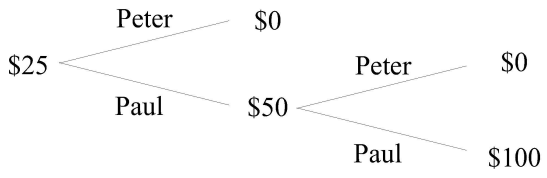
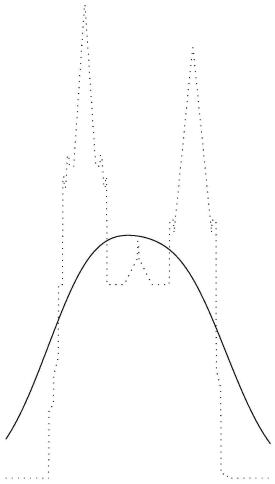
Test martingales, Bayes factors, and p-values

Glenn Shafer, Alexander Shen,
Nikolai Vereshchagin, and Vladimir Vovk





The Game-Theoretic Probability and Finance Project

Working Paper #33

First posted December 21, 2009. Last revised May 12, 2010.

Project web site:
http://www.probabilityandfinance.com

Abstract

A nonnegative martingale with initial value equal to one measures the evidence against a probabilistic hypothesis. The inverse of its value at some stopping time can be interpreted as a Bayes factor. It can be shown that if we exaggerate the evidence by considering the largest value attained so far by such a martingale, the exaggeration will not be great, and there are systematic ways to eliminate it. The inverse of the exaggerated value at some stopping time can be interpreted as a p-value. We give a simple characterization of all increasing functions that eliminate the exaggeration.

Contents

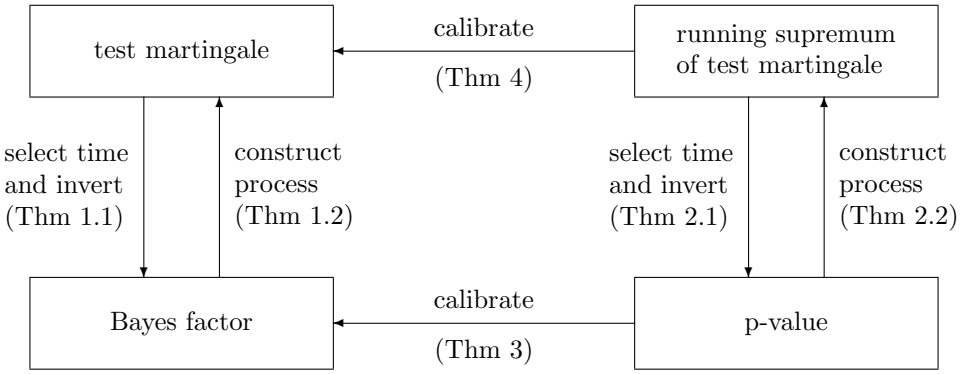


Figure 1: The relationship between a Bayes factor and a p-value can be thought of as a snapshot of the dynamic relationship between a nonnegative martingale (X_t) with initial value 1 and the process (X_t^*) that tracks its supremum. The snapshot could be taken at any time, but in our theorems we consider the final values of the martingale and its supremum process.

1 Introduction

Nonnegative martingales with initial value 1, Bayes factors, and p-values can all be regarded as measures of evidence against a probabilistic hypothesis. In this article, we review the well-known relationship between Bayes factors and nonnegative martingales and the less well-known relationship between p-values and the suprema of nonnegative martingales. Figure 1 provides a visual frame for the relationships we discuss.

Consider a random process (X_t) that initially has the value one and is a nonnegative martingale under a probabilistic hypothesis P (the time t may be discrete or continuous). We call such a martingale a *test martingale*. The values of a test martingale measure the changing evidence against P. When X_t becomes very large, P begins to look doubtful, but then X_u for some later time u may be lower and make P look better.

The notion of a test martingale (X_t) is related to the notion of a Bayes factor, which is more familiar to statisticians. A Bayes factor measures the degree to which a fixed body of evidence supports P relative to a particular alternative hypothesis Q; a very small value can be interpreted as discrediting P. If (X_t) is a test martingale, then the value $1/X_\tau$ for any stopping time τ is a Bayes factor. (For simplicity, the reader can replace everywhere a stopping time τ with a fixed value t.)

Suppose we exaggerate the evidence against P by considering not the current value X_t but the greatest value so far:

$$X_t^* := \sup_{s \le t} X_s.$$

A high X_t^* is not as impressive as a high X_t, but how should we understand the difference? Here are two complementary answers:

Answer 1 Although X_t^* is not a martingale, the final value $X_\infty^* \geq X_t^*$ still has a property associated with hypothesis testing: for every $\delta \in [0,1]$, $1/X_\infty^*$ has probability no more than δ of being δ or less. In this sense, $1/X_t^*$ is a p-value (perhaps conservative).

Answer 2 As we will show, there are systematic ways of adjusting X_t^* to eliminate the exaggeration. There exist, that is to say, functions f such that $\lim_{x\to\infty} f(x) = \infty$ and $f(X_t^*)$ is an unexaggerated measure of the evidence against P, inasmuch as there exists a test martingale Y_t always satisfying $Y_t \geq f(X_t^*)$ for all t.

Answer 2 will appeal most to readers familiar with the algorithmic theory of randomness, where the idea of treating a martingale as a dynamic measure of evidence is well established. Answer 1 may be more interesting to readers familiar with mathematical statistics, where the static notions of a Bayes factor and a p-value are often compared.

