BORDA WORKING PAPERS

José Manuel Gutiérrez

ON CONDITIONAL PROBABILITY AND BAYESIAN INFERENCE

Working paper No.: 2001

May 2020

http://borda.usal.es

ON CONDITIONAL PROBABILITY AND BAYESIAN INFERENCE

JOSÉ MANUEL GUTIÉRREZ

Faculty of Economics and Business, University of Salamanca

Abstract. Measurement theory has dealt with the applicability of the conditional probability formula to the updating of probability assignments when new information is incorporated. In this paper the original probability measure is taken as given, and an assumption on the relation between this probability and a possible conditional probability is imposed. Provided that the original probability is non-atomic, it is proved that there is one and only one transformed probability measure satisfying the assumption. Building on this result, we discuss the hypotheses underlying Bayesian inference. In the Bayesian parametric model, a joint probability distribution on the product of the sample space and the parameter space is assigned. As this probability distribution is shown to be non-atomic, we conclude that, apart from measure-theoretic representability hypotheses, the existence of this joint probability is the only nontechnical hypothesis underlying Bayesian parametric statistical inference.

1. Introduction

Conditional probability is, on the one hand, an intuitive concept, which captures the change in the original probability assignment when new information is known. On the other hand, the axiomatic definition of conditional probability is given by a formula that determines it from the original probability. Often both concepts are identified, and it is postulated that the incorporation of new information alters the original probability assignment according to this formula.

As always when an axiomatic definition is applied, it is worth discussing that applicability in each case. Indeed, when considering the frequentist interpretation of probability, there are plausible reasons for such applicability. In the case of the subjective interpretation of probability, as a degree of belief, typical of Bayesian statistical inference, arguments have been constructed to justify that the change in the assignment of probabilities when new information is incorporated must follow the conditional probability formula. These arguments start from a qualitative relation of the form $A \mid B \succsim C \mid D$, meaning "$A$ given B is qualitatively at least as probable as C given D", satisfying certain elaborated assumptions (see [7]). Then it is proved that there is one and only one probability P such that

$$A \mid B \succsim C \mid D \text{ iff } \frac{P(A \cap B)}{P(B)} \geq \frac{P(C \cap D)}{P(D)}$$

This result is to be understood within measurement theory, where the representation by probabilities of qualitative probability orderings of events is discussed;

Key words and phrases. Conditional probability, Bayesian inference, Non-atomic measure

Address. Facultad de Economía y Empresa. Universidad de Salamanca. Salamanca 37007. Spain.
E-mail. jmgut@usal.es.

usually finitely additive probabilities have been considered, although completely additive probabilities have also been studied (see [10]).

We consider in this paper a different starting point to justify the applicability of the axiomatic definition of conditional probability (i.e. the conditional probability formula). The original probability measure is taken as given, and an assumption on the relation between this original probability and a possible conditional probability is imposed (Aristotelian Assumption, (A.A) for short). Provided that the original probability is non-atomic, it is proved that there is one and only one transformed probability measure satisfying the assumption (Theorem 2.3).

We claim that the approach just mentioned is adequate to discuss the hypotheses underlying Bayesian statistics. For simplicity, we take momentarily all probability distributions to be representable in terms of densities. Suppose that $Y = (Y_1, ..., Y_n)$ is a random vector of n observations taking values on a sample space S. The parameter $\theta = (\theta_1, ..., \theta_k)$ with values in a parameter space $\Theta \subseteq \mathbb{R}^k$ indexes the various possible density functions $p(y \mid \theta)$ for Y; so $p(y \mid \theta)$ denotes the distribution of Y *when* θ *is known.* Bayesian statistics postulates that $p(y \mid \theta)$ represents a conditional distribution following the conditional probability formula. Thus (Y, θ) has a probability distribution (say with joint density $p(y, \theta)$; $p(y)$ and $p(\theta)$ stand for the density marginals) and

$$p(y \mid \theta)p(\theta) = p(y, \theta) \tag{1}$$

On the other hand, given the observed data $y = (y_1, ..., y_n)$, let $p(\theta \mid y)$ denote the distribution of the parameter θ *when* y *is known.* Bayesian statistics now postulates that $p(y \mid \theta)$ represents a conditional distribution following the conditional probability formula. Thus

$$p(\theta \mid y)p(y) = p(y, \theta) \tag{2}$$

Equating (1) and (2), Bayes' formula for the posterior distribution follows:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \tag{3}$$

Bayes' formula for the posterior distribution is certainly the basis of Bayesian statistics. In general, two hypotheses are underlying this formula:

(H1) There is a joint probability measure P on $S \times \Theta$[1].

