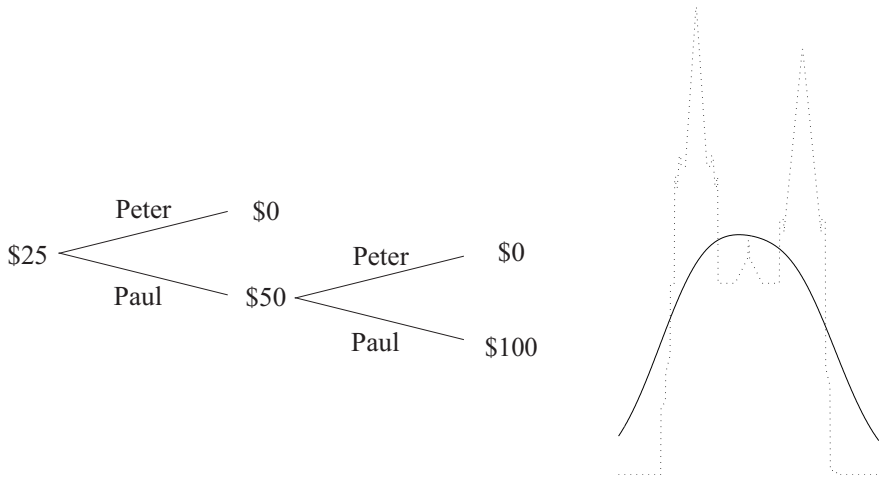
Test martingales, Bayes factors, and p-values

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 December 2, 2010.

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

Abstract

A nonnegative martingale with initial value equal to one measures evidence against a probabilistic hypothesis. The inverse of its value at some stopping time can be interpreted as a Bayes factor. If we exaggerate the evidence by considering the largest value attained so far by such a martingale, the exaggeration will be limited, 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

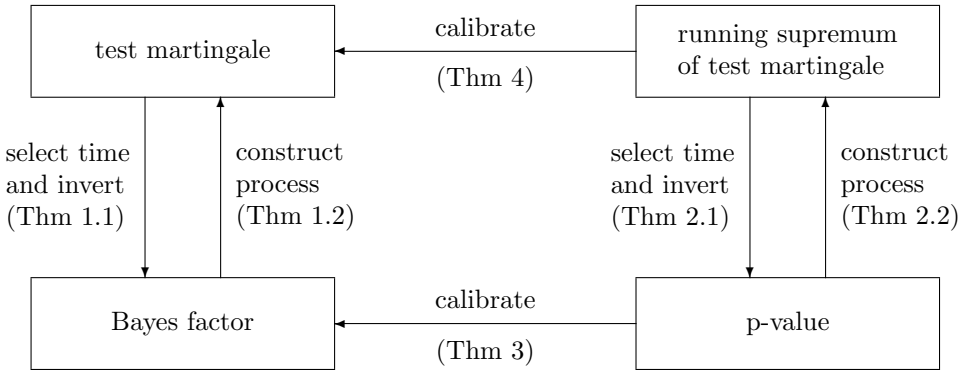


Figure 1: The relationship between a Bayes factor and a p-value can be thought of as a snapshot of the dynamic relationship between a nonnegative martingale (X_t) with initial value 1 and the process (X_t^*) that tracks its supremum. The snapshot could be taken at any time, but in our theorems we consider 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 (i.e., a simple statistical 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*. One statistical interpretation of the values of a test martingale is that they measure the changing evidence against P. The value X_t is the number of dollars a gambler has at time t if he begins with \$1 and follows a certain strategy for betting at the rates given by P; the nonnegativity of the martingale means that this strategy never risks a cumulative loss exceeding the \$1 with which it began. If X_t is very large, the gambler has made a lot of money betting against P, and this makes P 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 for any fixed time t, $1/X_t$ is a Bayes factor. We can also say, more generally, that the value $1/X_\tau$ for any stopping time τ is a Bayes factor. This is represented by the downward arrow on the left in Figure 1.

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 (downward arrow on the right in Figure 1) Although (X_t^*) is usually not a martingale, the final value $X_\infty^* := \sup_s X_s$ 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. For any t, X_t^*, because it is less than or equal to X_∞^*, has the same property. In this sense, $1/X_\infty^*$ and $1/X_t^*$ are p-values (perhaps conservative).

