Revisiting Bayes’ Rule and Evidence: 
Notional Events vs. Real Data

Aris Spanos
Department of Economics,
Virginia Tech

October 2014

Abstract

The paper undertakes a conceptual scrutiny of Bayes’ rule in terms of the nature and interpretation of its probabilistic components, which reveals that there is nothing obvious or self-evident about it. First, it is not an instantiation of the conditional probability formula because (i) it involves conditioning on the unobservable event $H$ (a hypothesis), and (ii) it ignores the gap between Plato’s world ($H$) and the real world of evidence $E$ by assuming that the two overlap in an overly simplistic way. Second, the analogical reasoning used to convert Bayes’ rule from events to random variables is highly misleading without the relevant quantifiers. Any attempt to render meaningful the traditional Bayesian interpretation of its key elements, by attaching such quantifiers, belies the likelihood principle. These findings have important implications, not only for Bayesian statistics, but also for Bayesian epistemology and inductive logics, more generally.

Key words: Bayes’ rule; conditional probability; events vs. random variables; Bayesian epistemology; inductive logics; confirmation logic; statistical model; fallacy of rejection/acceptance; likelihood principle; Bayes factor; random variables vs. events
1 Introduction

Despite its long history and crucial importance in fields like statistics, Bayesian epistemology and inductive logic, the pertinence of Bayes’ rule:

\[ P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}, \quad P(E) > 0, \quad (1) \]

where \( H \) denotes a hypothesis and \( E \) the relevant evidence, has not been subjected to adequate scrutiny. This is primarily because (1) is widely viewed as an instantiation of the conditional probability formula, and thus self-evident.

This paper raises questions of both substance and interpretation by undertaking a conceptual scrutiny of Bayes’ rule, based on taking a closer look at the nature of probabilistic concepts involved. The resulting éclaircissement calls into question the conventional wisdom on a number of different grounds, including:

(a) Bayes’ rule is not an instantiation of the conditional probability formula since:
   (i) its conditioning in \( P(E|H) \) is potentially problematic since \( H \) is unobservable,
   (ii) Bayes’ rule views \( H \) and \( E \) as overlapping events associated with the same chance set-up even though \( H \) belongs to Plato’s world and \( E \) to the real world.
(b) The traditional recasting of (1) in terms of random variables \( X \) and \( \theta \):

\[ f(\theta|x) = \frac{f(x|\theta) \cdot f(\theta)}{f(x)}, \quad f(x) > 0, \quad (2) \]

(see Howson and Urbach, 2006, p. 38, inter alia), where \( f(x|\theta) \) is interpreted as the conditional density of \( X \) given \( \theta \) and \( f(\theta) \) as the marginal density of \( \theta \), is formally incorrect without the quantifiers concerning the relevant values of \( X \) and \( \theta \). As shown in the sequel, when the quantifiers needed to render pertinent this interpretation are added, Bayes’ rule belies the Likelihood Principle; see Berger and Wolpert (1988).

One can dismiss the above issues as due to sloppy language and clumsy notation, but that will be a rushed decision because they conceal deeper foundational issues.

In addition to Bayesian statistics, the issues raised above should be of interest to Bayesian epistemology (Talbott, 2008), to inductive logic (Hawthorne, 2012) and confirmation logic (Eells and Fitelson, 2000); see also Joyce (2008). First, the current formal apparatus based on evidence \( E \) (potentially observable) and a hypothesis \( H \) (unobservable), viewed as events defined on the same chance set-up, is overly simplistic and highly artificial. Second, the impression that the event-based apparatus of Bayesian confirmation theory can accommodate cases where evidence comes in the form of data \( x_0 \) stems primarily from insufficient understanding of how events differ from random variables and how hypotheses differ from events; see Spanos (2010).

The paper sheds additional light on the crucial differences between the event-based and the random variables-based formalisms of probability, by unveiling the false analogical reasoning employed in relating the two. It is argued that the event-based framework is totally inadequate for the real world applications of data and evidence that statistics is concerned with. In particular, it precludes any serious discourse on confirmation that includes the more realistic case where evidence comes
in the form of data \( \mathbf{x}_0 := (x_1, x_2, \ldots, x_n) \), viewed as a realization of a sample (set of random variables) \( \mathbf{X} := (X_1, X_2, \ldots, X_n) \), and hypotheses are framed in terms of the unknown parameter(s) \( \theta \) in the context of a statistical model \( M_\theta(\mathbf{x}) \), \( \theta \in \Theta \), \( \mathbf{x} \in \mathbb{R}^n_\mathbf{X} \); note that random variables are denoted by capital and their values by small letters.

Section 2 provides a brief introduction to Bayes’ rule in terms of events \( \mathcal{E} \) and \( \mathcal{E}^c \), in an attempt to contrast it to the conditional probability formula and bring out some of the veiled but problematic features of the event-based formalism. Section 3 introduces the random variable-based framing and compares the two probabilistic formalisms. An attempt is made to provide enough background material in probability to render sections 2 and 3 self-contained. Section 4 discusses conditional distributions in the case of two random variables \( \mathbf{X} \) and \( \mathbf{Y} \) as a prelude to putting forward a proper rendering of the various components of Bayes’ rule. These results are then used in section 5 to call into question the basic components of the traditional interpretation of Bayes’ rule in the context of a statistical model \( M_\theta(\mathbf{x}) \).

2 The event based formalism of probability

This section summarizes the event-based formalism of probability theory with a view to compare and contrast it with the random-variable based in section 3.

Consider the probabilistic set-up described by the probability space \( (\Sigma, \mathcal{F}, \mathcal{P}(\cdot)) \), where \( \Sigma \) is the set of all possible distinct outcomes, \( \mathcal{F} \) is the event space, and \( \mathcal{P}(\cdot) \) a probability function. According to Kolmogorov’s (1933), \( (\Sigma, \mathcal{F}, \mathcal{P}(\cdot)) \) requires that:

(a) \( \Sigma \) is a well-defined set whose elements are all possible distinct outcomes,

(b) \( \mathcal{F} \) is a field (or algebra): a set of subsets of \( \Sigma \) which is closed under the set theoretic operations of union (\( \cup \)), intersection (\( \cap \)), and complementation (\( ^c \)), and represents events of interest and related events, and

(c) \( \mathcal{P}(\cdot) : \mathcal{F} \to [0, 1] \) a set function which satisfies the following axioms:

\[
\begin{align*}
[A1] & \quad \mathcal{P}(\Sigma) = 1, \\
[A2] & \quad \mathcal{P}(A) \geq 0, \text{ for any event } A \in \mathcal{F}, \\
[A3] & \quad \mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B), \text{ for } A \text{ and } B \text{ in } \mathcal{F} \text{ such that } A \cap B = \emptyset.
\end{align*}
\]

This mathematical set-up can be easily extended to allow for an infinite sequence of non-overlapping events defining a sigma-field \( \mathcal{F} \); see Billingsley (1995).

2.1 Conditional probabilities

For any two events \( A \) and \( B \) in \( \mathcal{F} \), the conditional probability formula is:

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ for } P(B) > 0.
\]

(3)

This formula treats the events \( A \) and \( B \) symmetrically, and thus:

\[
P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ for } P(A) > 0.
\]

(4)

Solving (3) and (4) for \( P(A \cap B) \) yields the multiplication formula:

\[
P(A \cap B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B).
\]

(5)
Substituting (5) into (3) yields an alternative but equivalent formula for conditional probability:

\[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \text{ for } P(B) > 0. \]  

(6)

When one uses the partition of \( S \) stemming from event \( A \) and its complement \( A^c \) to define the total probability formula:

\[ P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c), \text{ } P(A) > 0, \text{ } P(A^c) > 0, \]

a third formula for conditional probability arises:

\[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}, \text{ for } P(B) > 0. \]

(7)

More generally, the total probability formula holds for any set of events \( (A_1, A_2, ..., A_m) \):

\[ P(B) = \sum_{i=1}^{m} P(A_i) \cdot P(B|A_i), \]

when these events constitute a partition of \( S \) in the sense that:

\( A_1 \cup A_2 \cup ... \cup A_m = S, \text{ } A_i \cap A_j = \emptyset, \text{ for any } i \neq j, \text{ } i, j = 1, 2, ..., m. \)

**Example 1.** Consider the random experiment of tossing a coin twice:

\( S = \{(HH), (HT), (TH), (TT)\} \)

Let the events of interest be: \( A = \{(HH), (HT), (TH)\}, \text{ } B = \{(HT), (TH), (TT)\}. \)

These two events can be used to define the relevant field to be:

\( \mathcal{F} = \{S, \emptyset, A, B, A^c, B^c, A \cap B, A^c \cup B^c\} \)

(8)

where \( A^c = \{(TT)\}, \text{ } B^c = \{(HH)\}, \text{ } A^c \cup B^c = \{(HH), (TT)\}, \text{ } \) and \( A \cap B = \{HT\}, (TH)\}; \)

note that \( \mathcal{F} \) in (8) is closed under the set-theoretic operations of union (\( \cup \)), intersection (\( \cap \)), and complementation (\( ^c \)).