As we further show in this article, Answer 1 has a converse. For any random variable p that has the property that p has probability δ of being δ or less for every $\delta \in [0,1]$, there exists a test martingale (X_t) such that $p = 1/X_\infty^*$.

These results are probably known in one form or another to many people. But they have received less attention than they deserve, probably because the full picture emerges only when we bring together ideas from algorithmic randomness and mathematical statistics. Readers who are not familiar with both fields may want to read Section 2 for historical background. Others may turn directly to the mathematical exposition in Sections 3 to 7.

Section 3 is devoted to mathematical preliminaries; in particular, it defines test martingales and their wider conservative version, test supermartingales. Section 4 reviews the relationship between test supermartingales and Bayes factors, while Section 5 explains the relationship between the suprema of test supermartingales and p-values. Section 6 explains how p-values can be adjusted ("calibrated") so that they are not exaggerated relative to Bayes factors, and Section 7 explains how the maximal value attained so far by a test supermartingale can be similarly adjusted so that it is not exaggerated relative to the current value of a test supermartingale.

2 Some history

One source for the idea that a test martingale measures evidence is the work of Jean Ville, who introduced martingales into probability theory in his 1939 thesis [23]. Ville considered only test martingales and emphasized their betting interpretation. A test martingale under P is the capital process for a betting strategy that starts with a unit capital and bets at rates given by P, risking only the capital with which it begins. Such a strategy is an obvious way to test P: you refute the quality of P's probabilities by making money against them.

As Ville pointed out, the event that a test martingale tends to infinity has probability zero, and for every event of probability zero, there is a test martingale that tends to infinity if the event happens. Thus the classical idea that a probabilistic theory predicts events to which it gives probability equal (or nearly equal) to one can be expressed by saying that it predicts that test martingales will not become infinite (or very large). Ville's idea was popularized after World War II by Per Martin-Löf [14, 16] and subsequently developed by Claus-Peter Schnorr in the 1970s [21] and A. P. Dawid in the 1980s [5]. For details about the role of martingales in algorithmic randomness from von Mises to Schnorr, see [3]. For a historical perspective on the paradoxical behavior of martingales when they are not required to be nonnegative (or at least bounded below), see [4].

Ville's idea of a martingale was taken up as a technical tool in probability mathematics by Joseph Doob in the 1940s [13], and it subsequently became important as a technical tool in mathematical statistics, especially in sequential analysis and time series [11] and in survival analysis [1]. Mathematical statistics has been slow, however, to take up the idea of a martingale as a dynamic measure of evidence. Instead, statisticians emphasize a static concept of hypothesis testing.

Most literature on statistical testing remains in the static and all-or-nothing (reject or accept) framework established by Jerzy Neyman and Egon Pearson in 1933 [19]. Neyman and Pearson emphasized that when using an observation y to test P with respect to an alternative hypothesis Q we should reject P for values of y for which the likelihood ratio $P(y)/Q(y)$ is smallest or, equivalently, for which the reciprocal likelihood ratio $Q(y)/P(y)$ is largest.[1] If the observation y is a vector, say $y_1 \dots y_t$, where t continues to grow, then the reciprocal likelihood ratio $Q(y_1 \dots y_t)/P(y_1 \dots y_t)$ is a discrete-time martingale under P, but mathematical statisticians did not propose to interpret it directly. In the sequential analysis invented by Abraham Wald and George A. Barnard in the 1940s, the goal was to define an all-or-nothing Neyman-Pearson test by specifying a rule for stopping when $Q(y_1 \dots y_t)/P(y_1 \dots y_t)$ was large enough.

The increasing importance of Bayesian philosophy and practice starting in the 1960s has made the likelihood ratio $P(y)/Q(y)$ even more important. This ratio is now often called the Bayes factor for P against Q, because by Bayes's theorem, we obtain the ratio of P's posterior probability to Q's posterior probability by multiplying the ratio of their prior probabilities by this factor [10].

The notion of a p-value developed informally in statistics. In the late 19th and early 20th centuries, it was widely agreed in mathematical statistics that one should fix a threshold (subsequently called a *significance level*) for probabilities, below which a probability would be small enough to justify the rejection of a hypothesis. But because different people might fix this significance level differently, it was natural, in empirical work, to report the largest significance

[1] Here $P(y)$ and $Q(y)$ represent either probabilities assigned to y by the two hypotheses (discrete case) or, more generally, probability densities relative to a common reference measure. In the mathematical exposition that starts in Section 3, the two probabilistic hypotheses are represented by probability measures $\mathbf{P}$ and $\mathbf{Q}$ on a measurable space $(\Omega, \mathcal{F})$.

level for which the hypothesis would still have been rejected, and the English statisticians (e.g., Karl Pearson in 1900 [20] and R. A. Fisher in 1925 [7]) had the habit of calling this borderline probability "the value of P". Later, this became "P-value" or "p-value" [2].

After the work of Neyman and Pearson, which emphasized the probabilities of error associated with significance levels chosen in advance, mathematical statisticians often criticized applied statisticians for merely reporting p-values, as if a small p-value were a measure of evidence, speaking for itself without reference to a particular significance level. This disdain for p-values has been adopted and amplified by modern Bayesians, who have pointed to cases where p-values diverge widely from Bayes factors and hence are very misleading from a Bayesian point of view [22].

3 Mathematical preliminaries

In this section we define martingales, Bayes factors, and p-values. All three notions have two versions: the narrower "precise" version and the wider "conservative" version; the latter are often technically more useful. The conservative version of martingales is provided by supermartingales. As for Bayes factors and p-values, their main definitions will be conservative, but we will also define precise versions.