(H2) If $P(A \mid C)$ is given the interpretation "probability of event A *when event* C *is known*", then the conditional probability formula applies:

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)}$$

In the Bayesian parametric model, the joint probability P is shown to be non-atomic (Proposition 3.2). Taking (A.A) for granted, it follows from Theorem 2.3 that, at least in the parametric case, condition (H2) is redundant, and only (H1) is necessary for the Bayes' formula for the posterior distribution. We conclude that, apart from measure-theoretic representability hypotheses, the existence of that joint probability on $S \times \Theta$ is the only nontechnical hypothesis underlying Bayesian parametric statistical inference.

[1] The existence of a suitable joint probability is far from being a foregone conclusion from that of the marginals. The case of quantum mechanics is to the point. In that theory both $P(A)$ and $P(B)$ may exist and yet $P(A \cap B)$ need not (think of A referring to the position of a particle and B to its momentum).

2. The formula of conditional probability

In this section $(\Omega, \mathcal{A}, P)$ is a probability space, where Ω is a set, $\mathcal{A}$ is a σ-algebra in Ω and P is a (σ-additive) probability measure. Let $C \in \mathcal{A}$, with $P(C) > 0$.

Definition 2.1. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $C \in \mathcal{A}$ with $P(C) > 0$. The probability space $(\Omega, \mathcal{A}, P')$ is called a pre-conditional probability given C iff $P'(C) = 1$ and the following assumption hold:*
(A.A) If $A, B \in \mathcal{A}$ and $A, B \subseteq C$, then

$$P(A) = P(B) \textit{ implies } P'(A) = P'(B)$$

This definition arguably captures obvious requirements for any re-assignement of probabilities when we have the added information that the outcome is one of the elements of the event C. The requirement $P'(C) = 1$ says simply that "the outcome is one of the elements of the event C". Besides, the original assignment of probabilities has to have an influence on the new assignment, and not merely be thrown away. It has to be re-worked in an even-handed way, and (A.A) is in this sense a minimum requirement, expressing some sort of Aristotelian "treat like cases alike" principle.

Assumption (A.A) is rather mild, and it may be even unconstraining.

Example 2.1. *Consider that $\Omega := \{1, 2, 3, 4\}$, $\mathcal{A}$ is the set of all subsets of Ω, $P(1) := \frac{1}{10}$, $P(2) := \frac{3}{10}$, $P(3) := \frac{5}{10}$, $P(4) := \frac{1}{10}$, and $C := \{1, 2, 3\}$. Then any probability space $(\Omega, \mathcal{A}, P')$ is a pre-conditional probability given C, provided that C is a support of P' (i.e. $P'(C) = 1$).*

The set function $P(\cdot \mid C)$ on $\mathcal{A}$ defined by

$$P(A \mid C) := \frac{P(A \cap C)}{P(C)} \tag{4}$$

makes $(\Omega, \mathcal{A}, P(\cdot \mid C))$ into a pre-conditional probability given C. We are interested in the question of its uniqueness as a pre-conditional probability.

It is immediate that, if a probability space $(\Omega, \mathcal{A}, P')$ satisfies $P'(C) = 1$, then the following three conditions are equivalent:
(i) $P' = P(\cdot \mid C)$, as defined in (4)
(ii) If $A \in \mathcal{A}$ such that $A \subseteq C$, then

$$P'(A) = \frac{P(A)}{P(C)} \tag{5}$$

(iii) If $A, B \in \mathcal{A}$ such that $A, B \subseteq C$ and $P(B) > 0$, then $P'(B) > 0$ and

$$\frac{P'(A)}{P'(B)} = \frac{P(A)}{P(B)}$$

Recall that $A \in \mathcal{A}$ is an *atom* for P iff: (a) $P(A) > 0$ and (b) for every $B \in \mathcal{F}$ with $B \subseteq A$, either $P(B) = 0$ or $P(B) = P(A)$. A probability measure P which has no atoms is called *non-atomic*, and it is called *atomic* iff every $E \in \mathcal{A}$ such that $P(E) > 0$ contains an atom. If P is a probability measure, then there exist unique probability measures P_1 and P_2 and $\alpha \in [0, 1]$ such that $P = \alpha P_1 + (1 - \alpha) P_2$ and such that P_1 is atomic and P_2 is non-atomic (see [6] for further discussion in the general context of measures).

The following result is a particular case of a theorem of Sierpinski [9].

Theorem 2.1. *Let $(\Omega, \mathcal{A}, P)$ be a probability space with P non-atomic. If $E \in \mathcal{A}$ and $P(E) > 0$, then for every $\alpha \in [0, P(E)]$ there is an element $F \in \mathcal{A}$ with $F \subseteq E$ and $P(F) = \alpha$.*

Induction on k gives directly the next corollary of Theorem 2.1 (see [8]).