Assuming that the coin is fair, one can assign probabilities to all events in \( \mathcal{F} \):

\[ P(A) = .75, \text{ } P(B) = .75, \text{ } P(A \cap B) = .5, \text{ } P(A^c) = .25, \text{ } P(B^c) = .25, \text{ } P(A^c \cup B^c) = .5 \]

Hence, the conditional probability formulae in (3)-(4) yield:

\[ P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{3}, \text{ } P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{2}{3}. \]

**2.2 Bayes’ rule in terms of events**

The conditional probability formula in (6) is transformed into an *updating rule* by interpreting the two events \( A \) and \( B \) as a *hypothesis* \( H \) and *evidence* \( E \), respectively, to yield *Bayes’ rule*:

\[ P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}, \text{ } P(E) > 0. \]

(9)

Its components are traditionally interpreted as follows:

\[ P(H|E) \text{ represents the probability of } H \text{ given } E, \]
\[ P(E|H) \text{ represents the probability of } E \text{ given } H, \]
\[ P(H) \text{ represents the prior probability of } H, \]
\[ P(E) \text{ represents the marginal probability of } E. \]
(i) \( P(H|E) \) is the posterior probability of \( H \) given \( E \),
(ii) \( P(E|H) \) is the likelihood of \( E \) given \( H \),
(iii) \( P(H) \) is the prior probability of \( H \), and
(iv) \( P(E) \) is the initial probability of evidence \( E \).

The conventional wisdom that (9) constitutes an instance of (6) or (7) is misleading because the nature of the events involved and the assignment of probabilities raise legitimate questions of both substance and interpretation. Viewing Bayes’ rule in (9) in the context of the probability space \((S, \mathcal{F}, P(.))\) raises the following issues.

(A) For Bayes’ rule in (9) to represent an instantiation of the conditional probability formula (6), the events \( \mathcal{H} \) and \( \mathcal{E} \) are required to (i) be defined on the same event space \( \mathcal{F} \) and be (ii) observable and (iii) overlapping. Conditions (i)-(iii) are potentially problematic for Bayes’ rule since \( \mathcal{E} \) is, in principle, observable and lies in the real world, but \( \mathcal{H} \) is unobservable and lives in Plato’s world. Ignoring the gap between these two worlds by assuming they overlap in the above simplistic way raises a fundamental issue in empirical modeling. Indeed, this very issue had been a major stumbling block for the systematic development of statistics before the 1920s; see Fisher (1922), p. 311.

(B) It is not obvious how the likelihood function \( P(\mathcal{E}|\mathcal{H}) \) assigns a probability to \( \mathcal{E} \) by conditioning on an unobservable event \( \mathcal{H} \). How does one ‘condition on the occurrence of an unobservable event \( \mathcal{H} \)’ without running into an oxymoron? A generous possible interpretation might be that the conditioning is only notional in the sense that the hypothesis \( \mathcal{H} \) relates to a particular instance of the mechanism that gave rise to \( \mathcal{E} \). A generous interpretation might be that \( P(\mathcal{E}|\mathcal{H}) \) refers to the ‘objective probability of the occurrence of \( \mathcal{E} \) presuming that \( \mathcal{H} \) is true’. In practice, however, ‘presuming that \( \mathcal{H} \) is true’ could not represent the occurrence of an unobservable event as such. It could, however, be interpreted as an instantiation of a chance set-up pertaining to the mechanism giving rise to events like \( \mathcal{E} \); see section 5.1.

(C) Despite the bold move of merging hypotheses (Plato’s world) and evidence (real world) into the same \( S \), when it comes to acknowledging this overlap \( \mathcal{E} \cap \mathcal{H} \), Bayesians sidestep the issue by using the identity \( P(\mathcal{E} \cap \mathcal{H})=P(\mathcal{E}|\mathcal{H}) \cdot P(\mathcal{H}) \) in:

\[
P(H|E) = \frac{P(\mathcal{E} \cap \mathcal{H})}{P(\mathcal{E})}, \quad P(\mathcal{E}) > 0. \tag{10}
\]

It is presumed that the assignments \( P(\mathcal{E}|\mathcal{H}) \) and \( P(\mathcal{H}) \) are easier to justify!

(D) The most problematic of the probabilistic assignments (i)-(iv) is \( P(\mathcal{E}) \) because it’s not obvious where the probability could come from; see Earman (1992), p. 172. The Bayesians attempt to address this conundrum by defining (iv) in terms of (ii)-(iii) which they deem less questionable. In particular, they use \( \mathcal{H} \) and \( \mathcal{H}^c \) denoted by \( (\mathcal{H}^c, \) the complement of \( \mathcal{H} \) with respect to \( S \)), to define a partition of \( S \): \( S=\mathcal{H} \cup \mathcal{H}^c \), and then use:

\[
P(\mathcal{H} \cup \mathcal{H}^c) = P(\mathcal{H}) + P(\mathcal{H}^c) = P(S) = 1,
\]

to deduce the total probability formula:

\[
P(\mathcal{E}) = P(\mathcal{H}) \cdot P(\mathcal{E}|\mathcal{H}) + P(\mathcal{H}^c) \cdot P(\mathcal{E}|\mathcal{H}^c). \tag{11}
\]
The formula in (11) is often used to rewrite Bayes’ rule as:

\[ P(H|E) = \frac{P(E|H) \cdot P(H)}{P(H) \cdot P(E|H) + P(H^c) \cdot P(E|H^c)}, \quad P(E) > 0. \]  \tag{12}

The so-called Bayesian catchall factor \( P(E|H^c) \) in (12) has been severely criticized by Mayo (1996), pp. 116-118, as highly misleading in practice.

In summary, there is nothing obvious or self-evident about Bayes’ rule in (9). The conditional probabilities formula in (6), initially proposed by De Moivre (1718), is obvious and non-controversial because all events involved are observable and are naturally defined on the same probability space. Contrary to the conventional wisdom, the formula proposed by Bayes’ (1764) in (9) is not an instantiation of (6), because it raises several crucial issues of applicability, including (A)-(D) above.

2.3 Bayesian confirmation in terms of events

The Bayesian confirmation theory relies on comparing the prior with the posterior probability of a particular hypothesis \( H \):

[i] Confirmation: \( P(H|E) > P(H) \)

[ii] Disconfirmation: \( P(H|E) < P(H) \)

The degree of confirmation is evaluated using some measure \( c(H, E) \) of the ‘degree to which \( E \) raises the probability of \( H \)’. Examples of such Bayesian measures are:

\[ d(H, E) = P(H|E) - P(H), \quad m(H, E) = P(E|H) - P(E), \quad r(H, E) = \frac{P(H|E)}{P(H)}; \]

see Fitelson (1999). Using \( c(H, E) \) one can argue that:

Measure \( c(H, E) \) indicates that evidence \( E \) favors hypothesis \( H_1 \) over \( H_0 \), iff:

\[ c(H_1, E) > c(H_0, E). \]

For instance using the measure \( r(H, E) \) in the case of two competing hypotheses \( H_0 \) and \( H_1 \):

\[ \frac{P(H_1|E)}{P(H_1)} > \frac{P(H_0|E)}{P(H_0)} \quad \Rightarrow \quad \frac{P(E|H_1)}{P(E)} > \frac{P(E|H_0)}{P(E)} \quad \Leftrightarrow \quad \frac{P(E|H_1)}{P(E|H_0)} > 1 \]

where \( \frac{P(E|H_1)}{P(E|H_0)} \) is the (Bayesian) likelihood ratio. For comparison purposes let us contrast this to the ratio of posteriors:

\[ \frac{P(H_1|E)}{P(H_0|E)} = \frac{\frac{P(H_1)}{P(E|H_0)}}{\frac{P(H_0)}{P(E|H_0)}} = \left( \frac{P(E|H_1)}{P(E|H_0)} \right) \left( \frac{P(H_1)}{P(H_0)} \right) > 1, \]  \tag{13}

which is the product of the ‘likelihood ratio’ \( \frac{P(E|H_1)}{P(E|H_0)} \) and the ratio of the priors \( \frac{P(H_1)}{P(H_0)} \).

In light of the fact that the above Bayesian confirmation theory inherits all the weaknesses (A)-(D) raised above relating to the potential arbitrariness of the various probabilistic assignments pertaining to unobservable events, the whole confirmation endeavor seems like playing imaginary war games on a map without any actual connection to reality. This is primarily because the probabilistic assignments in (12) seem both arbitrary and non-testable.
3 The random variable-based formalism

What happens when the evidence $E$ comes in the form of data $x_0=(x_1, x_2, \ldots, x_n)$? Since data $x_0$ come in the form of numbers, we need recast the above mathematical formalization based on a probability space $(S, \mathcal{F}, P(\cdot))$ into a formulation that relates the events of interest to real numbers. In this section we take a digression to provide a self-contained discussion of the basic random variables formalism as it relates to evidence in terms of data $x_0$.

