On the de Finetti's representation theorem: an evergreen (and often misunderstood) result at the foundation of Statistics

Loris Serafino*

April 23, 2016

Abstract

This paper reconsider the fundamental de Finetti's representation theorem. It is stressed its role at the front-line between Probability Theory and Inferential Statistics and its relation to the fundamental problem of relating past observations with future predictions i. e. the problem of induction.

1 Introduction

The aim of this paper is to introduce and explain the conceptual role played by the *de Finetti's representation theorem* (henceforth dFRT) in the modern theory of statistical inference. dFRT had a strange destiny. Published first by Bruno de Finetti (Innsbruck, 1906 - Roma, 1985) in an paper in French language, it was rediscovered years later after translation into English (Barlow, 1992). It has been recognized as a fundamental result for Bayesian Statistics and interpreted as a kind of *justification* for the subjective interpretation of probability. For many reasons dFRT does not find a place in undergraduate Statistics textbooks. First, undergraduate Statistics textbooks follow mainly the frequentist approach with Bayesian methods relegated (if lucky) in one final chapter close to the appendix. Second, dFRT involves some mathematical technicalities that are not easily accessible to undergraduates. Third, and perhaps most important, dFRT has a conceptual relevance rather than practical one and this makes it usually more compelling for philosophers than for statisticians. It has to be properly *interpreted*, i.e. to assign it a meaning that properly locates it conceptually inside the theoretical framework of modern inferential statistics. The usual interpretation of dFRT stressed its role *both* as a formal justification of the "degrees of belief" school of Probability theory *and* as a link between the latter

*email: geoloris@gmail.com

and the frequentist School (Poirier, 2010). Due to its borderline and foundational role, dFRT has been approached with deference in some technical presentation. On the other hand, the scope and *power* of this theorem is usually under-represented in popular Statistics expositions or introductory textbooks. In what follows, I will re-explore and clarify the meaning of the dFRT, stressing its pivotal role in particular in the *induction process* that was also the *crux* and motivation behind de Finetti efforts concentrated in its theorem (Barlow, 1992). I hope this work can serve as a source of inspiration for statisticians as well as philosophers of science involved in foundation of probability and statistics courses. In class, dFRT can be an opportunity to push the students to think "out of the box" with respect to the traditional frequentist curriculum. The path followed here goes from some fundamental concept of Inferential Statistics to the very *core* of dFRT. It can be imagined as a "smooth" entry point in preparation for more advanced techniques, especially (but not limited to) for those interested in exploring the application of a subjective probability approach in Inferential Statistics.

1.1 Coins, independence and prediction

In life there are no difficult things to understand, everything depends on the path we follow on the way to reach the truth and clarify the terms used. For a full understanding of dFRT its important to review some basics facts about probability and inferential statistics [for the prerequisites the reader can refer to (Bernardo and Smith, 2009; Bolstad, 2013; Cox, 2006)]. In what follows I will use the term *induction* as a synonymous of *being able to probabilistically infer about future outcomes looking at the past relative frequencies*. At the outset some important clarification are in order. One difficulty of learning Statistics is that sometimes topics are presented in a way that mathematical formalism "covers" the meaning of the concepts so that the main point is not easy to be grasped by the novice. The good thing is that in Statistics we can explain a lot of things with the use of a very common and simple model: flipping a coin. At the outset I will focus on the coin-toss model, or more in general on the **Bernoulli random variable** case, for two reasons. First because historically this was the original setting considered by de Finetti and second because, despite its simplicity, it contains all the ingredients to fully appreciate the impact of this important mathematical result. The reader already used with these concepts can jump directly to section 3. So let's consider the canonical situation. We have a coin and we don't know if it is fair or not. We perform an experiment flipping the coin n times. We want to infer the probability of the $n+1$ toss. We can model the coin tossing process in the following way. First of all we introduce a random variable defined as:

$$X(E) = \begin{cases} 1 & \text{if } E_1 = \{\text{Head}\} \\ 0 & \text{if } E_2 = \{\text{Tail}\} \end{cases} \tag{1.1}$$

X is an example of a Bernoulli random variable characterize by the parameter $\theta = p(E_1)$. So we toss a coin n times and We the obtain a given list of zeros and ones. We want to use this list to estimate the probability of getting Head the next toss:

$$p(1|1, 0, 1 \ldots 0, 1, 1).$$

Despite its simplicity and ubiquity, the coin-tossing model has some important disadvantages and can lead students to some distorted ideas. The limitation is that when we consider a (not necessarily fair) coin, we are adopting an underlying *independent and identically distributed* assumption (*IID* henceforth) that is quite strong but that makes life very easy for the inferential exercise as I will clearly show below. But this does not represent the whole story. What if, for whatever reason, the tossing are not independent? What if they are not identically distributed[1]? Will we be able in these condition for example to predict the next toss given the past results?

$$p(x_{n+1}|x_1, \ldots, x_n) \tag{1.2}$$

Is induction possible in this case? Here is where the dFRT shows all its power since it clarify at least conceptually what we can and what we cannot do and know about induction in the the dangerous lands outside the safe IID enclosure. An important clarification has to be stated at the outset:

Remark 1. *Independence alone in general is not enough for induction, since*

$$p(x_{n+1}|x_1, \ldots, x_n) = p(x_{n+1}) \tag{1.3}$$

so this prevents the possibility to learn from the past. But if other than independent our random variables are also identically distributed, the IID case, then we can learn from the past using the relative frequency of the occurrence of the event of interest. This fact is usually given for granted in a first year undergraduate statistics course. As I will show below, the theoretical rationale behind this is another core result springing from the dFRT.