Recall that a *probability space* is a triplet $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω, and $\mathbf{P}$ is a probability measure on $\mathcal{F}$. A *random variable* X is a real-valued $\mathcal{F}$-measurable function on Ω; we allow random variables to take values $\pm\infty$. We use the notation $\mathbf{E}(X)$ for the integral of X with respect to $\mathbf{P}$ (undefined unless $\mathbf{P}\{X = \infty\} = 0$ or $\mathbf{P}\{X = -\infty\} = 0$), and $\mathbf{E}(X \mid \mathcal{G})$ for the conditional expectation of X given a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. A *random process* is a family (X_t) of random variables X_t; the index t is interpreted as time. We are mainly interested in the case where t is discrete, namely $t = 0, 1, 2, \ldots$; in particular, this was our assumption in Sections 1 and 2. However, our results (Theorems 1–4) will apply to both discrete t and continuous t (namely, $t \in [0, \infty)$).

Martingales and supermartingales

We will work with two standard definitions of martingales and supermartingales in a probability space:

1. $(X_t, \mathcal{F}_t)$, where t ranges over an ordered set ($\{0, 1, \ldots\}$ or $[0, \infty)$ in this article), is a *supermartingale* if $(\mathcal{F}_t)$ is a filtration (i.e., an indexed set of sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{F}_s \subseteq \mathcal{F}_t$ whenever $s < t$), (X_t) is a random process adapted with respect to $(\mathcal{F}_t)$ (i.e., each X_t is $\mathcal{F}_t$-measurable), each X_t is integrable, and
$$\mathbf{E}(X_t \mid \mathcal{F}_s) \le X_s$$

when $s < t$. A supermartingale is a *martingale* if, for all t and $s < t$,

$$\mathbf{E}(X_t \mid \mathcal{F}_s) = X_s. \tag{1}$$

2. A random process (X_t) is a *supermartingale* (resp. *martingale*) if $(X_t, \mathcal{F}_t)$ is a supermartingale (resp. martingale) where $\mathcal{F}_t$ is the σ-algebra generated by X_s, $s \leq t$.

For both definitions, the class of supermartingales contains that of martingales. In the case of continuous time we will assume that the paths of (X_t) are right-continuous (they will then automatically have left limits a.s.: see, e.g., [18], VI.3) and that the filtration $(\mathcal{F}_t)$ is *right-continuous*, in that, at each time t, $\mathcal{F}_t = \mathcal{F}_{t+} := \cap_{s>t}\mathcal{F}_s$ (we will, however, always indicate where this assumption is used, and we will never need the other of the "usual assumptions", the completeness of the σ-algebras $\mathcal{F}_t$).

We are particularly interested in *test* supermartingales: those that are nonnegative ($X_t \geq 0$ for all t) and have initial value X_0 equal to 1. A well-known fact about test supermartingales, first proven for discrete time and test martingales by Ville, is that

$$\mathbf{P}\{X_\infty^* \geq c\} \leq 1/c \tag{2}$$

for every $c \geq 1$ ([23], p. 100; [6], VI.1). We will call this the *maximal inequality*. This inequality shows that X_t can take value ∞ only with probability zero.

Bayes factors

A nonnegative measurable function $B : \Omega \to [0, \infty]$ is called a *Bayes factor for* $\mathbf{P}$ if $\int (1/B)\mathrm{d}\mathbf{P} \leq 1$; we will usually omit "for $\mathbf{P}$". A Bayes factor B is said to be *precise* if $\int (1/B)\mathrm{d}\mathbf{P} = 1$.

In order to relate this definition to the notion of Bayes factor discussed informally in Sections 1 and 2, we note first that whenever $\mathbf{Q}$ is a probability measure on $(\Omega, \mathcal{F})$, the Radon-Nikodym derivative $\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P}$ will satisfy $\int (\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P})\mathrm{d}\mathbf{P} \leq 1$, with equality if $\mathbf{Q}$ is absolutely continuous with respect to $\mathbf{P}$. Therefore, $B = 1/(\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P})$ will be a Bayes factor for $\mathbf{P}$. The Bayes factor B will be precise if $\mathbf{Q}$ is absolutely continuous with respect to $\mathbf{P}$; in this case B will be a version of the Radon-Nikodym derivative $\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{Q}$.

Conversely, whenever a nonnegative measurable function B satisfies $\int (1/B)\mathrm{d}\mathbf{P} \leq 1$, we can construct a probability measure $\mathbf{Q}$ that has $1/B$ as its Radon-Nikodym derivative with respect to $\mathbf{P}$. We first construct a measure $\mathbf{Q}_0$ by setting $\mathbf{Q}_0(A) := \int_A (1/B)\mathrm{d}\mathbf{P}$ for all $A \in \mathcal{F}$, and then obtain $\mathbf{Q}$ by adding to $\mathbf{Q}_0$ a measure that puts the missing mass $1 - \mathbf{Q}_0(\Omega)$ (which can be 0) on a set E (this can be empty or a single point) to which $\mathbf{P}$ assigns probability zero. (If $\mathbf{P}$ assigns positive probability to every element of Ω, we can add a new point to Ω.) The function B will be a version of the Radon-Nikodym derivative $\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{Q}$ if we redefine it by setting $B(\omega) := 0$ for $\omega \in E$ (remember that $\mathbf{P}(E) = 0$).

p-values

In order to relate p-values to supermartingales, we introduce a new concept, that of a p-test. A *p-test* is a measurable function $p : \Omega \to [0, 1]$ such that

$$\mathbf{P}\{\omega \mid p(\omega) \leq \delta\} \leq \delta \tag{3}$$

for all $\delta \in [0, 1]$. We say that p is a *precise p-test* if

$$\mathbf{P}\{\omega \mid p(\omega) \leq \delta\} = \delta \tag{4}$$

for all $\delta \in [0, 1]$.