The key to such a recasting is provided by the notion of a random variable $X$, defined as a real-valued function from the set of outcomes $(S)$ to the real line $(\mathbb{R})$:

$$X(\cdot): S \rightarrow \mathbb{R},$$

that does not belie the event structure of interest in $\mathcal{F}$. That is, all the events of the form $\{s: X(s) = x\}$, for all values of $X$, need to belong to $\mathcal{F}$ i.e.

$$\{s: X(s) = x\} \in \mathcal{F}, \forall x \in \mathbb{R},$$

where "$\forall$" denotes "for all" and $\mathbb{R}=(-\infty, \infty)$ denotes the real line.

**Example 1** (continued). Let us return to the random experiment of tossing a coin twice:

$$S = \{(HH), (HT), (TH), (TT)\}$$

with the events of interest: $A=\{(HH), (HT), (TH)\}$, $B=\{(HT), (TH), (TT)\}$, defining the relevant field to be:

$$\mathcal{F} = \{S, \emptyset, A, B, A^c, B^c, A \cap B, A^c \cup B^c\}$$

(15)

Let us define two real-valued functions from $S$ to the real line:

$$X(HH)=X(HT)=X(TH)=1, \quad X(TT)=0,$$

$$Y(TT)=Y(HT)=Y(TH)=1, \quad Y(HH)=0.$$  

Do these functions define random variables with respect to $\mathcal{F}$? To answer that question one needs to check condition (14).

$$\{s: X(s)=1\}=A \in \mathcal{F}, \quad \{s: X(s)=0\}=A^c \in \mathcal{F},$$

$$\{s: Y(s)=1\}=B \in \mathcal{F}, \quad \{s: Y(s)=0\}=B^c \in \mathcal{F}.$$  

That is, all the events of the form $\{s: X(s)=x\}$ belong to $\mathcal{F}$, $\forall x \in \mathbb{R}$, and $\{s: Y(s)=y\}$ belong to $\mathcal{F}$, $\forall y \in \mathbb{R}$. Note that for all values $x$ of $X$ different from 0 and 1, $\{s: X(s)=x\}=\emptyset \in \mathcal{F}$; similarly for $Y$. In contrast, the real-valued function:

$$Z(HH)=Z(HT)=1, \quad Z(TH)=Z(TT)=0,$$

is not a random variable relative to $\mathcal{F}$ in (15) because: $\{s: Z(s)=1\}=\{(HH), (HT)\} \notin \mathcal{F}$, and thus condition (14) does not hold.

What have we achieved so far by using the real-valued function $X(\cdot)$? First, the set-to-point function $P(\cdot)$ has been replaced by a numerical point-to-point function:

$$f(x) = P(s: X(s)=x), \forall x \in \mathbb{R}_X,$$
known as the **density function**.

In the case of the random variables $X$ and $Y$, their density functions are:

$$
\begin{array}{c|c|c|c|c}
 x & f(x) & y & f(y) \\
 0 & .25 & 0 & .25 \\
 1 & .75 & 1 & .75 \\
\end{array}
$$

Second, one can extend the above density functions to cases where we do not assume that the coin is fair, but instead:

$$
\theta = P(X=1), \quad \theta = P(Y=1), \quad 0 \leq \theta \leq 1,
$$

defining the density functions in terms of the unknown parameter $\theta$:

$$
\begin{array}{c|c|c|c|c}
 x & f(x) & y & f(y) \\
 0 & 1-\theta & 0 & 1-\theta \\
 1 & \theta & 1 & \theta \\
\end{array}
$$

Indeed, these density functions can be expressed more compactly in terms of a formula, known as the **Bernoulli density**:

$$
f(x; \theta) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1, \quad \text{and} \quad f(y; \theta) = \theta^y (1-\theta)^{1-y}, \quad y = 0, 1.
$$

The mathematical structure of the probability space $(S, \mathcal{F}, P(.))$ ensures that a density function satisfies the properties: [a] $f(x) \geq 0$, $\forall x \in \mathbb{R}$, [b] $\sum_{x \in \mathbb{R}_X} f_x(x) = 1$.

**Discrete vs. continuous random variables.** The above definition of a random variables is confined to functions that take only discrete values, i.e. $S$ is countable.

(b) For an uncountable $S$, the function $X(.) : S \to \mathbb{R}$, is a random variable relative to a particular event space of interest $\mathcal{F}$, if:

$$
\{s: X(s) \leq x\} \in \mathcal{F}, \quad \forall x \in \mathbb{R}.
$$

That is, in this case the events of interest are defined by the half-closed interval $\{s: X(s) \leq x\}$. In this case the connection between $P(.)$ and $f(.)$ is slightly more complicated because it goes through the cumulative distribution function $F(.)$:

$$
P(\{s: X(s) \leq x\}) = F(x) = \int_{-\infty}^{x} f(u) du, \quad \forall x \in \mathbb{R}.
$$

However: $f(x) = \frac{dF(x)}{dx} \rightarrow \int_{a}^{b} f(x) dx = F(b) - F(a)$, for all continuity points of $f(x)$.

Having said that, there is one crucial difference that needs to be brought out. In the discrete random variable case one can think of a density function as assigning probabilities to all events of the form $\{s: X(s) = x\}$, $x \in \mathbb{R}_X$, since $f(.) : \mathbb{R}_X \to [0, 1]$, but in the continuous random variable case it does not since: $f(.) : \mathbb{R}_X \to [0, \infty)$. To avoid this problem, in the continuous case probabilities are assigned to small intervals: $P(x \leq X \leq x+dx) \simeq f(x)dx$, for some small $dx > 0$, where $dx \to 0$ is the infinitesimal interval used in calculus.

Using the concept of a random variable one can recast the original formalization $(S, \mathcal{F}, P(.))$ into something more data friendly:

$$(S, \mathcal{F}, P(.)) \overset{X(.)}{\longrightarrow} \{f(x; \theta), \quad \theta \in \Theta, \ x \in \mathbb{R}_X\},$$

where $f(x; \theta)$ denotes a density function and $\theta$ its unknown parameters.
3.1 Joint distributions and random samples

The notion of a random sample was initially based on the vague idea of a selecting a ‘sample’ from a population so as to ensure its ‘representativeness’. When formalized, this idea has become a set of random variables $X := (X_1, \ldots, X_n)$ which are Independent and Identically Distributed (IID). For simplicity, let us define IID for the simple case of two random variables:

(I) The random variables $X$ and $Y$ are said to be Independent if:

$$f(x, y) = f_X(x) \cdot f_Y(y), \quad \forall (x, y) \in \mathbb{R}_x \times \mathbb{R}_y,$$

where $f(x, y)$ denotes the joint distribution, $f_X(x)$ and $f_Y(y)$ the marginal distributions of the random variables $X$ and $Y$.

(ID) The random variables $X$ and $Y$ are said to be Identically Distributed if:

$$f_X(x) = f_Y(y), \quad \forall (x, y) \in \mathbb{R}_x \times \mathbb{R}_y, \quad x = y \text{ and } \mathbb{R}_x = \mathbb{R}_y.$$

**Example 1** (continued). Consider the two random variables $X$ and $Y$ defined above. The marginal distributions were derived above, but how does one define the joint density $f(x, y)$? The answer is in terms of events of the joint occurrence of $X$ and $Y$ of the form:

$$\{s: X(s) = x, Y(s) = y\} = \{s: X(s) = x\} \cap \{s: Y(s) = y\}, \quad \forall (x, y) \in \mathbb{R}_x \times \mathbb{R}_y.$$ 

In the above case, these joint events are:

$$\{s: X(s) = 0, Y(s) = 0\} = \emptyset, \quad \{s: X(s) = 0, Y(s) = 1\} = \{(TT)\},
\{s: X(s) = 1, Y(s) = 0\} = \{(HH)\}, \quad \{s: X(s) = 1, Y(s) = 1\} = \{(HT), (TH)\}.$$

Given that $\mathcal{F}$ in (15) is a field, it means that all the above events belong to $\mathcal{F}$ because they constitute intersections of the original events associated with $X$ and $Y$ separately. Hence, their joint distribution is defined by assigning probabilities to each of these joint events, given in Table 1. Note that the marginal densities can be derived from the joint density by summing over rows for $f_X(x) = \sum_{y \in \mathbb{R}_y} f(x, y)$, and over columns for $f_Y(y) = \sum_{x \in \mathbb{R}_x} f(x, y)$.

<table>
<thead>
<tr>
<th>$y$ \ $x$</th>
<th>0</th>
<th>1</th>
<th>$f_Y(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.00</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td>1</td>
<td>.25</td>
<td>.50</td>
<td>.75</td>
</tr>
<tr>
<td>$f_X(x)$</td>
<td>.25</td>
<td>.75</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1

Given that the marginal distributions of the random variables $X$ and $Y$ coincide, it implies that they are Identically Distributed. But are they Independent (I)? The answer is no because: $f(0, 0) = 0 \neq f_X(0) \cdot f_Y(0) = (.25)(.25)$, and (16) does not hold.

