theorem we refer to Schervish (1995).

Theorem 2 can be alterantively stated as the existence of a $(\mathcal{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$-valued random variable P, distributed as Q, such that

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n \mid \mathsf{P}] = \prod_{i=1}^{n} \mathsf{P}(A_i)$$

namely, conditional on , the random variables $\{X\}_{i=0}^{\infty}$ are iid. Clearly, due to the potential uncountable nature of $\mathbb{X}$, the form of the conditional probabilities cannot always be simplified as in the $\mathbb{X} = \{0, 1\}$ case, where the random probability measure (RPM) P is equivalent to the random variable $\Theta \in [0, 1]$, where $\Theta = \mathsf{P}(\{1\})$. As in the binary case, it can be seen that, if we let

$$\mathsf{P}_n(A) := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A), \qquad A \in \mathcal{X},$$

be the ***empirical measure***, hence the measure Q is the distribution of the RPM, P, where $\mathbb{P}(\mathsf{P}_n \Rightarrow \mathsf{P}) = 1$, with $\Rightarrow$ denoting convergence in distribution. Furthermore, given an infinite sequence of exchangeable random variables, Q is unique. The distribution Q is sometimes referred to as the ***de Finetti's measure*** driving an exchangeable sequence.

The above discussion of Theorem 1 also follows in this case, thus, it serves as a justification of the ***subjective view of Bayesian statistics***. In other words, exchangeability justifies the existence of a random parameter instead of a fixed one. Furthermore, since virtually any $\mathsf{P} \in \mathcal{P}_{\mathbb{X}}$ can be seen as the limit of empirical measures, P_n, probabilities must then be computed on subjective judgments. With this, specifying an exchangeable sequence via a particular choice of de Finetti's measure, Q, can be interpreted as choosing a ***prior distribution*** on the random parameter, i.e. Θ or P, representing our ignorance about the phenomenon under study.

1.3. The Bayesian approach to inference

Equality (2) shows a duality between the ***finite dimensional distributions (fidis)*** corresponding to an infinite exchangeable sequence $\{X_i\}_{i=1}^{\infty}$ and de Finetti's measure Q. Hence, to improve our knowledge about the driving mechanism of $\{X_i\}_{i=1}^{\infty}$, in the light of observations, we could improve our knowledge about the law of Θ or P, i.e. Q.

1.3.1. The parametric case.

With the above purpose in mind one could consider a simplification. That is, instead of considering the space of all probability measures, $\mathcal{P}_{\mathbb{X}}$, we could restrict it to a ***finite parameter case***, through a parametric family, $\mathcal{P}_{\mathbb{O}}$. In other terms $\mathsf{Q}(\mathcal{P}_{\mathbb{O}}) = 1$. By doing this, the fidis of an exchangeable sequence, $\{X_i\}_{i=1}^{\infty}$, can be specified by a choice of a ***parametric model***, say $\mathsf{P}_\theta \in \mathcal{P}_{\mathbb{O}}$, and a ***prior distribution*** q on $(\mathbb{O}, \mathcal{B}_{\mathbb{O}})$, that is

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n] = \int_{\mathbb{O}} \left\{ \prod_{i=1}^{n} \mathsf{P}_\theta(A_i) \right\} \mathsf{q}(\mathrm{d}\theta), \quad \text{for all} \quad n \geq 0 \quad \text{and} \quad A_i \in \mathcal{X}. \tag{1.4}$$

From the aforementioned duality one could immediately deduce that ***predictive probabilities*** can be found as

$$\begin{aligned} \mathbb{P}[X_{n+1} \in A_{i+1} \mid X_1 \in A_1, \ldots, X_n \in A_n] &= \frac{\mathbb{P}[X_1 \in A_1, \ldots, X_{n+1} \in A_{n+1}]}{\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n]} \\ &= \frac{\mathbb{E}_{\mathsf{q}}\left[\prod_{i=1}^{n+1} \mathsf{P}_\Theta(A_i)\right]}{\mathbb{E}_{\mathsf{q}}\left[\prod_{i=1}^{n} \mathsf{P}_\Theta(A_i)\right]} \\ &= \mathbb{E}_{\mathsf{q}_{X^{(n)}}}\left[\mathsf{P}_\Theta(A_{n+1})\right], \end{aligned} \tag{1.5}$$