It is consistent with the established usage to call the values of a p-test *p-values*, at least if the p-test is precise. One usually starts from a measurable function $T : \Omega \to \mathbb{R}$ (the *test statistic*) and sets $p(\omega) := \mathbf{P}\{\omega' \mid T(\omega') \geq T(\omega)\}$; it is clear that a function p defined in this way, and any majorant of such a p, will satisfy (3). If the distribution of T is continuous, p will also satisfy (4). If not, we can treat the ties $T(\omega') = T(\omega)$ more carefully and set

$$p(\omega) := \mathbf{P}\{\omega' \mid T(\omega') > T(\omega)\} + \xi \mathbf{P}\{\omega' \mid T(\omega') = T(\omega)\},$$

where ξ is chosen randomly from the uniform distribution on $[0, 1]$; in this way we will always obtain a function satisfying (4) (where $\mathbf{P}$ now refers to the overall probability encompassing generation of ξ).

4 Supermartingales and Bayes factors

When $(X_t, \mathcal{F}_t)$ is a test supermartingale, $1/X_t$ is a Bayes factor for any value of t. It is also true that $1/X_\infty$, X_∞ being the supermartingale's limiting value, is a Bayes factor. Part 1 of the following theorem is a precise statement of the latter assertion; the former assertion follows from the fact that we can stop the supermartingale at any point t. Part 2 of the theorem states that we can construct a test martingale whose limiting value is reciprocal to a given precise Bayes factor.

Theorem 1. *1. If $(X_t, \mathcal{F}_t)$ is a test supermartingale, then X_∞ exists almost surely and $1/X_\infty$ is a Bayes factor.*

2. Suppose B is a precise Bayes factor. Then there is a test martingale (X_t) such that $B = 1/X_\infty$. Moreover, for any filtration $(\mathcal{F}_t)$ such that B is $\mathcal{F}_\infty$-measurable, there is a test martingale $(X_t, \mathcal{F}_t)$ such that $B = 1/X_\infty$ almost surely.

Proof. If $(X_t, \mathcal{F}_t)$ is a test supermartingale, the limit X_∞ exists almost surely by Doob's convergence theorem (see, e.g., [18], VI.6) and the inequality $\int X_\infty \mathrm{d}\mathbf{P} \leq 1$ holds by Fatou's lemma:

$$\int X_\infty \mathrm{d}\mathbf{P} = \int \liminf_{t\to\infty} X_t \mathrm{d}\mathbf{P} \leq \liminf_{t\to\infty} \int X_t \mathrm{d}\mathbf{P} = 1.$$

Now suppose that B is a precise Bayes factor and $(\mathcal{F}_t)$ is a filtration such that B is $\mathcal{F}_\infty$-measurable. We can set $X_t := \mathbf{E}(1/B \mid \mathcal{F}_t)$ (requiring (X_t) to be right-continuous in the case of continuous time: cf. [18], VI.4(3); it is essential here that $(\mathcal{F}_t)$ should be right-continuous). Then $X_\infty = 1/B$ almost surely by Lévy's zero-one law ([12], pp. 128–130; [18], VI.6, corollary). If $(\mathcal{F}_t)$ such that B is $\mathcal{F}_\infty$-measurable is not given in advance, we can define it by, e.g.,

$$\mathcal{F}_t := \begin{cases} \{\emptyset, \Omega\} & \text{if } t < 1 \\ \sigma(B) & \text{otherwise} \end{cases}$$

(where $\sigma(B)$ is the σ-algebra generated by B). □

Formally, a *stopping time* with respect to a filtration $(\mathcal{F}_t)$ is defined to be a nonnegative random variable τ taking values in $[0, \infty]$ such that, at each time t, the event $\{\omega \mid \tau(\omega) \le t\}$ belongs to $\mathcal{F}_t$. Let $(X_t, \mathcal{F}_t)$ be a test supermartingale. Doob's convergence theorem, which was used in the proof of Theorem 1, implies that we can define its value X_τ at τ by the formula $X_\tau(\omega) := X_{\tau(\omega)}(\omega)$ even when $\tau = \infty$ with positive probability. Doob's stopping theorem (see, e.g., [18], VI.13) shows that the *stopped process* $(X_t^\tau, \mathcal{F}_t) := (X_{t \wedge \tau}, \mathcal{F}_t)$ (where $a \wedge b := \min(a, b)$) is also a nonnegative supermartingale. From part 1 of Theorem 1 we can now deduce that $1/X_\tau$ is a Bayes factor since X_τ is the final value of the stopped process. (However, the fact that $1/X_\tau$ is a Bayes factor also follows directly from Doob's stopping theorem.)

5 Supermartingales and p-values

Part 1 of the theorem that we prove in this section says that the inverse of a supremum of a test supermartingale is a p-test. This is true when the supremum is taken over $[0, t]$ for some time point t (or over $[0, \tau]$ for a stopping time τ), but the strongest statement obtains when we consider the supremum over all time points (i.e., for $\tau := \infty$). Part 2 of the theorem says that when we are given a precise p-test, we can construct a test martingale that has the inverse of the p-test as its supremum.