[1]This is not an exotic possibility even in a simple coin tossing experiment. With some practice, after many tosses, a person can become able to affect the outcome for example introducing a bias in favor of head. The probability of getting head can thus change during the experiment, invalidating the IID assumption.

2 inferential statistics: a bird's eye view

The usual "statistical inference" tale follows some traditional steps: a) we are interested in a natural phenomenon that can be properly described by a given random variable; b) it follows that the outcome of a possible experiment regarding the phenomenon can be described by an appropriate statistical model; c) a *parametric* statistical model $\mathcal{M}$ is defined in terms of the parametric family of densities that depend on one or more parameters θ_i; d) we observe data $\{x_1, \ldots, x_n\}$ as a particular realization of the random sample $\{X_1, \ldots, X_n\}$; e) we use the sample to infer the value of the parameter(s); f) we use the fully specified model for prediction of future realization of the event of interest. The exact way in which this recipe is put into practice depends on the *paradigm* adopted. We know that in life matters rarely can be separated strictly in black and white, there is always a fuzzy shade of gray. This is also true for this long standing debate about the disagreement between frequentists *versus* Bayesians. By and large the main line of fracture lies in the way each group interpret probability statements and how this is reflected in the approach to statistical inference. Here a brief sketch of the two main schools of thought.

2.1 The frequentist approach

In this approach, probability is the long-run frequency of the occurrence of an event. Parameters in the statistical models are considered fixed but unknown quantities. In any statistical problem, we have data that are generally sampled from some population or data generating process that is repeatable. Probability statements cannot be made about parameters because they cannot meaningfully be considered as repeatable (Draper (2011)). The common condition used in undergraduate Statistics is that observation are IID because **when this property can be assumed** the statistical model of the joint distribution of $(X_1, \ldots, X_n)$ can be simplified tremendously by factorization:

$$p(x_1, \ldots, x_n) = \prod_i p(x_i, \theta). \tag{2.1}$$

Notice that in the formula above I put a comma between the sample value x_i and the parameter θ because for the frequentist a parameter is indeed just a parameter: a constant whose value is unknown. Any assumption about the term $p(x_1, \ldots, x_n)$ appearing in (2.1), the joint distribution of $(X_1, \ldots, X_n)$, is one of the fundamental starting points of the inferential process and at the same time the entry point for a full understanding of dFRT. Given this, frequentist estimation of θ is based on the concept of *likelihood function*. Given a realization of $\{X_i\}_{i=1\ldots n}$, i.e. $\{x_i\}_{i=1\ldots n}$, the likelihood is the joint distribution regarded as a function of the parameter:

$$L(\theta, x) = \prod_i \theta^{x_i}(1-\theta)^{n-x_i} = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}, \quad \theta \in [0,1] \tag{2.2}$$

The value of θ for which $L(\theta, x)$ reaches its maximum represents, for some well known theroetical consideration, a very good estimation of the real θ (Wasserman, 2013, chapter 9, p 122). This is what is known as the **maximum likelihood estimation** (MLE). For the Bernoulli case, it can be easily proved that:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\, L(\theta, x) = \frac{\sum_i x_i}{n} \tag{2.3}$$

It's worth stressing that for the frequentist θ is unknown but fixed, so letting the parameter varying in a range is regarded as just a mathematical artifice. The estimation of θ through MLE method leads to the relative frequency.

2.2 The Bayesian approach

For the Bayesian parameters are not just parameters in the above sense but random variables, something that in the mind of the researcher can assume different values with different probabilities attached to them. As any random variable, the parameter θ is specified by a distribution or a density called *prior* $\pi(\theta)$ that is based on the state of knowledge that the subject interested in the random experiment possesses about the parameters; it is here that the concept of *degree of belief* enters into the picture: the prior is a (not necessarily subjective) idea about the possible values of the parameter that can be different for different subject according to the different knowledge that they possess about the data-generating mechanism of event of interest. It is not important where the prior comes from, what is important for the Bayesian framework is how we "update" our knowledge by combining the prior and the information collected by a random experiment in the form of a set of data. This is given formally by the famous Bayes formula:

$$p(\theta|x_1, \dots x_n) = \frac{p(x_1, \dots x_n|\theta)\pi(\theta)}{p(x_1, \dots x_n)} = \frac{p(x_1, \dots x_n|\theta)\pi(\theta)}{\int_0^1 p(x_1, \dots x_n|\theta)\pi(\theta)d\theta}. \tag{2.4}$$

The denominator of the previous formula is the again the joint distribution of $(X_1, \dots, X_n)$. Here $p(x_1, \dots x_n)$ is called *marginal distribution* because it does not depend on the parameter since it has been marginalized out by integration:

$$p(x_1, \dots x_n) = \int_0^1 p(x_1, \dots x_n|\theta)\pi(\theta)d\theta. \tag{2.5}$$