for all $n \geq 1$, $A_i \in \mathcal{X}$, and where

$$\mathsf{q}_{X^{(n)}}(\mathrm{d}\theta) = \frac{\prod_{i=1}^{n} \mathsf{P}_\theta(A_i) \mathsf{q}(\mathrm{d}\theta)}{\mathbb{E}_{\mathsf{q}}\left[\prod_{i=1}^{n} \mathsf{P}_\Theta(A_i)\right]} \propto \prod_{i=1}^{n} \mathsf{P}_\theta(A_i)\, \mathsf{q}(\mathrm{d}\theta) \tag{1.6}$$

is termed the ***posterior distribution*** for the random parameter Θ given the sample $X^{(n)} := (X_1, \ldots, X_n)$. Depending on the context, and for the ease of notation, the above measure will sometimes be denoted by $\mathsf{q}_n(\cdot)$ or $\mathsf{q}(\theta \mid x^{(n)})$. In (1.4), the term in brackets is referred as the ***likelihood*** for the given parametric model. The conditional probability (1.6) can be seen as a result of ***Bayes rule*** for the given likelihood and prior distribution. Indeed, in a hierarchical notation one can depict the dependence structure induced by (1.4) through

$$\begin{aligned} X_i \mid \Theta &\overset{\text{iid}}{\sim} \mathsf{P}_\Theta, \quad i = 1, \ldots, n \\ \Theta &\sim \mathsf{q} \end{aligned} \tag{1.7}$$

which is the most common setup to specify a Bayesian model.

Assuming the existence of densities for (1.6), upon observing that $\mathsf{X}^{(n)} = x^{(n)}$, the updated knowledge about the distribution of Θ is given by

$$\mathsf{q}(\theta \mid x^{(n)}) \propto \left\{ \prod_{i=1}^{n} f_{\Theta|X}(\theta \mid x_i) \right\} \mathsf{q}(\theta)$$

and all inferences about Θ must be derived from it (the ***Bayesian paradigm***). The posterior distribution allows us to infer about any feature of intrest regarding the parameter, e.g. mean, variance, hypothesis testing, etc. In principle, this makes the Bayesian approach appealing to perform statistical induction about the *epistemic uncertainty*, once $\mathcal{P}_{\mathbb{O}}$ (the *random uncertainty* represented by the model) is fixed. However various complications arise, e.g. which family $\mathcal{P}_{\mathbb{O}}$ shall we restrict on, the posterior is difficult to obtain analytically or difficult to handle numerically, etc. The literature treating the parametric Bayesian approach is vast, see for instance Robert (2001) and many of the references therein.

1.3.2. The nonparametric case: a reduction of subjectivity

If we avoid the above simplification, namely we undertake the more general framework where $\mathcal{P}_{\mathbb{X}}$ has an infinite dimensional character, then we shall term such an approach to inference as ***Bayesian nonparametric (BNP)***. When $\mathbb{X}$ is finite,

such as the binary case encountered in Theorem 1, the space $\mathcal{P}_{\mathbb{X}}$ takes a simpler form, e.g. the space of Bernoulli distributions, and thus the need of such simplification is futile. Also, such cases are already treated under the umbrella of the parametric case.

Framing the general approach of Theorem 2 in the statistical induction discussion of Section 1.1, one immediately realises that one could learn about both the *epistemic* and the *aleatory uncertainty* driving the random phenomena under study, when undertaking the BNP approach. Namely, we do not only learn about a finite dimensional parameter resulting of restricting $\mathcal{P}_{\mathbb{X}}$, but rather we learn about the whole infinite dimensional structure behind it. Having said this, it is good to emphasise that the subjectivity is not removed but rather reduced to the solely choice of de Finetti's measure Q.

Just as in the parametric case, one could intuitively write the predictive probability, (1.5), as

$$\mathbb{P}[X_{n+1} \in A_{i+1} \mid X_1 \in A_1, \ldots, X_n \in A_n] = \mathbb{E}_{\mathsf{Q}_{X^{(n)}}}[\mathsf{P}(A_i)]$$

where $\mathsf{Q}_{X^{(n)}}(B) := \mathbb{P}(\mathsf{P} \in B \mid X^{(n)})$, for any $B \in \mathscr{P}_{\mathbb{X}}$, denotes the posterior probability of the RPM, P, given the sample, $X^{(n)}$.

As we will see in Lecture 2 we will learn different ways of defining RPMs, P, and their distributions Q. Using the analogous notation to description (1.7) we can write

$$\begin{aligned} X_i \mid \mathsf{P} &\overset{\text{iid}}{\sim} \mathsf{P} \\ \mathsf{P} &\sim \mathsf{Q}. \end{aligned} \tag{1.8}$$

Marginalising the RPM in (1.8) we have that $X_i \sim \mathbb{E}_{\mathsf{Q}}[\mathsf{P}]$ marginally, for any choice of de Finetti's measure Q.