In the case where the coin is not known to be fair, the joint density can be written in terms of the unknown parameters $\theta = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$, such that $\theta_{11} + \theta_{10} + \theta_{01} + \theta_{00} = 1$:

$$f(x, y; \theta) = \theta_{11}^y \cdot \theta_{00}^{(1-x)(1-y)} \cdot \theta_{01}^{(1-x)y} \cdot \theta_{10}^{(1-y)x}, \quad x = 0, 1, \quad y = 0, 1$$

(17)
The notion of a random sample is defined in terms of the joint distribution of the sample \( X := (X_1, X_2, ..., X_n) \), say \( f(x_1, x_2, ..., x_n; \theta) \), using two assumptions:

(I) **Independence**: the sample \( X \) is said to be Independent (I) if the joint distribution splits up into a product of marginal distributions:

\[
f(x; \theta) = f_1(x_1; \theta_1) \cdot f_2(x_2; \theta_2) \cdot \ldots \cdot f_n(x_n; \theta_n) := \prod_{k=1}^{n} f_k(x_k; \theta_k), \quad \forall x \in \mathbb{R}_X^n,
\]

(ID) **Identically Distributed**: the sample \( X \) is said to be Identically Distributed (ID) if the marginal distributions are identical:

\[
f_k(x_k; \theta_k) = f(x; \theta), \quad \forall k = 1, 2, ..., n.
\]

Note that this means two things, the density functions have the same form and the unknown parameters are common to all of them.

### 3.2 The notion of a statistical model

In practice, assuming a random (IID) sample will simplify the derivations, but its appropriateness is questionable in many scientific fields, especially when data are observational. To allow for dependence and/or heterogeneity statistical models are generally specified in terms of the (joint) distribution of the sample \( X := (X_1, ..., X_n) \), \( f(x; \theta), \ x \in \mathbb{R}_X^n \), where \( \mathbb{R}_X^n \) denotes the sample space (the set of all possible values of \( X \)). The generic form of a statistical model is:

\[
\mathcal{M}_\theta(x) = \{ f(x; \theta), \ \theta \in \Theta \}, \ x \in \mathbb{R}_X^n, \ \forall \theta \in \Theta,
\]  

(18)

where \( \theta \) denotes the unknown parameters taking values in \( \Theta \), the parameter space. The observed data \( x_0 := (x_1, ..., x_n) \) represent a single value in \( \mathbb{R}_X^n \) and are viewed a ‘typical’ realization of the sample \( X \), whose probabilistic structure is described by \( f(x; \theta), \ x \in \mathbb{R}_X^n \). From the frequentist perspective the parameters \( \theta \) are viewed as unknown constants, and the statistical model (18) to be well-defined, the number of unknown parameters in \( \theta \), say \( m \), should be much smaller (\( \ll \)) than the number of observations \( n \), i.e. \( \Theta \subset \mathbb{R}^m, \ m \ll n \).

**Example 3: simple Normal model.** Consider the case where each random variable \( X_k \) in the sample \( X := (X_1, ..., X_n) \) is assumed to be Normal (N), Independent and Identically Distributed (NIID), denoted by:

\[
\mathcal{M}_\theta(x) : \ X_k \sim \text{NIID}(\mu, \sigma^2), \ k = 1, 2, ..., n, ...
\]

(19)

For this model:

\[
f(x; \theta) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi} \sigma} \exp(-\frac{1}{2\sigma^2}(x_k-\mu)^2) = (\frac{1}{\sigma\sqrt{2\pi}})^n \exp \left\{-\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k-\mu)^2 \right\}.
\]

In the context of \( \mathcal{M}_\theta(x) \), data \( x_0 \) can be assigned a probability using the distribution of the sample \( f(x_0; \theta) \), which also defines the likelihood function via:

\[
L(\theta; x_0) \propto f(x_0; \theta), \ \forall \theta \in \Theta.
\]

(20)
"... define the likelihood as a quantity proportional to the probability that, from a population having the particular value of $\theta$, a sample having the observed value $x_0$, should be obtained." (Fisher, 1921, p. 227)

It is important to note that the two worlds (Plato’s vs. the real world) are kept apart in this context using the distinction: $\theta \in \Theta$ vs. $x_0 \in \mathbb{R}_X^n$. In addition, there is nothing conditional in this set-up. $f(x; \theta), x \in \mathbb{R}_X^n$ is not the conditional distribution of the sample given $\theta$; it varies with $x$, and it can be evaluated for different values of $\theta$, depending on the nature of the inferential procedure that invokes $f(x; \theta)$. In contrast, the likelihood function $L(\theta; x_0), \forall \theta \in \Theta$, varies with $\theta$, and it is evaluated at $X=x_0$, aiming to straddle the two worlds! The key difference between $P(E|H)$ and $L(\theta; x_0)$ is that the latter is defined in terms of the chance mechanism, described by $f(x; \theta), x \in \mathbb{R}_X^n$, giving rise to events like $x_0$. It is not some arbitrary assignment of probability as in the case of $P(E|H)$, or worse $P(E)$.

**Example 3** (continued). In the context of (19) with $\sigma^2$ known, Neyman-Pearson (N-P) hypotheses such as:

$$H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0,$$  

(21)

are framed in terms of the unknown parameter $\mu$, which do not represent events but different data generating scenarios. The question posed in (21) is: were data $x_0$ generated by some value of $\mu$ less than $\mu_0$ or not? Indeed, hypotheses are ultimately concerned with learning from data about the ‘true’ $\mu$, say $\mu^*$ that could have given rise to data $x_0$. In general, the expression ‘$\theta^*$’ denotes the true value of $\theta$ is a shorthand for saying that ‘data $x_0$ constitute a realization of the sample $X$ with distribution $f(x; \theta^*)$’. By defining the partition of $\Theta=(-\infty, \infty)$ in terms of $\Theta_0=(-\infty, \mu_0]$ and $\Theta_1=(\mu_0, \infty)$, and the associated partition of $M_0(x), M_0(x)=\{f(x; \mu), \mu \in \Theta_0\}$ and $M_1(x)=\{f(x; \mu), \mu \in \Theta_1\}$, the hypotheses in (21) can be framed equivalently, but more perceptively, as:

$$H_0: f(x; \mu^*) \in M_0(x) \text{ vs. } H_1: f(x; \mu^*) \in M_1(x), x \in \mathbb{R}^n.$$  

(22)

The test statistic $d(X)=\sqrt{m} \bar{x}_n-\mu_0), \bar{x}_n=\frac{1}{n} \sum^n_{i=1} x_i$, in terms of which the optimal (UMP) N-P test for (21) is specified (Lehmann, 1986), is, in essence, the standardized difference between $\mu^*$ and $\mu_0$, with $\mu^*$ replaced by its best estimator $\bar{x}_n$.

This suggests that hypothesis testing is posing questions whose objective is to learn from data $x_0$ about the ‘true’ data generating mechanism $M^*(x)=\{f(x; \theta^*)\}, x \in \mathbb{R}_X^n$, and not about the occurrence/non-occurrence of particular events; Spanos (2013a). Part of the reason why the confusion between events and hypotheses is widespread in philosophy of science stems from the confusion between the two probabilistic set ups, the event-based and the random variable-based discussed above. The confusion is exacerbated by conflating hypothesis testing with prediction, because the latter is about unobserved events. For instance, one might want to predict $x_{n+1}$, having observed $x_0:=(x_1, ..., x_n)$; see Spanos (2010) for additional confusions associated with the base-rate fallacy.
4 Conditional distributions and Bayes’ rule
How do the formulae for conditional distributions (densities) differ from Bayes’ rule when framed in terms of random variables?

4.1 Conditional densities with two random variables
Bayesian statistics textbooks convert the conditional probability formulae (3), (5) and (6) between events $A$ and $B$ in terms of density functions for random variables $X$ and $Y$ as follows:

$$f(y|x) = \frac{f(x,y)}{f(x)}, \text{ for } f(x) > 0,$$  \hspace{1cm} (23)

$$f(x,y) = f(x|y) \cdot f(y) = f(y|x) \cdot f(x), \text{ for } f(x) > 0, \text{ } f(y) > 0,$$  \hspace{1cm} (24)

$$f(y|x) = \frac{f(x|y) \cdot f(y)}{f(x)}, \text{ for } f(x) > 0;$$ \hspace{1cm} (25)


The formulae (23)-(25), however, are technically incorrect, as well as highly misleading for several reasons.

(a) One cannot simply replace $A$ and $B$ with $Y$ and $X$ because a random variable is not an event in itself. One can define events using the random variable $X$ by choosing one of its values, say $X=x$, or the sigma-field $\sigma(X)$ generated by $X$; see Billingsley (1995).

(b) A density function is defined at particular values of each random variable, and thus for the formulae (23)-(25) to make probabilistic sense one needs to add the missing quantifiers for the relevant values of $X$ and $Y$.