In order to understand part 2, we need to keep in mind that a p-test does not bring with it a structure indexed by time. In order to construct from the p-test a test martingale (X_t) whose supremum is the inverse of the p-test, we need also to construct a time scale. But the core idea can be explained more transparently in the case of a test with discrete levels, because constructing a time scale and a martingale is then simply a matter of constructing a sequence of bets. For simplicity, consider a function $T : \Omega \to \{0, 1, 2, \ldots\}$ such that, for each $n = 0, 1, 2, \ldots$, $\mathbf{P}\{T \ge n\} = 2^{-n}$. (This is a test of randomness in the sense of Martin-Löf [15]; Martin-Löf has "$\le 2^{-n}$" instead of "$= 2^{-n}$" in his general definition. The function 2^{-T} is essentially a p-test, if we ignore small constant factors.) The martingale X that achieves $2^{T(\omega)}$ as its supremum $\sup_t X_t(\omega)$ is constructed as the capital process of the following betting strategy: start with

capital $X_0 = 1$; gamble everything on the event $T \geq 1$, so that

$$X_1(\omega) = \begin{cases} 2 & \text{if } T(\omega) \geq 1 \\ 0 & \text{otherwise;} \end{cases}$$

then gamble everything on the event $T \geq 2$, so that

$$X_2(\omega) = \begin{cases} 4 & \text{if } T(\omega) \geq 2 \\ 0 & \text{otherwise;} \end{cases}$$

etc. Here is our formal result:

Theorem 2. *1. If $(X_t, \mathcal{F}_t)$ is a test supermartingale, $1/X^*_\infty = 1/\sup_t X_t$ is a p-test.*

2. If p is a precise p-test, there is a test martingale (X_t) such that $p = 1/X^*_\infty$.

Proof. The inequality $\mathbf{P}\{1/X^*_\infty \leq \delta\} \leq \delta$ for test supermartingales follows from the maximal inequality (2).

In the opposite direction, let p be a precise p-test. Set $\Pi := 1/p \in [1, \infty]$. Define a right-continuous random process (X_t), $t \in [0, \infty)$, by

$$X_t(\omega) = \begin{cases} 1 & \text{if } t \in [0, 1) \\ t & \text{if } t \in [1, \Pi(\omega)) \\ 0 & \text{otherwise.} \end{cases}$$

Since $X^*_\infty = \Pi$, it suffices to check that (X_t) is a test martingale. The time interval where this process is non-trivial is $t \geq 1$; notice that $X_1 = 1$ with probability one.

Let $t \geq 1$; we then have $X_t = t\,\mathbb{I}_{\{\Pi > t\}}$. Since X_t takes values in the two-element set $\{0, t\}$, it is integrable. The σ-algebra generated by X_t consists of 4 elements ($\emptyset$, Ω, the set $\Pi^{-1}((t, \infty])$, and its complement), and the σ-algebra $\mathcal{F}_t$ generated by X_s, $s \leq t$, consists of the sets $\Pi^{-1}(E)$ where E is either a Borel subset of $[1, t]$ or the union of $(t, \infty]$ and a Borel subset of $[1, t]$. To check (1), where $1 \leq s < t$, it suffices to show that

$$\int_{\Pi^{-1}(E)} X_t \mathrm{d}\mathbf{P} = \int_{\Pi^{-1}(E)} X_s \mathrm{d}\mathbf{P},$$

i.e.,

$$\int_{\Pi^{-1}(E)} t\,\mathbb{I}_{\{\Pi > t\}}\,\mathrm{d}\mathbf{P} = \int_{\Pi^{-1}(E)} s\,\mathbb{I}_{\{\Pi > s\}}\,\mathrm{d}\mathbf{P}, \tag{5}$$

where E is either a Borel subset of $[1, s]$ or the union of $(s, \infty]$ and a Borel subset of $[1, s]$. If E is a Borel subset of $[1, s]$, the equality (5) holds as its two sides are zero. If E is the union of $(s, \infty]$ and a Borel subset of $[1, s]$, (5) can be rewritten as

$$\int_{\Pi^{-1}((s,\infty])} t\,\mathbb{I}_{\{\Pi > t\}}\,\mathrm{d}\mathbf{P} = \int_{\Pi^{-1}((s,\infty])} s\,\mathbb{I}_{\{\Pi > s\}}\,\mathrm{d}\mathbf{P},$$

i.e., $t\mathbf{P}\{\Pi > t\} = s\mathbf{P}\{\Pi > s\}$, i.e., $1 = 1$. □

6 Calibrating p-values

An increasing (not necessarily strictly increasing) function $f : [0,1] \to [0,\infty]$ is called a *calibrator* if $f(p)$ is a Bayes factor for any p-test p. This notion was discussed in [24] and, less explicitly, in [22]. In this section we will characterize the set of all increasing functions that are calibrators; this result is a slightly more precise version of Theorem 7 in [24].

We say that a calibrator f *dominates* a calibrator g if $f(x) \le g(x)$ for all $x \in [0,1]$. We say that f *strictly dominates* g if f dominates g and $f(x) < g(x)$ for some $x \in [0,1]$. A calibrator is *admissible* if it is not strictly dominated by any other calibrator.

Theorem 3. *1. An increasing function $f : [0,1] \to [0,\infty)$ is a calibrator if and only if*