This equality can be easily derived by assuming that $\int_0^1 p(\theta|x_1, \dots x_n) = 1$, a fact that makes sense only in the Bayesians framework where θ is treated as a random variable. The term $p(x_1, \dots x_n|\theta)$ appearing on the right side of (2.4) is the likelihood function. Formlly it looks like the same we discussed before in the frequentist case but conceptually there are important differences. Furthermore, even Bayesians adopt a concept of independence to simplify the joint distribution of $(X_1, \dots, X_n)$ but now this assumes the following form of *conditional* independence:

$$p(x_1, \dots, x_n|\theta) = \prod_i p(x_i|\theta). \tag{2.6}$$

In this case we have to put a bar | in the above formula to stress than we are conditioning *given* the value of the random variable $\Theta = \theta$. There is a huge conceptual gap between formula (2.6) and formula (2.1) reflecting the fracture that opposes frequentist and Bayesian in the way they look at the inferential process.

Something at this point must be clear to the reader. Inference for the frequentist means to find an approximate value of the (unknown) constant θ extracting information from the collected sample at hand $(x_1, \dots, x_n)$. Inference for the Bayesian means to *improve his initial knowledge about the distribution of the parameter*. Both will use their findings about θ to use the statistical model for prediction of future events. This will be a recurrent theme if what follows. The issue of *trying to predict the future using the past information* is a important point at stake for both frequentists and Bayesian and it will be stressed again below since is a key ingredient in the

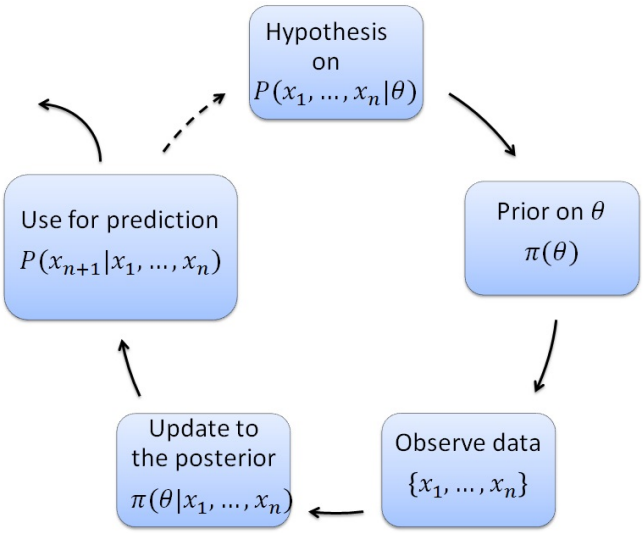

Figure 1: The Bayesian cycle for prediction.

elucidation of dFRT. Inferences about the parameter θ uses (2.4). What we need is to specify the prior and this of course can be a subjective ingredient. For the Bernoulli case, the Bayesian machinery uses the Beta distribution (Bolstad, 2013, chapter 8, p 143):

$$\pi(\theta) = Beta(\theta, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \text{for } 0 \leq \theta \leq 1 \tag{2.7}$$

This prior, combined with the likelihood function, generate the posterior according to Bayes rule (2.4). Different choice of the parameters a and b will generate different classes of priors. In figure (2) the initial prior (horizontal line) is the uniform density generated with the choice $a = 1$ and $b = 1$. The black curve refers to the density of a the Beta(47,55), representing the posterior in the case of fair coin .

3 To the de Finetti's representation theorem

The starting point to the dFRT is different and lies in the more general concept of *exchangeability* instead of independence. Informally this means that given the set of sampled observations $\{x_i\}_{i=1...n}$, the order of these observations does not matter. This applies in the case of usual understanding of multiple tosses of a coin. In the "coin tossing\number of heads" experiment modeled via the random variable (1.1), the observed list of outcomes, for example $\{0, 1, 0, 0, 1, 0\}$, are expected to be exchangeable since the probability of this sequence does not change if we change the

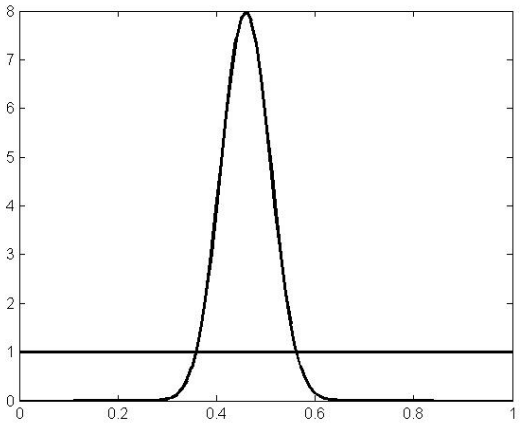

Figure 2: From a uniform prior (horizontal line) and after 100 tosses of a fair coin where $\#H \sim \#T$. The resulting posterior is a Beta.

order of digits. *What seems relevant here is not the order but the numbers of 1's.* Exchangeability expresses a kind symmetry of beliefs about the random quantities in which the future observation are expected to be similar to past observation (Stern, 2011; O'Neill, 2009; Bernardo, 1996; Freedman, 1995). The next, is a very important:

Remark 2. *The condition of exchangeability is weaker than independence but it is stronger than the identically distributed property. It can be easily proven that IID random variables are exchangeable (Poirier (2010); Heath and Sudderth (1976)).*

There are many situations in which this assumption is reasonable like in the coin tossing experiment, and others where is not true or questionable. Consider the following example. A football player who is practicing to score in a penalty: The sequence scored penalties $FAIL, FAIL, FAIL, GOAL, GOAL$ has presumably a higher probability than $GOAL, GOAL, FAIL, FAIL, FAIL$, because the player accuracy improves with practice so we can expect the future will be different from the past. For the mathematician taste, here I give a more formal definition of exchangeability:

Definition 3. *A set of random variable* $\{X_n\}$ *is said to be exchangeable if, given the joint density* $p(x_1, \ldots, x_n)$*, we have*

$$p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \tag{3.1}$$

for all permutations σ *of* $1, ..., n$*.*

If, like in the example of the IID coin toss, we are in presence of exchangeability, this has important conceptual consequences in terms of predictive inference. Being exchangeable means that the past is similar to the future and this symmetry can be translated saying that knowing the past is telling us something about the future and helps to predict the future. As already mentioned, this is strictly related to the problem of predictive inference, that is to estimate:

$$p(x_{n+1}|x_1, \ldots, x_n) \tag{3.2}$$

Equipped with the concept of an exchangeable sequence we can now state the dFRT for Bernoulli distributed random variables. Various forms of extension and generalization of exchangeability and de Finetti result can be found in literature. The interested reader can refer to (De Finetti, 1937; Diaconis and Freedman, 1980; Hewitt and Savage, 1955).

Theorem 4 (De Finetti, 1930)**.** *let* $\{X_i\}_{i=1}^{\infty}$ *be a sequence of finitely exchangeable random variables i.e.* $\forall n > 0$ *each finite sub-sequence* $\{X_i\}_{i=1}^{n}$ *is exchangeable. Then there exists a random variable* Θ *and a distribution function* $\mathcal{F}(\theta)$ *such that:*

$$p\left(\lim_{n\to\infty} \frac{\sum X_i}{n} = \Theta\right) = 1 \quad with \quad \Theta \sim \mathcal{F}(\theta) \tag{3.3}$$

and

$$p(x_1, \ldots, x_n) = \int_0^1 \left[\prod_i \theta^{x_i}(1-\theta)^{1-x_i}\right] d\mathcal{F}(\theta) \tag{3.4}$$

A detailed proof can be found in Chapter 4 of (Bernardo and Smith, 2009). Here I will try to motivate the relevance of dFRT for the induction problem with some examples and computer simulations. First let me clarify some points. As previously mentioned, the main ingredient of inferential statistics is given by the hypotheses over the structure joint probability distribution $p(x_1, \ldots, x_n)$. The dFRT tells us that under exchangeability (not necessarily IID) the correct form of this joint probability is given by (3.4). $\mathcal{F}(\theta)$ in (3.3) is sometimes referred to as the *mixing distribution* of the exchangeable random variable. $d\mathcal{F}(\theta)$ can be thought of as equivalent to $\pi(\theta)d\theta$ (in the sense of the Stieltjes integral) when $\mathcal{F}(\theta)$ is continuous (Spanos, 1999, chapter 10, p 524). This said, (3.4) becomes:

$$p(x_1, \ldots, x_n) = \int_0^1 \left[\prod_i \theta^{x_i}(1-\theta)^{1-x_i}\right] \pi(\theta)d\theta \tag{3.5}$$

where $\pi(\theta)$ can be interpreted as the density function related to Θ.

Another important point of the theorem rests primarily on the *existence* result (3.3). it assures the existence of a random variable that encapsulate the maximum possible knowledge about the underlying data-generating mechanism that produces data. The (3.3) represents in fact the *Law of Large Numbers for exchangeable random variables*, a very important result embedded inside the dFRT. Under exchangeability condition the relative frequency $\sum X_i/n$ tends to a random variable, not necessarily "degenerate" (i.e. constant with probability one) as in the IID classical *Large Number Law* case. A point usually stressed since the first de Finetti philosophical interpretation is that dFRT justifies the use of a probability distribution over the parameter $\Theta = p(X = 1)$. Summarizing, the condition of exchangeability implies:

- There exists a random variable Θ such that:

$$p(x_1, \ldots, x_n | \Theta = \theta) = \theta^k (1-\theta)^{n-k};$$

- Θ is the limit of the relative frequencies: this is the the more general *Law of Large Numbers for exchangeable random variables* and

$$\mathcal{F}(\theta) = \lim_{n \to \infty} p\left(\frac{\sum_{i=1}^{n} X_i}{n} \leq \theta \right)$$

- if $\mathcal{F}$ has density, $d\mathcal{F}(\theta) = \pi(\theta) d\theta$ where $\pi(\theta)$ is the density of Θ. Before observing the data, any hypothesis about $\pi(\theta)$ (right or wrong that it can be) corresponds to the prior: it is the idea about the underlying structure of the parameter Θ before the data are collected;

Combining dFRT and Bayes rule, and after some calculus "gymnastic", it is possible to show that:

$$p(x_{n+1} = 1 | x_1, \ldots, x_n) = \int_0^1 \theta \pi(\Theta | x_1, \ldots, x_n) d\theta = E\left(\pi(\Theta | x_1, \ldots, x_n)\right). \qquad (3.6)$$

This means that, after the posterior is obtained, the best prediction about a future observation is the expected value of the posterior (figure 3).