(c) Although the conditioning in (3)-(4) in terms of events is symmetric, the conditional density is non-symmetric with respect to the random variables $X$ and $Y$.

The intuitive reason why the recasting of these joint and conditional formulae from events to random variables goes astray is that random variables take more than one value and it matters how one treats all its values. Hence, the claim that in the case of discrete random variables the formulae (23)-(25) coincide with those in terms of events has no merit. Indeed, for the discussion of conditional probabilities and Bayes’ rule the distinction between discrete and continuous random variables is largely irrelevant; just replace summation ($\sum$) with integration ($\int$) for the latter.

In light of (a)-(c), the proper way to define (23) is:

$$f(y|X=x) = \frac{f(X=x,y)}{f(X=x)}, \text{ for } f(x) > 0, \forall y \in \mathbb{R}_Y.$$ \hspace{1cm} (26)

The key issues in defining a conditional density are:

(i) The conditional density $f(y|X=x)$ retains only one value of the conditioning random variable $X$, the value $X=x$, but all values of the random variable $Y$.

(ii) For each value $X=x$, $f(y|X=x), \forall y \in \mathbb{R}_Y$, defines a different conditional distribution, each of which constitutes a proper density function since:
in terms of random variables because:

\[ f(y|X=x) \geq 0, \quad \forall y \in \mathbb{R}, \quad \text{and} \quad \sum_{y \in \mathbb{R}_Y} f(y|X=x)dy=1. \]

(iii) Due to (i)-(ii), the multiplication rule in terms of events (5) is **no longer valid** in terms of random variables because:

\[ \{ f(y|X=x) \cdot f(X=x), \forall y \in \mathbb{R}_Y \} \neq \{ f(x|Y=y) \cdot f(Y=y), \forall x \in \mathbb{R}_X \} \] (27)

The multiplication rule only holds when \( X \) and \( Y \) are allowed to take all their values:

\[ f(x, y) = f(y|X=x) \cdot f(X=x) = f(x|Y=y) \cdot f(Y=y), \quad \forall x \in \mathbb{R}_X, \forall y \in \mathbb{R}_Y. \] (28)

The double quantifiers in (28), however, essentially defines the joint distribution.

**Example 2.** To illustrate what the conditional density formula in (26) represents, let us evaluate \( f(y|X=x) \), \( \forall y \in \mathbb{R}_Y \) and \( f(x|Y=y), \forall x \in \mathbb{R}_X \) in the case of the joint distribution in table 2.

<table>
<thead>
<tr>
<th>( y \setminus x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>( f_y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.20</td>
<td>.10</td>
<td>.15</td>
<td>.45</td>
</tr>
<tr>
<td>1</td>
<td>.10</td>
<td>.25</td>
<td>.05</td>
<td>.40</td>
</tr>
<tr>
<td>2</td>
<td>.01</td>
<td>.06</td>
<td>.08</td>
<td>.15</td>
</tr>
<tr>
<td>( f_x(x) )</td>
<td>.31</td>
<td>.41</td>
<td>.28</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2

These derivations will demonstrate that, in general, \( f(X=x, y) \) and \( f(x, Y=y) \) are not interchangeable and the formulae (23)-(25) are highly misleading without the relevant quantifiers! Indeed, the non-symmetric treatment of \( Y \) and \( X \) in conditioning, renders the traditional conversion of Bayes’ rule fallacious.

\[
\begin{align*}
f(y|x=1) &= \begin{cases} 
\frac{f(x=1,y=0)}{f_x(x=1)} &= \frac{20}{31}, & y=0 \\
\frac{f(x=1,y=1)}{f_x(x=1)} &= \frac{10}{31}, & y=1 \\
\frac{f(x=1,y=2)}{f_x(x=1)} &= \frac{01}{31}, & y=2 
\end{cases} \\
\begin{array}{c|c|c|c}
 y & 0 & 1 & 2 \\
\\
 f(y|x=1) & .645 & .323 & .032 \\
\end{array}
\]

\[
\begin{align*}
f(y|x=2) &= \begin{cases} 
\frac{f(x=2,y=0)}{f_x(x=2)} &= \frac{10}{31}, & y=0 \\
\frac{f(x=2,y=1)}{f_x(x=2)} &= \frac{25}{31}, & y=1 \\
\frac{f(x=2,y=2)}{f_x(x=2)} &= \frac{06}{31}, & y=2 
\end{cases} \\
\begin{array}{c|c|c|c}
 y & 0 & 1 & 2 \\
\\
 f(y|x=2) & .244 & .610 & .146 \\
\end{array}
\]

\[
\begin{align*}
f(y|x=3) &= \begin{cases} 
\frac{f(x=3,y=0)}{f_x(x=3)} &= \frac{15}{28}, & y=0 \\
\frac{f(x=3,y=1)}{f_x(x=3)} &= \frac{05}{28}, & y=1 \\
\frac{f(x=3,y=2)}{f_x(x=3)} &= \frac{08}{28}, & y=2 
\end{cases} \\
\begin{array}{c|c|c|c}
 y & 0 & 1 & 2 \\
\\
 f(y|x=3) & .536 & .179 & .285 \\
\end{array}
\]

Similarly, the conditional density of \( X \) given \( Y=y \) take the form:

\[
f(X|Y=y) = \frac{f(x, Y=y)}{f(Y=y)}, \quad \text{for} \quad f(y) > 0, \quad \forall x \in \mathbb{R}_X. \] (30)
When evaluated using the joint density in table 2 yields:

\[
\begin{align*}
  f(x|y=0) &= \begin{cases} 
    \frac{f(x=1,y=0)}{f_y(y=0)} = 0.20, & x=1 \\
    \frac{f(x=2,y=0)}{f_y(y=0)} = 0.10, & x=2 \\
    \frac{f(x=3,y=0)}{f_y(y=0)} = 0.15, & x=3 \\
    \frac{f(x=1,y=1)}{f_y(y=1)} = 0.20, & x=1 \\
    \frac{f(x=2,y=1)}{f_y(y=1)} = 0.15, & x=2 \\
    \frac{f(x=3,y=1)}{f_y(y=1)} = 0.05, & x=3 \\
    \frac{f(x=1,y=2)}{f_y(y=2)} = 0.01, & x=1 \\
    \frac{f(x=2,y=2)}{f_y(y=2)} = 0.06, & x=2 \\
    \frac{f(x=3,y=2)}{f_y(y=2)} = 0.08, & x=3 
  \end{cases} \\
  f(x|y=1) &= \begin{cases} 
    \frac{f(x=1,y=0)}{f_y(y=0)} = 0.45, & x=1 \\
    \frac{f(x=2,y=0)}{f_y(y=0)} = 0.35, & x=2 \\
    \frac{f(x=3,y=0)}{f_y(y=0)} = 0.35, & x=3 \\
    \frac{f(x=1,y=1)}{f_y(y=1)} = 0.40, & x=1 \\
    \frac{f(x=2,y=1)}{f_y(y=1)} = 0.25, & x=2 \\
    \frac{f(x=3,y=1)}{f_y(y=1)} = 0.15, & x=3 \\
    \frac{f(x=1,y=2)}{f_y(y=2)} = 0.01, & x=1 \\
    \frac{f(x=2,y=2)}{f_y(y=2)} = 0.06, & x=2 \\
    \frac{f(x=3,y=2)}{f_y(y=2)} = 0.08, & x=3 
  \end{cases} \\
  f(x|y=2) &= \begin{cases} 
    \frac{f(x=1,y=0)}{f_y(y=0)} = 0.45, & x=1 \\
    \frac{f(x=2,y=0)}{f_y(y=0)} = 0.35, & x=2 \\
    \frac{f(x=3,y=0)}{f_y(y=0)} = 0.35, & x=3 \\
    \frac{f(x=1,y=1)}{f_y(y=1)} = 0.01, & x=1 \\
    \frac{f(x=2,y=1)}{f_y(y=1)} = 0.06, & x=2 \\
    \frac{f(x=3,y=1)}{f_y(y=1)} = 0.08, & x=3 \\
    \frac{f(x=1,y=2)}{f_y(y=2)} = 0.01, & x=1 \\
    \frac{f(x=2,y=2)}{f_y(y=2)} = 0.06, & x=2 \\
    \frac{f(x=3,y=2)}{f_y(y=2)} = 0.08, & x=3 
  \end{cases}
\end{align*}
\]

(31)

\[
\begin{array}{c|ccc}
  x & 1 & 2 & 3 \\
  \hline
  f(x|y=0) & .445 & .222 & .333 \\
  f(x|y=1) & .250 & .625 & .125 \\
  f(x|y=2) & .067 & .400 & .533 
\end{array}
\]

It is important to reiterate that the above results do not depend on the fact that the example used is in terms of discrete random variables. The only difference when \((X, Y)\) are continuous is that the summation will be replaced by integration.