Answer 2 (leftward arrow at the top of Figure 1) As we will show, there are systematic ways of shrinking X_t^* (*calibrating* it, as we shall say) 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 evidence against P, inasmuch as there exists a test martingale (Y_t) always satisfying $Y_t \ge f(X_t^*)$ 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 (see, e.g., [25], Section 4.5.7). 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.

For the sake of conceptual completeness, we note that Answer 1 has a converse. For any random variable p that has probability δ of being δ or less for every $\delta \in [0,1]$, there exists a test martingale (X_t) such that $p = 1/X_\infty^*$. This converse is represented by the upward arrow on the right of our figure. It may be of limited practical interest, because the time scale for (X_t) may be artificial.

Parallel to the fact that we can shrink the running supremum of a test martingale to obtain an unexaggerated test martingale is the fact that we can inflate a p-value to obtain an unexaggerated Bayes factor. This is the leftward arrow at the bottom of our figure. It was previously discussed in [41] and [35].

These relationships are probably all 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 find the historical discussion in Section 2 helpful.

Although our theorems are not deep, we state and prove them using the full formalism of modern probability theory. Readers more comfortable with the conventions and notation of mathematical statistics may want to turn first to Section 8, in which we apply these results to testing whether a coin is fair.

The theorems depicted in Figure 1 are proven in Sections 3 to 7. Section 3 is devoted to mathematical preliminaries; in particular, it introduces the concept

of a test martingale and the wider and in general more conservative concept of a test supermartingale. 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 inflated 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 shrunk so that it is not exaggerated relative to the current value of a test supermartingale.

There are two appendices. Appendix A explains why test supermartingales are more efficient tools than test martingales in the case of continuous time. Appendix B carries out some calculations that are used in Section 8.

2 Some history

Jean Ville introduced martingales into probability theory in his 1939 thesis [39]. Ville considered only test martingales and emphasized their betting interpretation. As we have explained, 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 [27, 28] and subsequently developed by Claus-Peter Schnorr in the 1970s [34] and A. P. Dawid in the 1980s [11]. For details about the role of martingales in algorithmic randomness from von Mises to Schnorr, see [8]. For historical perspective on the paradoxical behavior of martingales when they are not required to be nonnegative (or at least bounded below), see [9].

Ville's idea of a martingale was taken up as a technical tool in probability mathematics by Joseph Doob in the 1940s [26], and it subsequently became important as a technical tool in mathematical statistics, especially in sequential analysis and time series [21] 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 [31]. Neyman and Pearson emphasized that when using an observation y to test P with respect to an alternative hypothesis Q, it is optimal to 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. (Here $P(y)$ and $Q(y)$ represent either probabilities assigned to y by the two hypotheses or, more generally, probability densities relative to a common reference measure.) 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 still is to define an all-or-nothing Neyman-Pearson test satisfying certain optimality conditions, although the reciprocal likelihood ratio plays an important role (when testing P against Q, this goal is attained by a rule that rejects P when $Q(y_1 \dots y_t)/P(y_1 \dots y_t)$ becomes large enough and accepts P when $Q(y_1 \dots y_t)/P(y_1 \dots y_t)$ becomes small 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 [20].

The notion of a p-value developed informally in statistics. From Jacob Bernoulli onward, everyone who applied probability theory to statistical data agreed that one should fix a threshold (later 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 threshold differently, it was natural, in empirical work, to report the smallest threshold for which the hypothesis would still have been rejected, and British statisticians (e.g., Karl Pearson in 1900 [32] and R. A. Fisher in 1925 [16]) sometimes called this borderline probability "the value of P". Later, this became "P-value" or "p-value" [3].

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 [35, 43].

3 Mathematical preliminaries

In this section we define martingales, Bayes factors, and p-values. All three notions have two versions: a narrow version that requires an equality and a wider version that relaxes this equality to an inequality and is considered conservative because the goal represented by the equality in the narrow version may be more than attained; the conservative versions are often technically more useful. The conservative version of a martingale is a supermartingale. As for Bayes factors and p-values, their main definitions will be conservative, but we will also define

narrow 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 $\mathbf{P}$ and $\mathbf{E}(X \mid \mathcal{G})$ for the conditional expectation of X given a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$; these notations are used only when X is integrable (i.e., when $\mathbf{E}(X^+) < \infty$ and $\mathbf{E}(X^-) < \infty$; in particular, $\mathbf{P}\{X = \infty\} = \mathbf{P}\{X = -\infty\} = 0$). 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 discrete time (say $t = 0, 1, 2, \ldots$), but our results (Theorems 1–4) will also apply to continuous time (say $t \in [0, \infty)$).