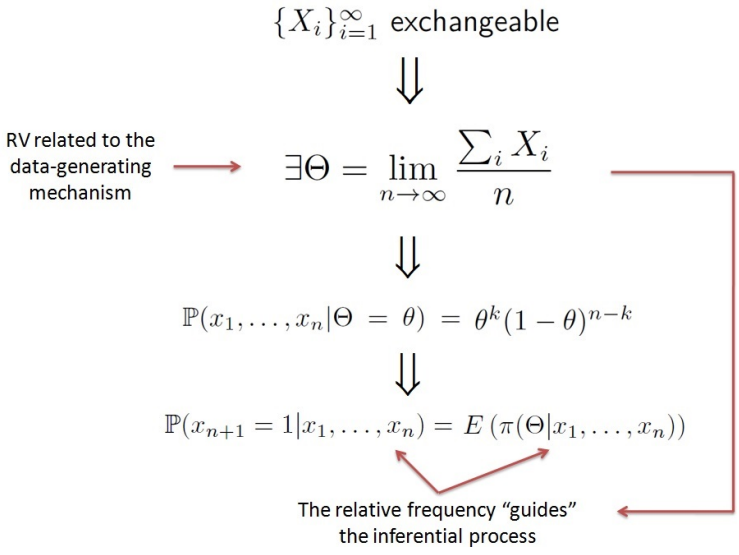

Figure 3: dFRT: the overall picture.

It should be clear now how much dFRT is important for the problem of induction. Before observing the sample $\{x_1, \ldots, x_n\}$, what we call *prior* is an idea about the distribution of Θ that corresponds to an idea about its density $\pi(\theta)$. This is strictly related the the underlying data-generating mechanism that describe the structure of the joint distribution $p(x_1, \ldots, x_n)$. The idea about $\pi(\theta)$ can be more or less "close" to the correct, real, distribution of θ, but after observing the data, something important happens for the possibility of the induction process (i.e. making probabilistic statements about the future using observation from the past):

Remark 5. *Given the exchangeability hypotheses, and whatever is the observer's idea about the prior density $\pi(\theta)$, the induction about the probability of the next observation of the event of interest given the data, will be strongly guided by the relative frequency of the observed event of interest.*

In what follows I will motivate it with some examples where I will stress how the theorem helps to solve the theoretical problem of induction.

3.1 Case I: $\{X_i\}_{i=1}^{\infty}$ IID.

In this case, since $\{X_i\}_{i=1}^{n}$ are IID, by the law of large numbers we have that $\frac{\sum X_i}{n}$ converges to a degenerate random variable Θ, that is a random variable for winch there exist one value θ such that $P(\Theta = \theta_0) = 1$ and such that $E(X_i) = \theta_0$. This case is equivalent to say that (in what follows we assume $k = \sum_i x_i$):

$$p(X_1 = x_1, \ldots, X_n = x_n) = \theta_0^k (1 - \theta_0)^{n-k} \tag{3.7}$$

This is a quit special situation. In general things are more complicated as I will mention below. In this case Θ does not have a density but, according to the discussion above, we can still manage the integral (3.4) "as if" it's density is represented by a *Dirac delta function* (for a "refresh" of its properties, see appendix A):

$$\pi(\theta) = \delta(\theta - \theta_0) \tag{3.8}$$

and the corresponding step distribution:

$$\mathcal{F}(x) = \int_{-\infty}^{x} \delta(\theta - \theta_0) d\theta. \tag{3.9}$$

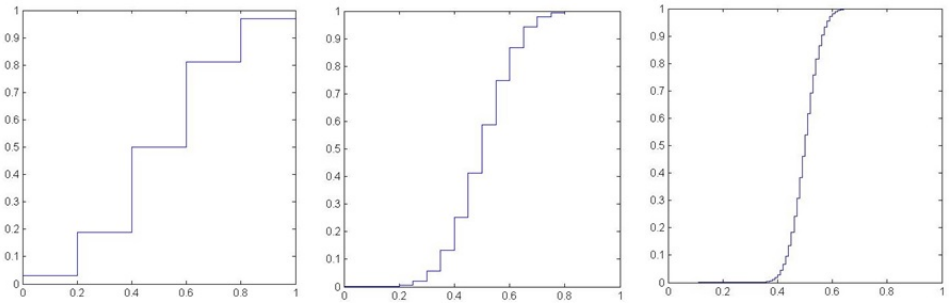

Figure 4: Distribution of $\sum_i^n X_i/n$ in the case $p(\sum X_i = k) = \binom{n}{x}(0.5)^n(0.5)^{n-k}$ for different values of n. From the left to the right $n = 5, 20, 100$. It clearly shows that the limit distribution corresponds to the degenerate case $\Theta = \theta_0$ with, in this example, $\theta_0 = 0.5$.