### 4.2 Bayes’ rule in terms of two random variables

What does the non-validity of the multiplication formula in (27) imply for Bayes’ rule:

\[
P(H|E) = \frac{P(E \cap H)}{P(E)} = \frac{P(E|H) \cdot P(H)}{P(E)}, \quad P(E) > 0,
\]

(32)

when converted into random variables and density functions?

To avoid additional complications, let us ignore the issues (A)-(D) raised above concerning the non-observability of \(H\), and pretend that \(H\) and \(E\) constitute well-defined generic events on a probability space \((S, \mathcal{F}, P(.))\). Replacing \(H\) with \(Y\) and \(E\) with \(X\), the proper conversion of (32) is:

\[
f(y|X=x) = \frac{f(X=x, y)}{f(X=x)} = \frac{f(X=x|y) f(y)}{f(X=x)}, \quad \text{for } f(x) > 0, \quad \forall y \in \mathbb{R}_Y.
\]

(33)

To avoid potential misunderstandings it is very important to emphasize that the above formula holds irrespective of whether the random variables \((X, Y)\) are discrete or continuous. Moreover, any attempt to interpret (33) in the case of discrete random variables as coinciding with (32) is erroneous due to the missing quantifier.

The question that naturally arises is: what does the key component \(\{f(X=x|y), \quad \forall y \in \mathbb{R}_Y\}\) in (33) represent? Contrary to cursory claims by many Bayesians, it is neither the conditional density of \(X\) given \(Y=y\), \(\{f(y|X=x), \forall y \in \mathbb{R}_Y\}\), nor the conditional density of \(Y\) given \(X=x\), \(\{f(X|Y=y), \forall x \in \mathbb{R}_X\}\). Instead, \(\{f(X=x|y), \forall y \in \mathbb{R}_Y\}\) in (33) represents a ‘cannibalized’ probability function that is not even a proper density in the sense that it sums to one.
Example 2 (continued). The evaluation of \( f(X=x|y), y \in \mathbb{R}_Y \), for the joint distribution in table 2 gives rise to table 3.

| \( y \) | \( f(X=2,y) \) | \( f(X=2|y) \) | \( f(y) \) |
|-------|-------------|-------------|--------|
| 0     | .10         | .222        | .45    |
| 1     | .25         | .625        | .40    |
| 2     | .06         | .400        | .15    |

| \( x \) | \( f(x=0|y) \) | \( f(x=1|y) \) | \( f(x=2|y) \) |
|---------|---------------|---------------|---------------|
| 0.10    |               |               |               |
| 0.222   |               |               |               |
| 0.45    |               |               |               |

Table 3

\[ \sum_{y \in \mathbb{R}_Y} f(X=2|y) = 1.247, \quad \sum_{y \in \mathbb{R}_Y} f(X=2,y) = .41 = f(X=2), \]  \hspace{1cm} (35)

(i) \( f(X=2,y) = f(X=2|y) \cdot f(y) \) is not a proper joint density, and when summed over all values of \( Y \), yields \( f(X=2) \), which is not the marginal distribution of \( X \). 

(ii) \( f(X=x|y), y \in \mathbb{R}_Y \) takes one piece from each of the three conditional density functions \( f(x|Y=y) \), associated with \( X=2 \), and all three values of \( Y \):

\[
\begin{align*}
  f(x=2|y=0) &= \frac{f(x=2,y=0)}{f(y=0)} = .222, \\
  f(x=2|y=1) &= \frac{f(x=2,y=1)}{f(y=1)} = .625, \\
  f(x=2|y=2) &= \frac{f(x=2,y=2)}{f(y=2)} = .4.
\end{align*}
\]

These represent pieces from the three conditional densities in (31). Note that when the probabilities from these pieces are rescaled by \( f(X=2) \), they do sum up to one:

\[ \sum_{y \in \mathbb{R}_Y} \frac{f(X=2|y) \cdot f(y)}{f(X=2)} = 1. \]  \hspace{1cm} (36)

(iii) In contrast to (33), the components of (32) represent proper joint, conditional and marginal probabilities.

5 Bayes’ rule in the context of a statistical model

How does \( f(X=x|y), y \in \mathbb{R}_Y \) being an improper density function affect the interpretation of Bayes’ rule (9) in the context of a statistical model \( M_\theta(x) \)?

5.1 Revisiting Bayes’ rule

In the context of a statistical model \( M_\theta(x), x \in \mathbb{R}_X^n, \forall \theta \in \Theta \), with a prior \( \pi(\theta), \forall \theta \in \Theta \), Bayes’ rule in (9) – when framed properly – takes the form:

\[
\pi(\theta|x_0) = \frac{f(x_0|\theta) \cdot \pi(\theta)}{f(x_0)} = \frac{f(x_0|\theta) \cdot \pi(\theta)}{\int_{\theta \in \Theta} f(x_0|\theta) \cdot \pi(\theta) d\theta}, \quad \forall \theta \in \Theta, \]  \hspace{1cm} (36)

for \( f(x_0) > 0 \), where data \( x_0 \) represents a point in the sample space \( \mathbb{R}_X^n \). Invariably, however, Bayesian textbooks:

[a] neglect the subscript \( 0 \) for \( x_0 \), and/or [b] leave out the quantifier \( \forall \theta \in \Theta \).

For instance, according to Ghosh et al. (2006), Bayes’ rule takes the form:

\[
\pi(\theta|x) = \frac{f(x|\theta) \cdot \pi(\theta)}{\int_{\theta \in \Theta} f(x|\theta) \cdot \pi(\theta) d\theta}, \]  \hspace{1cm} (37)

15
“where \( \pi(\theta) \) is the prior density function and \( f(x|\theta) \) is the density of \( X \), interpreted as the conditional density of \( X \) given \( \theta \). The numerator is the joint density of \( \theta \) and \( X \) and the denominator is the marginal density of \( X \).” (p. 31)

When (37) is compared to (36), however, it is clear that these claims are incorrect:

[i] \( f(x_0|\theta) \) is not the conditional density of \( X \) given \( \theta \), because it has a fixed \( X=x_0 \), and a varying \( \theta \), \( \forall \theta \in \Theta \); see \( f(X=x|y) \), \( y \in \mathbb{R}_Y \) in table 3. Moreover, the conditional density of \( X \) given \( \theta \) calls for the quantifier \( \forall x \in \mathbb{R}_X^n \), and not \( \forall \theta \in \Theta \).

[ii] \( f(x_0|\theta) \cdot \pi(\theta) \), \( \forall \theta \in \Theta \) is neither a proper density nor the joint density of \( \theta \) and \( X \), because the latter requires a double quantifier: \( f(x, \theta) \), \( \forall \theta \in \Theta \), \( \forall x \in \mathbb{R}_X^n \).

These problems stem from the fact that the ‘multiplication rule’ for events in (5) holds for density functions only when the double quantifier is added:

\[
f(x, \theta) = f(x|\theta) \cdot f(\theta) = f(\theta|x) \cdot f(x), \quad \forall \theta \in \Theta, \quad \forall x \in \mathbb{R}_X^n.
\]

Note that the right hand side of (38) represents a reparameterization of the joint density. Conditional densities are defined for a particular value of the conditioning variable and all the values of the other random variable, i.e.

\[
f(x|\theta) = \frac{f(x, \theta)}{f(\theta)}, \quad \forall x \in \mathbb{R}_X^n, \quad f(\theta|x_0) = \frac{f(x_0, \theta)}{f(x_0)}, \quad \forall \theta \in \Theta.
\]

Hence, \( f(x, \theta) \), \( \forall \theta \in \Theta \) cannot be turned around to define \( f(x|\theta_0) \), \( \forall x \in \mathbb{R}_X^n \) or vice-versa, as in the case of events in (3)-(4).

[iii] \( \int_{\theta \in \Theta} f(x_0|\theta) \cdot \pi(\theta) \, d\theta = f(x_0) \) is just a scalar, and not the marginal density of \( X \), which is defined by \( f(x) \), \( \forall x \in \mathbb{R}_X^n \).

It will be a mistake to dismiss the above quotation from Ghosh et al. (2006) as just a case of using sloppy language and clumsy notation, since this interpretation is typical of Bayesian textbooks more generally; see Bernardo and Smith (1994), p. 129, Lindley (1965), p. 118, O’Hagan (1994), p. 4, and Robert (2004), pp. 8-9, inter alia. It will be equally ill-advised to dismiss [i]-[iii] as restating the obvious that ‘we all know that the numerator of Bayes’ rule is not a proper density function’. It is worth remembering that in the context of Bayesian inference, both \( X \) and \( \theta \) are viewed as random variables (vectors) with proper distributions. Hence, when \( \pi(\theta) \) is the prior density function and \( f(x|\theta) \) is ... the conditional density of \( X \) given \( \theta \), they should to be proper densities, without any scaling, unless \( \pi(\theta) \) is improper.

In summary, the key differences between the posterior in (36) from the usual conditional probability formula (33) are:

[a] \( f(x_0|\theta) \) is not a proper conditional distribution, but a set of pieces from different conditional distributions associated with different values of \( \theta \), e.g. \( f(y|x=2) \) in (29) is very different from \( f(x=2|y) \), \( \forall y \in \mathbb{R}_Y \), in (34), section 4.1.