$$\int_0^1 \frac{\mathrm{d}x}{f(x)} \le 1. \tag{6}$$

2. Any calibrator is dominated by an admissible calibrator.

3. A calibrator is admissible if and only if it is left-continuous and

$$\int_0^1 \frac{\mathrm{d}x}{f(x)} = 1. \tag{7}$$

Proof. Part 1 is proven in [24] (Theorem 7), but we will give another argument, perhaps more intuitive. The condition "only if" is obvious: every calibrator must satisfy (6) in order to transform the "exemplary" p-test $p(\omega) = \omega$ on the probability space $([0,1], \mathcal{F}, \mathbf{P})$, where $\mathcal{F}$ is the Borel σ-algebra on $[0,1]$ and $\mathbf{P}$ is the uniform probability measure on $\mathcal{F}$, into a Bayes factor. To check "if", suppose (6) holds and take any p-test p. The expectation $\mathbf{E}(1/f(p))$ depends on p only via the values $\mathbf{P}\{p \le c\}$, $c \in [0,1]$, and this dependence is monotonic: if a p-test p_1 is *stochastically smaller* than another p-test p_2 in the sense that $\mathbf{P}\{p_1 \le c\} \ge \mathbf{P}\{p_2 \le c\}$ for all c, then $\mathbf{E}(1/f(p_1)) \ge \mathbf{E}(1/f(p_2))$. This can be seen, e.g., from the well-known formula $\mathbf{E}(\xi) = \int_{c=0}^{\infty} \mathbf{P}\{\xi > c\}$, where ξ is a nonnegative random variable. The condition (6) means that the inequality $\mathbf{E}(1/f(p)) \le 1$ holds for our exemplary p-test p; since p is stochastically smaller than any other p-test, this inequality holds for any p-test.

Part 3 follows from part 1, and part 2 follows from parts 1 and 3. □

Equation (7) gives a recipe for producing admissible calibrators f: take any left-continuous decreasing function $g : [0,1] \to [0,\infty]$ such that $\int_0^1 g(x)\mathrm{d}x = 1$ and set $f(x) := 1/g(x)$, $x \in [0,1]$. We see in this way, for example, that

$$f(x) := x^{1-\alpha}/\alpha \tag{8}$$

is an admissible calibrator for every $\alpha \in (0,1)$; if we are primarily interested in the behavior of $f(x)$ as $x \to 0$, we should take a small value of α. This class of calibrators was found independently in [24] and [22].

The calibrators (8) shrink to 0 significantly slower than x as $x \to 0$. But there are evidently calibrators that shrink as fast as $x \ln^{1+\alpha}(1/x)$, or $x \ln(1/x) \ln^{1+\alpha} \ln(1/x)$, etc., where α is a positive constant. For example,

$$f(x) := \begin{cases} \alpha^{-1}(1+\alpha)^{-\alpha} x \ln^{1+\alpha}(1/x) & \text{if } x \le e^{-1-\alpha} \\ \infty & \text{otherwise} \end{cases} \tag{9}$$

is a calibrator for any $\alpha > 0$.

7 Calibrating the suprema of test supermartingales

Let us call an increasing function $f : [1, \infty) \to [0, \infty)$ a *martingale calibrator* if it satisfies the following property:

> For any probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and any test supermartingale $(X_t, \mathcal{F}_t)$ in this probability space there exists a test supermartingale $(Y_t, \mathcal{F}_t)$ such that $Y_t \ge f(X_t^*)$ for all t.

There are at least 32 equivalent definitions of a martingale calibrator: we can independently replace each of the two entries of "supermartingale" in the definition by "martingale", we can independently replace $(X_t, \mathcal{F}_t)$ by (X_t) and $(Y_t, \mathcal{F}_t)$ by (Y_t), and we can optionally allow t to take value ∞. The equivalence will be demonstrated in the proof of Theorem 4. Our convention is that $f(\infty) := \lim_{x \to \infty} f(x)$ (but remember that $X_t^* = \infty$ only with probability zero, even for $t = \infty$).

As in the case of calibrators, we say that a martingale calibrator f is *admissible* if there is no other martingale calibrator g such that $g(x) \ge f(x)$ for all $x \in [1, \infty)$ (g *dominates* f) and $g(x) > f(x)$ for some $x \in [1, \infty)$.

Theorem 4. *1. An increasing function $f : [1, \infty) \to [0, \infty)$ is a martingale calibrator if and only if*

$$\int_0^1 f(1/x) \mathrm{d}x \le 1. \tag{10}$$

2. Any martingale calibrator is dominated by an admissible martingale calibrator.

3. A martingale calibrator is admissible if and only if it is right-continuous and

$$\int_0^1 f(1/x) \mathrm{d}x = 1. \tag{11}$$

Proof. We start from the statement "if" of part 1. Suppose an increasing function $f : [1, \infty) \to [0, \infty)$ satisfies (10) and $(X_t, \mathcal{F}_t)$ is a test supermartingale.

By Theorem 3, $g(x) := 1/f(1/x)$, $x \in [0,1]$, is a calibrator, and by Theorem 2, $1/X^*_\infty$ is a p-test. Therefore, $g(1/X^*_\infty) = 1/f(X^*_\infty)$ is a Bayes factor, i.e., $\mathbf{E}(f(X^*_\infty)) \le 1$. As in the proof of Theorem 1, we set $Y_t := \mathbf{E}(f(X^*_\infty) \mid \mathcal{F}_t)$ obtaining a nonnegative martingale $(Y_t, \mathcal{F}_t)$ with initial value $Y_0 \le 1$ satisfying $Y_\infty = f(X^*_\infty)$ a.s. Since

$$Y_t = \mathbf{E}(f(X^*_\infty) \mid \mathcal{F}_t) \ge \mathbf{E}(f(X_t) \mid \mathcal{F}_t) = f(X_t)$$

(this includes $t = \infty$) and we can make $(Y_t, \mathcal{F}_t)$ a test martingale by dividing each Y_t by $Y_0 \le 1$, the statement "if" in part 1 of the theorem is proven. Notice that our argument shows that f is a martingale calibrator in any of the 32 senses; this uses the fact that (Y_t) is a test supermartingale whenever $(Y_t, \mathcal{F}_t)$ is a test supermartingale.