In this case we have:

$$\int_0^1 \left\{\theta^k(1-\theta)^{n-k}\right\} \delta(\theta - \theta_0)d\theta = \theta_0^k(1-\theta_0)^{n-k}, \tag{3.10}$$

where the natural choice for approximating θ_0 is

$$\hat{\theta}_0 = \frac{k}{n}. \tag{3.11}$$

This is the case where frequentist and Bayesian meet.

3.2 Case II: $\{X_i\}_{i=1}^{\infty}$ exchangeable but not independent.

A very instructive example due to Bayes (Schervish, 2012, p.29). Let's imagine that we have a sequence of Bernoulli random variable $(X_1, X_2, \dots)$ such that

$$p\left(\sum_i^n X_i = k\right) = \frac{1}{n+1}, \quad \text{for } k = 1 \dots n. \tag{3.12}$$

In this case $\{X_i\}_{i=1}^{\infty}$ are exchangeable, they are also identically distributed since $p(X_i = 1) = \int_0^1 \theta d\mathcal{F}$ but they are not independent since for example $p(X_2|X_1) \neq p(X_2)$. dFRT applies, so we can specify the joint probability using (3.4). Since $\{X_i\}_{i=1}^{\infty}$ are not IID the dFRT tells us that $\sum_i^n X_i/n$ still converges to a random variable whose "structure" is now more complicated than the degenerate case saw in the classical IID example above. It can be easily shown analytically that $\mathcal{F}(\theta) = \theta$.

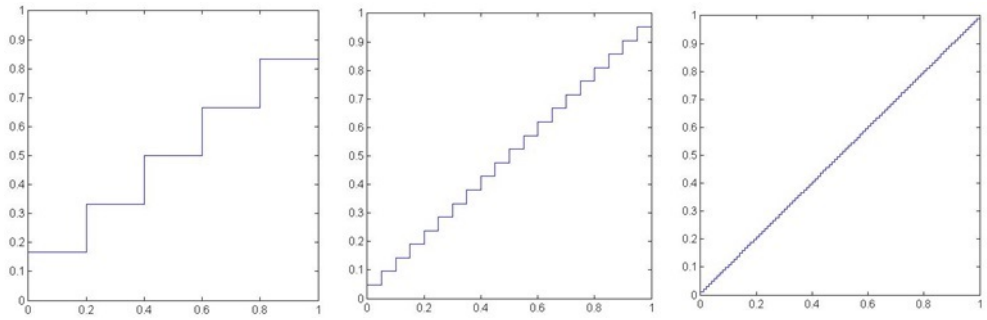

Figure 5: Distribution of $\sum_i^n X_i/n$ in the case $p(\sum X_i = k) = \frac{1}{n+1}$ for different values of n. From the left to the right $n = 5, 20, 100$. It clearly shows that the limit distribution is equal to θ.

This is equivalent to say that the prior $\pi(\theta) = 1$, i.e θ has a uniform distribution in $(0, 1)$. Here I will motivate this intuitively with the help of some computer simulations. Figure (5) shows the distribution function of $\sum_i^n X_i/n$ for our random variables at different values of n. It clearly shows what happens if $n$becomes large. From picture 3.2 we observe (and it can be easily proved) that the expected value of the posterior tends to 0.5. This implies that, if the underlying data generating mechanism is characterized by a Θ distributed as $\mathcal{F}(\theta) = \theta$, we expect to observe around 50% of $0's$ and 50% of $1's$. This is reasonable since the number of ways an almost 50/50 case can happen under this condition is overwhelmingly big compared to other arrangements as n grows. As stressed before, relative frequencies will lead to a correct prediction.

3.3 The general case.

In general the possible structure of $p(S_n = k)$ for the binary case is limited because of the constrains of the probability properties. The previous cases are only particular situations and in general it is possible to show that the formula for the general form of $p(S_n = k)$ is given by:

$$p(S_n = k) = \binom{n}{k} \int_0^1 \theta^k (1-\theta)^{n-k} \frac{\Gamma\left(\frac{b}{c}\right)\Gamma\left(\frac{r}{c}\right)}{\Gamma\left(\frac{b}{c}+\frac{r}{c}\right)} d\theta, \tag{3.13}$$

where Γ stands for the *Gamma function*, n is the number of successes over the total n and and $a, b, r, c$are suitable parameters (Helfand (2013)). The (3.13) is called *Pólya* urn model (for more details the reader can refer to (Johnson and Kotz, 1977)). It follows that:

$$\lim_{n\to\infty} \frac{\sum X_i}{n} = \Theta \sim \int_0^{\theta} \frac{1}{Beta\left(\frac{b}{c}, \frac{r}{c}\right)} u^{\frac{b}{c}-1}(1-u)^{\frac{r}{c}-1} du \tag{3.14}$$

3.4 An "extreme" case

Let's now consider the situation depicted in figure (6). Given X_i ∼Bernoulli(θ), such that $p(X_i = 1) = \theta$ and:

$$p(X_{n+1} = 1|X_n = 1) = 1 \quad \text{and} \quad p(X_{n+1} = 0|X_n = 0) = 1 \tag{3.15}$$

$\{X_i\}_{i=1}^{\infty}$ are exchangeable and satisfy the dFRT conditions, the relative frequency can direct successfully the induction process.