3.1 Martingales and supermartingales

The time scale for a martingale or supermartingale is formalized by a filtration. In some cases, it is convenient to specify this filtration when introducing the martingale or supermartingale; in others it is convenient to specify the martingale or supermartingale and derive an appropriate filtration from it. So there are two standard definitions of martingales and supermartingales in a probability space. We will use them both:

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
$$\mathbf{E}(X_t \mid \mathcal{F}_s) \leq X_s \quad \text{a.s.}$$
when $s < t$. A supermartingale is a *martingale* if, for all t and $s < t$,
$$\mathbf{E}(X_t \mid \mathcal{F}_s) = X_s \quad \text{a.s.} \tag{1}$$

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 always assume that the paths of (X_t) are right-continuous almost surely (they will then automatically have left limits almost surely: see, e.g., [13], VI.3(2)). We will also assume that the filtration $(\mathcal{F}_t)$ in $(X_t, \mathcal{F}_t)$ satisfies the *usual conditions*, namely that each σ-algebra $\mathcal{F}_t$ contains all subsets of all $E \in \mathcal{F}$ satisfying $\mathbf{P}(E) = 0$ (in particular, the probability space is complete) and that $(\mathcal{F}_t)$ is *right-continuous*, in that, at each time t, $\mathcal{F}_t = \mathcal{F}_{t+} := \cap_{s>t}\mathcal{F}_s$. If the original filtration $(\mathcal{F}_t)$ does not satisfy the usual conditions (this will often be the case when $\mathcal{F}_t$ is the σ-algebra generated by X_s, $s \leq t$), we can redefine $\mathcal{F}$ as the $\mathbf{P}$-completion $\mathcal{F}^{\mathbf{P}}$ of $\mathcal{F}$ and redefine $\mathcal{F}_t$ as $\mathcal{F}^{\mathbf{P}}_{t+} := \cap_{s>t}\mathcal{F}^{\mathbf{P}}_s$, where $\mathcal{F}^{\mathbf{P}}_s$ is the σ-algebra generated by $\mathcal{F}_s$ and

the sets $E \in \mathcal{F}^{\mathbf{P}}$ satisfying $\mathbf{P}(E) = 0$; $(X_t, \mathcal{F}_t)$ will remain a (super)martingale by [13], VI.3(1).

We are particularly interested in *test supermartingales*, defined as supermartingales that are nonnegative ($X_t \geq 0$ for all t) and satisfy $\mathbf{E}(X_0) \leq 1$, and *test martingales*, defined as martingales that are nonnegative and satisfy $\mathbf{E}(X_0) = 1$. Earlier, we defined test martingales as those having initial value 1; this can be reconciled with the new definition by setting $X_t := 1$ for $t < 0$. 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 $c \geq 1$ ([39], p. 100; [13], VI.1). We will call this the *maximal inequality*. This inequality shows that X_t can take the value ∞ only with probability zero.

3.2 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 $\mathbf{P}(E) = 0$).

3.3 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 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 time 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. We include this result for mathematical completeness rather than because of its practical importance; the construction involves arbitrarily introducing a filtration, which need not correspond to any time scale with practical meaning. In its statement, we use $\mathcal{F}_\infty$ to denote the σ-algebra generated by $\cup_t \mathcal{F}_t$.