Corollary 2.2. *Let P be non-atomic, and suppose $E \in \mathcal{A}$ such that $P(E) > 0$. Let α_i for $i = 1, ..., k$ be real numbers with $\alpha_i > 0$ and $\sum\limits_{i=1}^{k} \alpha_i = P(E)$. Then E can be decomposed as a union of disjoint sets $E_i \in \mathcal{A}$ with $P(E_i) = \alpha_i$ for $i = 1, ..., k$.*

Provided that a probability measure is non-atomic, we are going to see that any pre-conditional probability is determined by the conditional probability formula.

Theorem 2.3. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $C \in \mathcal{A}$ with $P(C) > 0$. Suppose that $(\Omega, \mathcal{A}, P')$ is a pre-conditional probability given C. If P is non-atomic, then $P' = P(\cdot \mid C)$ as defined in (4).*

Proof. Let $A \in \mathcal{A}$ such that $A \subseteq C$. In order to prove (5), it can be assumed, without loss of generality, that $P(A) > 0$. The proof will be divided into three steps.

(a) Consider the case $\frac{P(A)}{P(C)} = \frac{1}{q}$, where $q \in \mathbb{N}$, $q > 0$.

Applying Corollary 2.2 to C, with $\alpha_i = \frac{1}{q}P(C)$ for $i = 1, ..., q$, there exist disjoint sets $C_1, ..., C_q \in \mathcal{A}$ such that $\bigcup\limits_{i=1}^{q} C_i = C$ and $P(C_i) = \frac{1}{q}P(C) = P(A)$ for $i = 1, ..., q$. By (A.A), $P'(C_i) = P'(A)$ for $i = 1, ..., q$, and thus $P'(A) = \frac{1}{q}P'(\bigcup\limits_{i=1}^{q} C_i) = \frac{1}{q}$. Therefore $P'(A) = \frac{P(A)}{P(C)}$, which is our claim.

(b) Consider the case $\frac{P(A)}{P(C)} = \frac{p}{q} \in \mathbb{Q}$, where $p, q \in \mathbb{N}$, $p, q > 0$, $p \leq q$.

Applying Corollary 2.2 to A, with $\alpha_i = \frac{1}{p}P(A)$ for $i = 1, ..., p$, there exist disjoint sets $A_1, ..., A_p \in \mathcal{A}$ such that $\bigcup\limits_{i=1}^{p} A_i = A$ and $P(A_i) = \frac{1}{p}P(A) = \frac{1}{q}P(C)$ for $i = 1, ..., p$. Since $\frac{P(A_i)}{P(C)} = \frac{1}{q}$ for $i = 1, ..., p$, it follows from case (a) that $P'(A_i) = \frac{P(A_i)}{P(C)}$ for $i = 1, ..., p$. Therefore $P'(A) = P'(\bigcup\limits_{i=1}^{p} A_i) = \frac{p}{q} = \frac{P(A)}{P(C)}$.

(c) Consider the general case $\frac{P(A)}{P(C)} = \beta \in \,]0, 1]$.

There is a strictly increasing sequence (β_n) in $]0, \beta[\cap \mathbb{Q}$ such that $\lim\limits_{n\to\infty} \beta_n = \beta$. Write $\gamma_n := \frac{P(C)}{P(A)}\beta_n$ for $n = 1, 2, ...$; obviously $\gamma_n \in \,]0, 1[$. We proceed to define inductively an expansive sequence (A_n) in $\mathcal{A}$, with $A_n \subseteq A$ and $P(A_n) = \beta_n P(C)$. For $n = 1$, by Theorem 2.1, there is $A_1 \in \mathcal{A}$, $A_1 \subseteq A$, such that $P(A_1) = \gamma_1 P(A) = \beta_1 P(C)$. For $n = 2$, by Theorem 2.1, there is $\widetilde{A}_2 \in \mathcal{A}$, $\widetilde{A}_2 \subseteq (A \setminus A_1)$, such that $P(\widetilde{A}_2) = \frac{\gamma_2 - \gamma_1}{1-\gamma_1}P(A \setminus A_1)$; let $A_2 := A_1 \cup \widetilde{A}_2$. We have

$$P(A_2) = \gamma_1 P(A) + \frac{\gamma_2 - \gamma_1}{1 - \gamma_1}(P(A) - \gamma_1 P(A)) = \gamma_2 P(A) = \beta_2 P(C)$$