Let us now check that any martingale calibrator (in any of the senses) satisfies (10). By any of our definitions of a martingale calibrator, we have $\int f(X^*_t) \mathrm{d}\mathbf{P} \le 1$ for all test martingales (X_t) and all $t < \infty$. It is easy to see that in Theorem 2, part 2, we can replace X^*_∞ with, say, $X^*_{\pi/2}$ by replacing the test martingale (X_t) whose existence it asserts with

$$X'_t := \begin{cases} X_{\arctan t} & \text{if } t < \pi/2 \\ X_\infty & \text{otherwise.} \end{cases}$$

Applying this modification of Theorem 2, part 2, to the precise p-test $p(\omega) := \omega$ on $[0,1]$ equipped with the uniform probability measure we obtain

$$\int_0^1 f(1/x) \mathrm{d}x = \mathbf{E}(f(X^*_{\pi/2})) \le \mathbf{E}(Y_{\pi/2}) \le 1.$$

This completes the proof of part 1.

Part 3 is now obvious, and part 2 follows from parts 1 and 3. □

As in the case of calibrators, we have a recipe for producing admissible martingale calibrators f provided by (11): take any left-continuous decreasing function $g : [0,1] \to [0,\infty)$ satisfying $\int_0^1 g(x) \mathrm{d}x = 1$ and set $f(y) := g(1/y)$, $y \in [1,\infty)$. In this way we obtain the class of admissible martingale calibrators

$$f(y) := \alpha y^{1-\alpha}, \quad \alpha \in (0,1),$$

analogous to (8) and the class

$$f(y) := \begin{cases} \alpha(1+\alpha)^\alpha \frac{y}{\ln^{1+\alpha} y} & \text{if } y \ge e^{1+\alpha} \\ 0 & \text{otherwise,} \end{cases} \quad \alpha > 0,$$

analogous to (9).

Acknowledgements

A. Philip Dawid and Steven de Rooij's help is gratefully appreciated. Steven's thoughts on the subject of this article have been shaped by discussions with Peter Grünwald. Our work on the article has been supported in part by ANR grant NAFIT ANR-08-EMER-008-01 and EPSRC grant EP/F002998/1.

Technical appendix

In this appendix we will mainly discuss the case of continuous time; we will see that in this case the notion of a test martingale is not fully adequate for the purpose of hypothesis testing (Lemma 2). Fix a filtration $(\mathcal{F}_t)$; in this appendix we will only consider supermartingales $(X_t, \mathcal{F}_t)$, and we will abbreviate $(X_t, \mathcal{F}_t)$ to (X_t), or even to X_t or X.

In discrete time, there is no difference between using test martingales and test supermartingales for hypothesis testing: every test martingale is a test supermartingale, and every test supermartingale is dominated by a test martingale (according to Doob's decomposition theorem, [18], VII.1); therefore, using test supermartingales only allows discarding evidence as compared to test martingales. In continuous time, the difference between test martingales and test supermartingales is essential, as we will see below (Lemma 2). For hypothesis testing we need "local martingales", a modification of the notion of martingales introduced by Itô and Watanabe [8] and nowadays used perhaps even more often than martingales themselves in continuous time. This is the principal reason why in this article we use test supermartingales so often starting from Section 3.

Remember that a random process (X_t) is a *local* member of a class $\mathcal{C}$ of random processes (such as martingales or supermartingales) if there exists a sequence $\tau_1 \le \tau_2 \le \cdots$ of stopping times (called a *localizing sequence*) such that $\tau_n \to \infty$ a.s. and each stopped process $X_t^{\tau_n} := X_{t \wedge \tau_n}$ belongs to the class $\mathcal{C}$. A standard argument (see, e.g., [6], VI.29) shows that there is no difference between test supermartingales and local test supermartingales:

Lemma 1. *Every local test supermartingale (X_t) is a test supermartingale.*

Proof. Let $\tau_1, \tau_2, \ldots$ be a localizing sequence, so that $\tau_n \to \infty$ as $n \to \infty$ a.s. and each X^{τ_n}, $n = 1, 2, \ldots$, is a test supermartingale. By Fatou's lemma for conditional expectations, we have, for $0 \le s < t$:

$$\mathbf{E}(X_t \mid \mathcal{F}_s) = \mathbf{E}\left(\lim_{n\to\infty} X_t^{\tau_n} \mid \mathcal{F}_s\right) \le \liminf_{n\to\infty} \mathbf{E}\left(X_t^{\tau_n} \mid \mathcal{F}_s\right) \le \liminf_{n\to\infty} X_s^{\tau_n} = X_s.$$

In particular, taking $s = 0$ we obtain $\mathbf{E}(X_t) \le 1$. □

An adapted process (A_t) is called *increasing* if $A_0 = 0$ and its every path is right-continuous and increasing (as usual, not necessarily strictly increasing). According to the Doob-Meyer decomposition theorem ([6], Theorem VII.12), every test supermartingale (X_t) can be represented as the difference $X_t = Y_t - A_t$ of a local test martingale (Y_t) and an increasing process (A_t). Therefore, for the purpose of hypothesis testing in continuous time, local test martingales are as powerful as test supermartingales: every local test martingale is a test supermartingale, and every test supermartingale is dominated by a local test martingale.

In discrete time there is no difference between local test martingales and test martingales ([6], (VI.31.1)). In continuous time, however, the difference is essential. A standard example ([9]; see also [18], VI.21, and [6], VI.29) of a local

martingale which is not a martingale is $L_t := 1/\|W_t + e\|$, where W_t is Brownian motion in $\mathbb{R}^3$ and e is a vector in $\mathbb{R}^3$ such that $\|e\| = 1$ (e.g., $e = (1, 0, 0)$); L_t being a local martingale can be deduced from $1/\|\cdot\|$ (the Newtonian kernel) being a harmonic function on $\mathbb{R}^3 \setminus \{0\}$. The random process (L_t) is a local test martingale such that $\sup_t \mathbf{E}(L_t^2) < \infty$; nevertheless it fails to be a martingale. See, e.g., [17] (Example 1.140) for detailed calculations.