Theorem 1. *1. If $(X_t, \mathcal{F}_t)$ is a test supermartingale, then $X_\infty := \lim_{t\to\infty} X_t$ 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 $B = 1/X_\infty$ a.s. 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 ([13], VI.6), and the inequality $\int X_\infty \mathrm{d}\mathbf{P} \leq 1$ holds by Fatou's lemma:

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

Now suppose that B is a precise Bayes factor and $(\mathcal{F}_t)$ is a filtration (not necessarily satisfying the usual conditions) such that B is $\mathcal{F}_\infty$-measurable; for concreteness, we consider the case of continuous time. Define a test martingale $(X_t, \mathcal{F}^{\mathbf{P}}_{t+})$ by setting $X_t := \mathbf{E}(1/B \mid \mathcal{F}^{\mathbf{P}}_{t+})$; versions of conditional expectations

can be chosen in such a way that (X_t) is right-continuous: cf. [13], VI.4. Then $X_\infty = 1/B$ almost surely by Lévy's zero-one law ([24], pp. 128–130; [30], VI.6, corollary). It remains to notice that $(X_t, \mathcal{F}_t)$ will also be a test martingale. 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 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. The *stopped process* $(X^\tau_t, \mathcal{F}_t) := (X_{t \wedge \tau}, \mathcal{F}_t)$, where $a \wedge b := \min(a, b)$, will also be a test supermartingale ([13], VI.12). Since X_τ is the final value of the stopped process, it follows from part 1 of Theorem 1 that $1/X_\tau$ is a Bayes factor. (This also follows directly from Doob's stopping theorem, [30], VI.13.)

5 Supermartingales and p-values

Now we will prove 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 any stopping time τ, but the strongest way of making the point is to consider the supremum over all time points (i.e., for $\tau := \infty$).

We will also show how to construct a test martingale that has the inverse of a given p-test as its supremum. Because the time scale for this martingale is artificial, the value of the construction is more mathematical than directly practical; it will help us prove Theorem 4 in Section 7. But it may be worthwhile to give an intuitive explanation of the construction. This is easiest when the p-test has discrete levels, because then we merely construct a sequence of bets. Consider a p-test p that is equal to 1 with probability 1/2, to 1/2 with probability 1/4, to 1/4 with probability 1/8, etc.:

$$\mathbf{P}\{p = 2^{-n}\} = 2^{-n-1}$$

for $n = 0, 1, \ldots$. To see that a function on Ω that takes these values with these probabilities is a p-test, notice that when $2^{-n} \le \delta < 2^{-n+1}$,

$$\mathbf{P}\{p \le \delta\} = \mathbf{P}\{p \le 2^{-n}\} = 2^{-n} \le \delta.$$

Suppose that we learn first whether p is 1. Then, if it is not 1, we learn whether it is 1/2. Then, if it is not 1/2, whether it is 1/4, etc. To create the test martingale $X_0, X_1, \ldots$, we start with capital $X_0 = 1$ and bet it all against p

being 1. If we lose, $X_1 = 0$ and we stop. If we win, $X_1 = 2$, and we bet it all against p being $1/2$, etc. Each time we have even chances of doubling our money or losing it all. If $p = 2^{-n}$, then our last bet will be against $p = 2^{-n}$, and the amount we will lose, 2^n, will be X_∞^*. So $1/X_\infty^* = p$, as desired.

Here is our formal result:

Theorem 2. *1. If $(X_t, \mathcal{F}_t)$ is a test supermartingale, $1/X_\infty^*$ 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 $\mathbf{P}\{1/X_\infty^* \le \delta\} \le \delta$ for test supermartingales follows from the maximal inequality (2).

In the opposite direction, let p be a precise p-test. Set $\Pi := 1/p$; this function takes values 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 $t \ge 1$; notice that $X_1 = 1$ with probability one.

Let $t \ge 1$; 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, $s \le t$, 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 $1 \le s < t$, 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 [41] and, less explicitly, in [35]. 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 [41].

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 [41] (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 $\mathbf{E}(\xi) = \int_0^\infty \mathbf{P}\{\xi > c\}\mathrm{d}c$, where ξ is a nonnegative random variable:

$$\mathbf{E}(1/f(p_1)) = \int_0^\infty \mathbf{P}\{1/f(p_1) > c\}\mathrm{d}c \ge \int_0^\infty \mathbf{P}\{1/f(p_2) > c\}\mathrm{d}c = \mathbf{E}(1/f(p_2)).$$

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 [41] and [35].

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 an admissible calibrator for any $\alpha > 0$.