[b] Conditioning on an unobservable \( \theta \) raises technical issues which are beyond the scope of the present paper, but see Renyi (1970), p. 259.
5.2 Bayes’ rule vs. the Likelihood Principle

Is there a form of Bayes’ rule that justifies the traditional interpretation of its various components? There is. In light of (38), the quotation from Ghosh et. al (2006) could be rendered formally correct when the double quantifier is added to (37):

$$\pi(\theta|x) = \frac{f(x|\theta) \cdot \pi(\theta)}{\int_{\theta \in \Theta} f(x|\theta) \cdot \pi(\theta) d\theta}, \forall \theta \in \Theta, \forall x \in \mathbb{R}^n_X.$$  \hspace{1cm} (40)

For (40), all the interpretations in the above quotation are valid since $f(x|\theta)$, $\pi(\theta)$ and $f(x)=\int_{\theta \in \Theta} f(x|\theta) \cdot \pi(\theta) d\theta$ are proper densities. However, due to the presence of the quantifier $\forall x \in \mathbb{R}^n_X$, the form of Bayes’ rule in (40) belies the Likelihood Principle.

**Likelihood Principle.** For inference purposes the only relevant sample information pertaining to $\theta$ is contained in the likelihood function $L(x_0|\theta)$, $\forall \theta \in \Theta$. Moreover, if $x_0$ and $y_0$ are two sample realizations contain the same information about $\theta$ if they are proportional to one another (Berger and Wolpert, 1988, p. 19).

This means that an inference procedure that allows for any realization $x \in \mathbb{R}^n_X$ other than data $x_0$ contravenes the Likelihood Principle. Such procedures include all forms of frequentist inference, including estimation, hypothesis testing and prediction, since the relevant error probabilities associated with such procedures depend crucially on the distribution of the sample $f(x;\theta)$, $\forall x \in \mathbb{R}^n_X$; see Cox and Hinkley (1974).

In light of the above discussion, Bayesians cannot pretend that (37) is a formally correct rendition of Bayes’ rule. They have to decide between:

[a] the formally correct rendering of Bayes’ rule in (36) that coheres with the Likelihood Principle, but calls into question the traditional Bayesian interpretation of its components $f(x_0|\theta)$, $\forall \theta \in \Theta$ and $f(x_0)$, or

[b] adopt the modified (formally correct) interpretation in (40) that coheres with the Bayesian interpretation of these components, but is at odds with the Likelihood Principle. Keeping both options renders the Bayesian approach logically incoherent!

5.3 How do Bayesians get away with such obfuscations?

The question that naturally arises at this stage is: ‘how do Bayesian statisticians get away with ignoring the issues [i]-[iii] in the context of Bayesian inference?’

The simple answer is that the improprieness of $f(x_0|\theta)$ goes unnoticed because in Bayesian inference there is never any actual conditioning. That is, the conditioning invoked by Bayesians is **notional**, not actual. Bayes’ rule involves a sleight of hand in the form of **reinterpretation** the distribution of the sample $f(x,\theta)$, $\forall x \in \mathbb{R}^n_X$, as the conditional density of $X$ given $\theta$, i.e. $f(x|\theta)$, $\forall x \in \mathbb{R}^n_X$ : “$f(x|\theta)$ is the density of $X$, interpreted as the conditional density of $X$ given $\theta$.” Ghosh et al. (2006), p. 31. In light of that reinterpretation, Bayesians use $f(x_0|\theta)$ to define the likelihood function:

$$L(\theta|x_0) \propto f(x_0|\theta), \forall \theta \in \Theta,$$  \hspace{1cm} (41)

which is notionally different from the frequentist definition in (20). Hence, when all is said and done, Bayesian inference uses nothing more than a reinterpreted form of distribution of the sample $f(x,\theta)$ to justify viewing $\theta$ as a random variable.
In light of the above discussion, the foundations of Bayesian inference need to be reexamined more thoroughly, paying attention to both the substance and the interpretation of the various components of Bayes’s rule. Pretending that Bayes’ rule follows from the conditional probability formulae will not do for the various reasons raised above. Bayes’ rule, the cornerstone of Bayesian inference, is grounded on a notional conditioning which, when formally justified to agree with the rhetoric, it contravenes the likelihood principle.

In addition, the notional reinterpretation of \( f(x, \theta) \) into \( f(x | \theta) \) renders Bayesian inference vulnerable to statistical misspecification; any of the probabilistic assumptions defining the statistical model \( M_\theta(x) \) are invalid for data \( x_0 \). A statistically misspecified model is based on an erroneous \( f(x, \theta) \) which yields an erroneous likelihood \( L(\theta | x_0) \), which in turn will lead to the wrong posterior \( \pi(\theta | x_0) \propto L(\theta | x_0) \cdot \pi(\theta) \), \( \forall \theta \in \Theta \), and thus unreliable inferences.

A typical Bayesian attitude toward the problem of statistical misspecification is exemplified by Kadane (2011) who argues:

"If you accept the argument of this book, likelihoods are just as subjective as priors, and there is no reason to expect scientists to agree on them in the context of an applied problem." (p. 445)

In truth, the validity of a statistical model (defining the likelihood) is neither subjective nor a matter of agreement. In contrast to the prior, the probabilistic assumptions defining \( M_\theta(x) \) are directly testable vis-a-vis data \( x_0 \), and the invalidity can be assessed using misspecification testing; Mayo and Spanos (2004).

### 5.4 An alternative formal justification for Bayes’ rule?

One possible way to provide a formal justification for Bayes’s rule, which avoids misusing the multiplication formulae for conditional distributions in (38), is to begin with a generic function of two arguments:

\[
\phi(\cdot, \cdot) : [\mathbb{R}_X^n \times \Theta] \rightarrow \mathbb{R}.
\]

The function \( \phi(\cdot, \cdot) \), for a fixed \( \theta \), say \( \theta_0 \), defines the distribution of the sample:

\[
\phi(x, \theta_0) = f(x, \theta_0), \forall x \in \mathbb{R}_X^n.
\]

In turn, \( f(x, \theta_0), \forall x \in \mathbb{R}_X^n \), determines the relevant sampling distribution \( F_n(t) \) of any statistic \( T_n = g(X_1, ..., X_n) \) via:

\[
F_n(t) = P(T_n \leq t) = \int_{\{x: g(x) \leq t\}} f(x; \theta_0) dx, t \in \mathbb{R},
\]

since they all involve a fixed (true or hypothesized) \( \theta_0 \); see Spanos (2013).

For a fixed \( x \), say data \( x_0 \), \( \phi(\cdot, \cdot) \) defines the frequentist likelihood function:

\[
\phi(x_0; \theta) = L(x_0, \theta), \forall \theta \in \Theta.
\]

It is important to note that such a function already exists (implicitly) in frequentist inference when one derives the Fisher information matrix \( I_n(\theta) \), since it requires derivatives with respect to \( \theta \), \( \left( \frac{\partial^2 \ln h(x, \theta)}{\partial \theta \partial \theta} \right) \), and then expectations with respect to \( X \):
\[ L_{n}(\theta) = E \left(-\frac{\partial^2 \ln f(x, \theta)}{\partial \theta \partial \theta} \right). \] Note that neither \(\ln f(x, \theta_0), \forall x \in \mathbb{R}^n\) nor \(L(x_0, \theta), \forall \theta \in \Theta\) would accord with both operations since \(\frac{\partial^2 \ln f(x, \theta)}{\partial \theta \partial \theta}\) and \(E \left(-\frac{\partial^2 \ln L(x_0, \theta)}{\partial \theta \partial \theta} \right)\) are dubious.

From the Bayesian perspective the function \(h(x, \theta), \forall x \in \mathbb{R}^n\) and \(\forall \theta \in \Theta\), could be interpreted as the joint distribution of \(X\) and \(\theta\). To avoid the problems raised by conditioning on unobservable random variables, one could condition on the \(\sigma\)-field generated by \(\theta\), denoted by \(\sigma(\theta)\), which is meaningful because the conditioning is with respect to all possible events generated by \(\theta\), not particular values which are unobservable. That is, one could reparameterize the joint distribution using conditioning on \(\sigma(\theta)\) to derive:

\[ h(x, \theta) = f(x|\sigma(\theta)) \cdot \pi(\theta), \forall x \in \mathbb{R}^n, \forall \theta \in \Theta, \]

where \(f(x|\sigma(\theta))\) is the conditional distribution of \(X\) given \(\theta\), with \(\theta\) being treated as a random variable by retaining its ‘randomness’ (distribution).