Suppose now that $A_1, ..., A_n \in \mathcal{A}$ are defined, such that $A_{i-1} \subseteq A_i \subseteq A$ for $i = 2, ..., n$ and $P(A_n) = \beta_n P(C)$. By Theorem 2.1, there is $\widetilde{A}_{n+1} \in \mathcal{A}$, $\widetilde{A}_{n+1} \subseteq (A \setminus A_n)$, such that $P(\widetilde{A}_{n+1}) = \frac{\gamma_{n+1} - \gamma_n}{1-\gamma_n}P(A \setminus A_n)$; let $A_{n+1} := A_n \cup \widetilde{A}_{n+1}$. We have

$$P(A_{n+1}) = \gamma_n P(A) + \frac{\gamma_{n+1} - \gamma_n}{1 - \gamma_n}(P(A) - \gamma_n P(A)) = \gamma_{n+1} P(A) = \beta_{n+1} P(C)$$

which shows that the expansive sequence (A_n) is defined as intended. Since $\frac{P(A_n)}{P(C)} = \beta_n \in \mathbb{Q}$, it follows from case (b) that $P'(A_n) = \beta_n$ for $n = 1, 2, ...$ Therefore $P'(\bigcup\limits_{n=1}^{\infty} A_n) = \lim\limits_{n\to\infty} \beta_n = \beta$. On the other hand,

$$P(\bigcup_{n=1}^{\infty} A_n) = (\lim_{n\to\infty} \beta_n) P(C) = \beta P(C) = P(A)$$

Hence, from (A.A), we have $P'(A) = P'(\bigcup\limits_{n=1}^{\infty} A_n)$, and so $P'(A) = \beta = \frac{P(A)}{P(C)}$. ∎

Obviously (Example 2.1) the condition of P being non-atomic cannot be dropped in Theorem 2.3.

3. Bayesian parametric inference

In standard Bayesian parametric inference we consider a probability space $(S \times \Theta, \mathcal{B}_{n+k}, P)$, where S is a Borel set in $\mathbb{R}^n$, Θ is a (generalized) interval in $\mathbb{R}^k$, $\mathcal{B}_{n+k}$ is the Borel σ-algebra on $S \times \Theta$ and P is a (σ-additive) probability measure. Here S is interpreted as the *sample space* where the response vector Y takes values and Θ as the *parameter space*, each parameter θ determining a probability distribution for Y. Recall that the marginal distributions P_Y and P_θ are defined by $P_Y(A) := P(A \times \Theta)$, $P_\theta(B) := P(S \times B)$ for the corresponding Borelian sets A in S and B in Θ. In accordance to practice (see [3] and [2]; note that improper prior distributions are not being considered) we assume that in the parametric case P_θ is non-atomic. We shall refer to $(S \times \Theta, \mathcal{B}_{n+k}, P)$ as the Bayesian parametric model.

For proofs of the following proposition see [1] or [4].

Proposition 3.1. *Any atom of a Borel measure on a second countable Hausdorff space includes a singleton of positive measure.*

Our last result is now immediate.

Proposition 3.2. *Let $(S \times \Theta, \mathcal{B}_{n+k}, P)$ be the Bayesian parametric model. Then P is non-atomic.*

Proof. By Proposition 3.1, if P had an atom, then it would include a singleton of positive measure, which contradicts that P_θ is non-atomic. ∎

If the Bayesian parametric model is considered as a valid formulation of a statistical problem (essentially, if $S \times \Theta$ can be given a joint probability distribution), we conclude (taking (A.A) for granted) from Theorem 2.3 and Proposition 3.2 that (H.2) follows, and thus Bayes' formula for the posterior distribution can be applied (provided that the measure-theoretic hypotheses for the suitable representation of the probability distributions hold; see for instance [5]). Loosely speaking, the existence of a joint probability on $S \times \Theta$ is the only nontechnical hypothesis underlying Bayesian parametric statistical inference.

References

[1] Aliprantis, C.D.; Border K.C. (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd edition. Springer (Berlin, New York).

[2] Bernardo, J.M.; Smith, A.F.M. (2006): *Bayesian Theory*, 2nd edition. Wiley (Chichester).

[3] DeGroot, M.H. (1970): *Optimal Statistical Decisions*. McGraw-Hill (New York).

[4] Dudley, R.M.; Norvaisa, R. (2011): *Concrete Functional Calculus*. Springer (New York).

[5] Ghosal, S.; van der Vaart, A. (2017): *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press (Cambridge).

[6] Johnson, R.A. (1970): "Atomic and nonatomic measures", *Proceedings of the American Mathematical Society*, **25**, 650-655.

[7] Krantz, D.H.; Luce, R.D.; Suppes, P.; Tversky, A. (1971): *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. Academic Press (New York).

[8] Pfeffer, W.F. (1977): *Integrals and Measures*. Dekker (New York, Basel).

[9] Sierpinski, W. (1922): "Sur les fonctions d'ensemble additives et continues", *Fundamenta Mathematicae*, **3**, 240-246.

[10] Villegas, C. (1964): "On qualitative probability σ-algebras", *Annals of Mathematical Statistics*, **35**, 1787-1796.