The local martingale $L_t := 1/\|W_t + e\|$ provides an example of a test supermartingale which cannot be replaced, for the purpose of hypothesis testing, by a test martingale. According to another version of the Doob-Meyer decomposition theorem ([18], VII.31), a supermartingale (X_t) can be represented as the difference $X_t = Y_t - A_t$ of a martingale (Y_t) and an increasing process (A_t) if and only if (X_t) belong to the class (DL). The latter is defined as follows: a supermartingale is said to be in (DL) if, for any $a > 0$, the system of random variables X_τ, where τ ranges over the stopping times satisfying $\tau \le a$, is uniformly integrable. It is known that (L_t), despite being uniformly integrable (as a collection of random variables L_t), does not belong to the class (DL) ([18], VI.21 and the note in VI.19). Therefore, (L_t) cannot be represented as the difference $L_t = Y_t - A_t$ of a martingale (Y_t) and an increasing process (A_t). Test martingales cannot replace local test martingales in hypothesis testing also in the stronger sense of the following lemma.

Lemma 2. *Let $\delta > 0$. It is not true that for every local test martingale (X_t) there exists a test martingale (Y_t) such that $Y_t \ge \delta X_t$ for all t.*

Proof. Let $X_t := L_t = 1/\|W_t + e\|$, and suppose there is a test martingale (Y_t) such that $Y_t \ge \delta X_t$ for all t. Let $\epsilon > 0$ be arbitrarily small. Since (Y_t) is in (DL) ([18], VI.19(a)), for any $a > 0$ we can find $C > 0$ such that

$$\sup_\tau \int_{\{Y_\tau \ge C\}} Y_\tau \mathrm{d}\mathbf{P} < \epsilon\delta,$$

τ ranging over the stopping times satisfying $\tau \le a$. Since

$$\sup_\tau \int_{\{X_\tau \ge C/\delta\}} X_\tau \mathrm{d}\mathbf{P} \le \sup_\tau \int_{\{Y_\tau \ge C\}} (Y_\tau/\delta) \mathrm{d}\mathbf{P} < \epsilon,$$

(X_t) is also in (DL), which we know to be false. □

References

[1] Odd Aalen, Per Kragh Andersen, Ørnulf Borgan, Richard Gill, and Niels Keiding. History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, 5(1), June 2009. Available at `www.jehps.net`.

[2] John Aldrich. P-VALUE and prob-value. *Earliest Known Uses of Some of the Words of Mathematics*, `http://jeff560.tripod.com/p.html`.

[3] Laurent Bienvenu, Glenn Shafer, and Alexander Shen. On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics*, 5(1), June 2009. Available at `www.jehps.net`.

[4] Bernard Bru, Marie-France Bru, and Kai Lai Chung. Borel and the St. Petersburg martingale. *Electronic Journal for History of Probability and Statistics*, 5(1), June 2009. Available at `www.jehps.net`.

[5] A. Philip Dawid. Statistical theory: the prequential approach. *Journal of the Royal Statistical Society A*, 147:278–292, 1984.

[6] Claude Dellacherie and Paul-André Meyer. *Probabilities and Potential B: Theory of Martingales*. North-Holland, Amsterdam, 1982.

[7] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

[8] Kiyosi Itô and Shinzo Watanabe. Transformation of Markov processes by multiplicative functionals. *Annales de l'institut Fourier*, 15:15–30, 1965.

[9] Guy Johnson and L. L. Helms. Class D supermartingales. *Bulletin of the American Mathematical Society*, 69:59–62, 1963.

[10] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[11] Tze Leung Lai. Martingales in sequential analysis and time series, 1945–1985. *Electronic Journal for History of Probability and Statistics*, 5(1), June 2009. Available at `www.jehps.net`.

[12] Paul Lévy. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, Paris, 1937. Second edition: 1954.

[13] Bernard Locker. Doob at Lyon. *Electronic Journal for History of Probability and Statistics*, 5(1), June 2009. Available at `www.jehps.net`.

[14] Per Martin-Löf. Algorithmen und zufällige Folgen. Vier Vorträge von Per Martin-löf (Stockholm) gehalten am Mathematischen Institut der Universität Erlangen-Nürnberg, 1966. This document, dated 16 April 1966, consists of notes taken by K. Jacobs and W. Müller from lectures by Martin-Löf at Erlangen on April 5, 6, 14, and 15. There are copies in several university libraries in Germany and the United States. Available at `http://www.probabilityandfinance.com/misc/erlangen.pdf`.

[15] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[16] Per Martin-Löf. The literature on von Mises' Kollektivs revisited. *Theoria*, 35:12–37, 1969.

[17] Péter Medvegyev. *Stochastic Integration Theory*. Oxford University Press, Oxford, 2007.

[18] Paul A. Meyer. *Probability and Potentials*. Blaisdell, Waltham, MA, 1966.

[19] Jerzy Neyman and Egon Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, 231:289–337, 1933.

[20] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

[21] Claus-Peter Schnorr. *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*. Springer, Berlin, 1971.

[22] Thomas Sellke, M. J. Bayarri, and James Berger. Calibration of p-values for testing precise null hypotheses. *American Statistician*, 55:62–71, 2001.

[23] Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

[24] Vladimir Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society B*, 55:317–351, 1993.