Using the \(h(\cdot, \cdot)\) function, Bayes’ rule takes the alternative form:

\[ \pi(\theta; x_0) = \frac{f(x_0|\sigma(\theta)) \cdot \pi(\theta)}{\int_{\theta \in \Theta} f(x_0|\sigma(\theta)) \cdot \pi(\theta) d\theta}, \forall \theta \in \Theta, \]

where the numerator is viewed as \(h(x, \theta)\) with \(x\) fixed: \(h(x_0, \theta) = f(x_0|\sigma(\theta)) \cdot \pi(\theta), \forall \theta \in \Theta\). That is, the posterior distribution in (45) is just a scaled version of \(h(x_0, \theta)\).

This derivation of Bayes’ rule has several advantages over the traditional one:
(a) It deals with the problem of conditioning on the unobservable \(\theta\),
(b) it addresses the derivation difficulties arising from the misuse of the multiplication formulae in (38),
(c) it acknowledges that \(f(x_0|\sigma(\theta))\) is not a proper conditional distribution,
(d) it explains why the denominator \(f(x_0) = \int_{\theta \in \Theta} f(x_0|\sigma(\theta)) \cdot \pi(\theta) d\theta\) is needed: to provide the necessary scaling for the numerator to integrate to one,
(e) it represents an instance of the correct rendition of Bayes’ rule in (40) when the latter is evaluated at data \(x_0\).

The last point brings out the fact that both (40) and (45) are at odds with viewing \(\pi(\theta|x)\) as a conditional distribution. The double quantifier in (40) essentially transforms it into the joint distribution \(\pi(\theta, x) = h(x, \theta), \forall x \in \mathbb{R}^n, \forall \theta \in \Theta\). Similarly, (45) fortifies that by indicating that any inference based on \(f(x_0|\sigma(\theta)) \cdot \pi(\theta)\) stems from the joint distribution \(h(\theta, x)\) as it relates to data \(x_0\). In both cases, conditioning is assigned the subordinate role of providing the parameterization of interest. This questions the most cherished arrogate of the Bayesian rhetoric, which is that inference is all about updating ones’ beliefs concerning \(\theta\) by conditioning on data \(x_0\).

5.5 Bayesian confirmation in the context of a statistical model

For Bayesian confirmation in the context of \(M_\theta(x)\), the likelihood \(L(\theta|x_0), \theta \in \Theta\) is determined by the probabilistic assumptions of \(M_\theta(x)\), evidence \(E\) is replaced with data \(x_0\), and hypotheses \(H\) as events are replaced with hypotheses framed in terms of the unknown parameter(s) \(\theta \in \Theta\).
In particular, Bayesian hypothesis testing revolves around the ratio of posteriors which, for different values of $\theta \in \Theta$, say $H_0$: $\theta = \theta_0$ and $H_1$: $\theta = \theta_1$, takes the form:

$$\frac{\pi(\theta_0|x_0)}{\pi(\theta_1|x_0)} = \left(\frac{L(\theta_0|x_0)}{L(\theta_1|x_0)}\right) \left(\frac{\pi(\theta_0)}{\pi(\theta_1)}\right), \tag{46}$$

representing the product of the ‘likelihood ratio’ $L(\theta_0|x_0)/L(\theta_1|x_0)$ and the ‘ratio of the priors’ $\pi(\theta_0)/\pi(\theta_1)$; this is directly analogous to (13).

The measure most often used to appraise the evidence for $H_0$ is the likelihood ratio, (euphemistically) referred to by Bayesians as the Bayes factor:

$$BF(x_0; \theta_0) = \frac{L(\theta_0|x_0)}{L(\theta_1|x_0)},$$

in conjunction with certain rules of thumb concerning different thresholds indicating the strength of evidence. A Bayes factor result $BF(x_0; \theta_0) > k$, for $k \geq 3.2$, indicates that data $x_0$ favors $H_0$ with the ‘strength of evidence’ increasing with $k$. In particular, for $3.2 \leq k < 10$ the evidence is substantial, for $10 \leq k < 100$ the evidence is strong, and for $k \geq 100$ is decisive; see Robert (2007).

The key weakness of the Bayesian approach as an evidential account is that the step from $[L(\theta_0|x_0)/L(\theta_1|x_0)] > k$ to inferring that $x_0$ provides weak or strong evidence for $H_0$, depending on the value of $k > 1$, renders the Bayes factor (likelihoodist ratio) highly susceptible to two serious fallacies:

[i] the fallacy of rejection: (mis)interpreting reject $H_0$ [evidence against $H_0$] as evidence for a particular $H_1$; this can easily arise when a test has very high power (e.g. $n$ is very large).

[ii] the fallacy of acceptance: (mis)interpreting accept $H_0$ [no evidence against $H_0$] as evidence for $H_0$; this can easily arise when a test has very low power (e.g. $n$ is very small); see Spanos (2013b). The source of this weakness is that the Bayes factor stems from its inability to provide an evidential account because:

(a) is often invariant to the sample size $n$, (e.g. when a sufficient statistic exists),

(b) it ignores all other values of $X$ by invoking the likelihood principle.

As a result, the Bayes factor dismisses the generic capacity (power) of the test to detect different discrepancies from $H_0$, in favor of comparing the likelihoods of different values of $\theta$ using arbitrary thresholds based on information other than the data; see Kass and Raftery (1995).

Whether data $x_0$ provide evidence for or against a particular hypothesis $H_0$ depends crucially on its generic capacity. The intuition behind this is that a rejection of $H_0$ based on a test with low power for detecting a particular discrepancy $\gamma \neq 0$ provides stronger evidence for the presence of $\gamma$ than using a test with much higher power; see Spanos (2013b). As pointed out by Mayo (1996), this intuition is completely at odds with the Bayesian and likelihoodist intuitions; see Howson and Urbach (2006). Mayo went on to propose a frequentist evidential account based on harnessing this intuition in the form of a post-data severity evaluation of the frequentist accept/reject results. This is based on custom-tailoring the generic capacity of the test in question
to establish the discrepancy $\gamma$ warranted by data $x_0$. As shown in Mayo and Spanos (2006), this post-data severity evaluation can be used to address these fallacies in the context of frequentist testing. In addition, the problem of model validation (establishing statistical adequacy) can be addressed using thorough misspecification testing. This ensures the validity of the inductive premises and secures the error reliability of inference: the actual error probabilities approximate closely the nominal ones.

In contrast, the Bayesian and likelihoodist approaches have no principled way to circumvent the fallacies of acceptance and rejection. Often, when faced with serious foundational problems, Bayesians seek refuge in the notion of a ‘loss function’ as the final arbiter, in an obvious attempt to sidestep the real issues. The aim is to transmute the original problem into a question of selecting an ‘appropriate’ loss function that depends on the particulars of the problem, and leave it at that; fudging the original problem. For instance, in his attempt to sidestep the problem of distinguishing between statistical and substantive significance raised by the Bayes factor, Robert (2014), p. 224, proposes to invoke a vague ‘loss function’ ... to be determined by circumstances. Such advocates disregard Fisher’s discerning observation that:

“In the field of pure research no assessment of the cost of wrong conclusions, or of delay in arriving at more correct conclusions can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence.” (Fisher, 1935, pp. 25-26)

6 Summary and conclusions

The conceptual scrutiny of Bayes’ rule (9), in terms of the nature and interpretation of its probabilistic components, revealed that there is nothing obvious or self-evident about it. It is not an instantiation of the conditional probability formula because (i) it involves conditioning on the unobservable event $H$ and (ii) it ignores the gap between Plato’s world ($H$) and the real world ($E$) by assuming that the two overlap in an overly simplistic way. In addition, the analogical reasoning used to convert Bayes’ rule from events to random variables is highly misleading without the relevant quantifiers. Any attempt to render meaningful the traditional Bayesian interpretation of its components, by attaching the relevant quantifiers (see (40)), belies the likelihood principle. Bayesian statisticians get away with such obfuscations because in practice they use a reinterpreted form of the distribution of the sample $f(x; \theta)$, which involves just a notional conditioning on $\theta$ to yield $f(x|\theta)$. These problems can be addressed when a proper form of conditioning for the unobservable $\theta$ is employed, but that calls for rethinking the traditional interpretation of the posterior distribution.

The above findings have important implications, not only for Bayesian statistics, but also for Bayesian epistemology and inductive logics, more generally. The current domination of formal epistemology and confirmation theory by the events-based framework makes the discussions in the philosophy of science literature seem disconnected from real scientific methodology. This is primarily because the various probabilistic assignments in Bayes’ rule seem both arbitrary and largely non-testable.
It also makes it hard to apply any lessons from philosophy to the real world, and inhibits communication between philosophers and scientists.

Much richer and considerably more relevant epistemologies and confirmation theories can be constructed in the context of a statistical model $M_\theta(x)$ and the accompanied model-based frequentist inference, supplemented with the post-data severity evaluation that provides an evidential account of inference. Such an epistemology will replace the current ‘pragmatic self-defeat test (Dutch Book Arguments) for epistemic rationality’ (Talbott, 2008), with an empirical justification: the soundness of the premises of $M_\theta(x)$ vis-a-vis data $x_0$ providing the justification for ampliative induction. The soundness of the premises can be objectively established using misspecification testing.

References