7 Calibrating the running 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 almost surely.

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., $\mathbf{E}(f(X^*_\infty)) \le 1$. Similarly to the proof of Theorem 1, we set $Y_t := \mathbf{E}(f(X^*_\infty) \mid \mathcal{F}_t)$ obtaining a nonnegative martingale $(Y_t, \mathcal{F}_t)$ satisfying $Y_\infty = f(X^*_\infty)$ a.s. We have $\mathbf{E}(Y_0) \le 1$; the case $\mathbf{E}(Y_0) = 0$ is trivial, and so we assume $\mathbf{E}(Y_0) > 0$. Since

$$Y_t = \mathbf{E}(f(X^*_\infty) \mid \mathcal{F}_t) \ge \mathbf{E}(f(X^*_t) \mid \mathcal{F}_t) = f(X^*_t) \quad \text{a.s.}$$

(the case $t = \infty$ was considered separately) and we can make $(Y_t, \mathcal{F}_t)$ a test martingale by dividing each Y_t by $\mathbf{E}(Y_0) \in (0,1]$, 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 (super)martingale whenever $(Y_t, \mathcal{F}_t)$ is a test (super)martingale.

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

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

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

$$1 \ge \int f(X^*_{\pi/2})\mathrm{d}\mathbf{P} = \int f(1/p)\mathrm{d}\mathbf{P} = \int_0^1 f(1/x)\mathrm{d}x.$$

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

$$f(y) := \alpha y^{1-\alpha}, \quad \alpha \in (0,1), \tag{12}$$

analogous to (8) and the class

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

analogous to (9).

In the case of discrete time, Theorem 4 has been greatly generalized by Dawid et al. ([12], Theorem 1). The generalization, which required new proof

techniques, makes it possible to apply the result in new fields, such as mathematical finance ([12], Section 4).

In this article, we have considered only tests of simple statistical hypotheses. We can use similar ideas for testing composite hypotheses, i.e., sets of probability measures. One possibility is to measure the evidence against the composite hypothesis by the current value of a random process that is a test supermartingale under all probability measures in the composite hypothesis; we will call such processes *simultaneous test supermartingales*. For example, there are non-trivial processes that are test supermartingales under all exchangeable probability measures simultaneously ([42], Section 7.1). Will martingale calibrators achieve their goal for simultaneous test supermartingales? The method of proof of Theorem 4 does not work in this situation: in general, it will produce a different test supermartingale for each probability measure. The advantage of the method used in [12] is that it will produce one process, thus demonstrating that for each martingale calibrator f and each simultaneous test supermartingale X_t there exists a simultaneous test supermartingale Y_t such that $Y_t \geq f(X_t^*)$ for all t (the method of [12] works pathwise and makes the qualification "almost surely" superfluous).

More flexible method: a separate test supermartingale for each probability measure in the composite hypothesis. The method of proof Theorem 4 now works.

8 Examples

Although our results are very general, we can illustrate them using the simple problem of testing whether a coin is fair. Formally, suppose we observe a sequence of independent identically distributed binary random variables $x_1, x_2, \ldots$, each taking values in the set $\{0, 1\}$; the probability $\theta \in [0, 1]$ of $x_1 = 1$ is unknown. Let P_θ be the probability distribution of $x_1, x_2, \ldots$; it is a probability measure on $\{0, 1\}^\infty$. In most of this section, our null hypothesis is that $\theta = 1/2$.

We consider both Bayesian testing of $\theta = 1/2$, where the output is a posterior distribution, and non-Bayesian testing, where the output is a p-value. We call the approach that produces p-values the *sampling-theory approach* rather than the frequentist approach, because it does not require us to interpret all probabilities as frequencies; instead, we can merely interpret the p-values using Cournot's principle ([36], Section 2). We have borrowed the term "sampling-theory" from D. R. Cox and A. P. Dempster [10, 14], without necessarily using it in exactly the same way as either of them do.

We consider two tests of $\theta = 1/2$, corresponding to two different alternative hypotheses.

1. First we test $\theta = 1/2$ against $\theta = 3/4$. This is unrealistic on its face; it is hard to imagine accepting a model that contains only these two simple hypotheses. But some of what we learn from this test will carry over to sensible and widely used tests of a simple against a composite hypothesis.