4 What if $\{X_i\}_{i=1}^{\infty}$ are not exchangeable?

Let's try to summarize the story so far. dFRT is an important theoretical tool since it can shed light on the meaning and role of the prior in the whole Bayesian cycle. It shows how the IID case represents just one among many different possibilities about the distribution over the parameter of interest θ. It is important to stress that if we are interested in a predictive exercise:

$$p(x_{n+1}|x_1, \dots x_n), \tag{4.1}$$

even if the initial hypotheses about the joint distribution differ, after the data have been collected, opinionns tend to "'converge". This is clearly shown in the following example.

Before the data X_i are collected, we have:

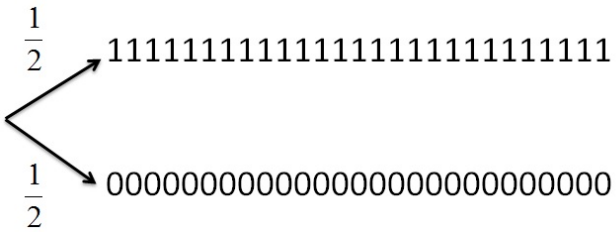

Figure 6: An "extreme" case of exchangeable RV.

- IID assumption: $p(1,0,1,0,1,0,1,0,1,0) = \frac{1}{2^5} = \frac{1}{1024}$
- Uniform prior assumption: $p(1,0,1,0,1,0,1,0,1,0) = \frac{1}{2772}$.

Given the observed set of data $1,0,1,0,1,0,1,0,1,0$, prediction about the next observation will be:

- IID assumption: $p(1|1,0,1,0,1,0,1,0,1,0) = \frac{1}{2}$
- after updating the uniform: $p(1|1,0,1,0,1,0,1,0,1,0) = \frac{1}{2}$

This is a well known fact in Statistical practice: irrespective of the idea about the prior, different posterior will tend to be close to each other after data are collected (figure 8). In particular *the expected value of the posterior will be close to the relative frequency* and so it will be the prediction about the probability of the next observation of an event. So the relative frequency plays an important role but *this is true if we are in presence of exchangeable random variables.* If they lack this property, relative frequency in general is no longer able to lead the induction process. I will show this with the following example.

Given $X_i \sim$Bernoulli(θ), such that

$$p(X_{n+1} = 1|X_n = 0) = 1 \quad \text{and} \quad p(X_{n+1} = 0|X_n = 1) = 1 \tag{4.2}$$

The evolution with non zero probability are depicted in (7). X_i are not exchangeable and *induction fails* since the relative frequency now is not a guide to the estimation of the underlying mechanism that produced the data. Indeed for example $p(1|1,0,1,0,1,0,1,0,1,0) = 1$ and not the value suggested using the relative frequency that is 0.5 in this case.

Summarizing: *if a sequence of random variable is exchangeable, the relative frequency of data leads to a proper evaluation of the predictive probability. If the random variables are not exchangeable, in general the relative frequency will not guide to a proper inferential conclusion.* This is the case where the Bayes inferential machinery (2.4) *goes haywire.* This can be synthesize in the following final:

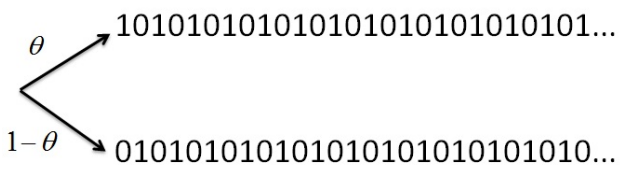

Figure 7: A case of not exchangeable random variables. Induction will fail in this case.

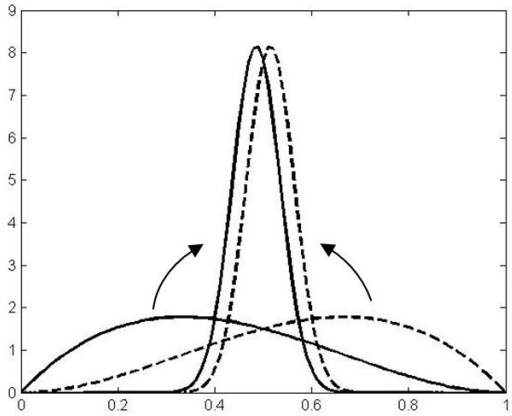

Figure 8: Different opinions about θ will converge, with the expected value close to the relative frequency.

Remark 6. *If X_i are not exchangeable, the relative frequency is in general no longer able to direct the induction process to a proper conclusions.*

5 The moral of the story

Frequentist and Bayesian inference are usually pictured as irreconcilable paradigms in Statistics and the main difference between the two (parameters fixed versus parameters as random variables) often stressed as a the fracture between the two visions. The de Finetti's representation theorem is usually introduced in the context of Bayesian inference and it is considered to play a role in the "justification" of the prior distribution of the parameter of interest. In this expository work I tried to challenge this view with an understanding of the theorem that stresses its role at the front line between probability theory and inferential statistics, and its relation to the problem of relating past observations with future predictions. To conclude, a list of the main key-points:

- exchangeability is the key property for induction;
- the use of relative frequencies for prediction during the induction process makes sense only in the presence of exchangeability;
- de Finetti's theorem clarify the role played by the relative frequency in the Bayesian framework;
- the IID case is a particular case;

- for non-exchangeable random variables, relative frequencies will fail to guide the induction process;
- the theorem can be extended to arbitrary real-valued exchangeable sequences (De Finetti, 1937). Finite version and generalizations can be found in (Diaconis and Freedman, 1980). Further generalizations in (Hewitt and Savage, 1955).