2. Second, we test $\theta = 1/2$ against the composite hypothesis $\theta \neq 1/2$. In the spirit of Bayesian statistics and following Laplace ([22]; see also [38], Section 870, and [37]), we represent this composite hypothesis by the uniform distribution on $[0, 1]$, the range of possible values for θ. (In general, the composite hypotheses of this section will be composite only in the sense of Bayesian statistics; from the point of view of the sampling-theory approach, these are still simple hypotheses.)

For each test, we give an example of calibration of the running supremum of the likelihood ratio. In the case of the composite alternative hypothesis, we also discuss the implications of using the inverse of the running supremum of the likelihood ratio as a p-value.

To round out the picture, we also discuss Bayesian testing of the composite hypothesis $\theta \leq 1/2$ against the composite hypothesis $\theta > 1/2$, representing the former by the uniform distribution on $[0, 1/2]$ and the latter by the uniform distribution on $(1/2, 1]$. Then, to conclude, we discuss the relevance of the calibration of running suprema to Bayesian philosophy.

Because the idea of tracking the supremum of a martingale is related to the idea of waiting until it reaches a high value, our discussion is related to a long-standing debate about "sampling to reach a foregone conclusion", i.e., continuing to sample in search of evidence against a hypothesis and stopping only when some conventional p-value finally dips below a conventional level such as 5%. This debate goes back at least to the work of Francis Anscombe in 1954 [4]. In 1961, Peter Armitage described situations where even a Bayesian can sample to a foregone conclusion ([6]; [7], Section 5.1.4). Yet in 1963 [15], Ward Edwards and his co-authors insisted that this is not a problem: "The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience." For further information on this debate, see [43]. We will not attempt to analyze it thoroughly, but our examples may be considered a contribution to it.

8.1 Testing $\theta = 1/2$ against a simple alternative

To test our null hypothesis $\theta = 1/2$ against the alternative hypothesis $\theta = 3/4$, we use the likelihood ratio

$$X_t := \frac{P_{3/4}(x_1, \dots, x_t)}{P_{1/2}(x_1, \dots, x_t)} = \frac{(3/4)^{k_t}(1/4)^{t-k_t}}{(1/2)^t} = \frac{3^{k_t}}{2^t}, \tag{13}$$

where k_t is the number of 1s in $x_1, \dots, x_t$ (and $P_\theta(x_1, \dots, x_t)$ is the probability under P_θ that the first t observations are $x_1, \dots, x_t$; such informal notation was already used in Section 2). The sequence of successive values of this likelihood ratio is a test martingale (X_t).

According to (12), the function

$$f(y) := 0.1y^{0.9} \tag{14}$$

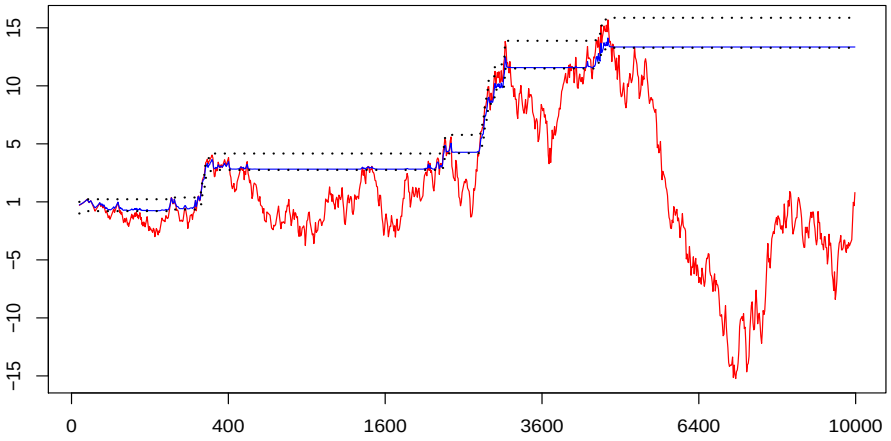


Figure 2: The red line is a realization over 10,000 trials of the likelihood ratio for testing $\theta = 1/2$ against $\theta = 3/4$. The horizontal axis gives the number of observations so far. The vertical axis is logarithmic and is labelled by powers of 10. The likelihood ratio varies wildly, up to 10^{15} and down to 10^{-15}. Were the sequence continued indefinitely, it would be unbounded in both directions.