  "*This [theorem] is one of the most beautiful and important results in modern statistics. Beautiful, because it is so general and yet so simple*(Lindley and Phillips, 1976)"

Appendix A Dirac delta

It is obvious that probability density is definite only for absolute continuous variables. However, in some "pathological" situation can be useful to extend the concept of density. The figure 4 below depicts the posterior shapes (described by Betas distribution) for different valises of n. The when N increases the base of the bell-shaped density will be narrower and narrower and the top higher and higher. For N very big we can imagine that the density will tend to something with an infinitesimally narrow base whereas the height goes to infinity. The limit density when $N \to \infty$ is not a "traditional" density but an exotic mathematical object called a *Dirac delta generalized distribution*. The Dirac delta, also called *generalized function*, is usually indicated as $\delta(\theta - \theta_0)$ and formally it can be described as follows:

$$\delta(\theta - \theta_0) = \begin{cases} 0 & \theta \neq \theta_0 \\ \infty & \theta = \theta_0 \end{cases} \tag{A.1}$$

with the property that

$$\int_{\mathbb{R}} \delta(x)dx = 1 \tag{A.2}$$

The delta function is not a distribution, technically it is not even a mathematical function. Instead it can make sense to use it inside integrals in operation involving limits of sequences of normalized (integral= 1) functions behaving like the Beta in (A). If we have a sequence of such functions $\delta_n(x)$ it holds that:

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f(x)\delta_n(x - x_0)dx = f(x_0) \tag{A.3}$$

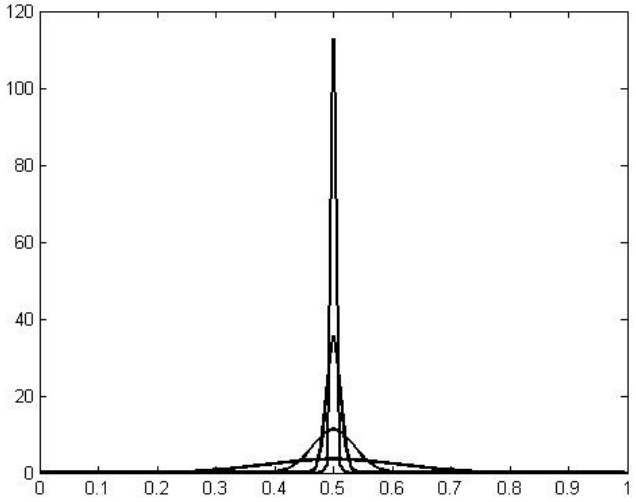

Figure 9: The limit density of $p(\theta)$ will be a δ function

It should be emphasize that the integral on the left-hand side of is not a Riemann integral but a limit. It can be treated as a *Stieltjes integral* if desired. $\delta(x)dx$ is replaced by $dH(x)$, where $dH(x)$ is the *Heaviside step function.*

References

Barlow, R. (1992). Introduction to de finetti (1937) foresight: its logical laws, its subjective sources. In *Breakthroughs in statistics*, pp. 127–133. Springer.

Bernardo, J. M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences 4*, 111–122.

Bernardo, J. M. and A. F. Smith (2009). *Bayesian theory*, Volume 405. John Wiley & Sons.

Bolstad, W. M. (2013). *Introduction to Bayesian statistics.* John Wiley & Sons.

Cox, D. R. (2006). *Principles of statistical inference.* Cambridge University Press.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, Volume 7, pp. 1–68.

Diaconis, P. and D. Freedman (1980). Finite exchangeable sequences. *The Annals of Probability*, 745–764.

Draper, D. (2011). Bayesian modeling, inference and prediction.

Freedman, D. (1995). Some issues in the foundation of statistics. *Foundations of Science 1*(1), 19–39.

Heath, D. and W. Sudderth (1976). De finetti's theorem on exchangeable variables. *The American Statistician 30*(4), 188–189.

Helfand, N. (2013). Polya's urn and the beta-bernoulli process. *University of Chicago REU*.

Hewitt, E. and L. J. Savage (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 470–501.

Johnson, N. L. and S. Kotz (1977). *Urn Models and Their Application. An Approach to Modern Discrete Probability Theory*. John Wiley & Sons, New York-London-Sydney.

Lindley, D. V. and L. Phillips (1976). Inference for a bernoulli process (a bayesian view). *The American Statistician 30*(3), 112–119.

O'Neill, B. (2009). Exchangeability, correlation, and bayes' effect. *International Statistical Review 77*(2), 241–250.

Poirier, D. J. (2010). Exchangeability, representation theorems, and subjectivity. *Handbook of Bayesian Econometrics*.

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media.

Spanos, A. (1999). Probability theory and statistical inference. *Cambridge Books*.

Stern, J. M. (2011). Symmetry, invariance and ontology in physics and statistics. *Symmetry 3*(3), 611–635.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.