is a martingale calibrator. So there exists a test martingale (Y_t) such that

$$Y_t \geq \max_{n=1,\dots,t} 0.1 X_n^{0.9}. \tag{15}$$

Figure 2 shows an example in which the martingale calibrator (14) preserves a reasonable amount of the evidence against $\theta = 1/2$. To construct this figure, we generated a sequence $x_1, \dots, x_{10,000}$ of 0s and 1s, choosing each x_t independently with the probability θ for $x_t = 1$ always equal to $\ln 2 / \ln 3 \approx 0.63$. Then we formed the lines in the figure as follows:

- The red line is traced by the sequence of numbers $X_t = 3^{k_t}/2^t$. If our null hypothesis $\theta = 1/2$ were true, these numbers would be a realization of a test martingale, but this hypothesis is false (as is our alternative hypothesis $\theta = 3/4$).
- The upper dotted line is the running supremum of the X_t:

$$X_t^* = \max_{n=1,\dots,t} \frac{3^{k_n}}{2^n} = (\text{best evidence so far against } \theta = 1/2)_t.$$

- The lower dotted line, which we will call F_t, shrinks this best evidence using our martingale calibrator: $F_t = 0.1(X_t^*)^{0.9}$.
- The blue line, which we will call Y_t, is a test martingale under the null hypothesis that satisfies (15): $Y_t \geq F_t$.

According to the proof of Theorem 4, $\mathbf{E}(0.1(X^*_\infty)^{0.9} \mid \mathcal{F}_t)/\mathbf{E}(0.1(X^*_\infty)^{0.9})$, where the expected values are with respect to $P_{1/2}$, is a test martingale that satisfies (15). Because these expected values may be difficult to compute, we have used in its stead in the role of Y_t a more easily computed test martingale that is shown in [12] to satisfy (15).

Here are the final values of the processes shown in Figure 2:

$$X_{10,000} = 2.2 \qquad X^*_{10,000} = 7.3 \times 10^{15}$$
$$F_{10,000} = 1.9 \times 10^{13} \qquad Y_{10,000} = 2.2 \times 10^{13}.$$

The test martingale Y_t legitimately and correctly rejects the null hypothesis at time $10,000$ on the basis of X_t's high earlier values, even though the Bayes factor $X_{10,000}$ is not high. The Bayes factor $Y_{10,000}$ gives overwhelming evidence against the null hypothesis, even though it is more than two orders of magnitude smaller than $X^*_{10,000}$.

As the reader will have noticed, the test martingale X_t's overwhelming values against $\theta = 1/2$ in Figure 2 are followed, around $t = 7,000$, by overwhelming values (order of magnitude 10^{-15}) against $\theta = 3/4$. Had we been testing $\theta = 3/4$ against $\theta = 1/2$, we would have found that it can also be rejected very strongly even after calibration. The fact that (X_t) and $(1/X_t)$ both have times when they are very large is not accidental when we sample from $P_{\ln 2/\ln 3}$. Under this measure, the conditional expected value of the increment $\ln X_t - \ln X_{t-1}$, given the first $t-1$ observations, is

$$\frac{\ln 2}{\ln 3} \ln \frac{3}{2} + \left(1 - \frac{\ln 2}{\ln 3}\right) \ln \frac{1}{2} = 0.$$

So $\ln X_t$ is a martingale under $P_{\ln 2/\ln 3}$. The conditional variance of its increment is

$$\frac{\ln 2}{\ln 3} \left(\ln \frac{3}{2}\right)^2 + \left(1 - \frac{\ln 2}{\ln 3}\right) \left(\ln \frac{1}{2}\right)^2 = \ln 2 \ln \frac{3}{2}.$$

By the law of the iterated logarithm,

$$\limsup_{t\to\infty} \frac{\ln X_t}{\sqrt{2 \ln 2 \ln \frac{3}{2}\, t \ln \ln t}} = 1 \text{ and } \liminf_{t\to\infty} \frac{\ln X_t}{\sqrt{2 \ln 2 \ln \frac{3}{2}\, t \ln \ln t}} = -1$$

almost surely. This means that as t tends to ∞, $\ln X_t$ oscillates between approximately $\pm 0.75\sqrt{t \ln \ln t}$; in particular,

$$\limsup_{t\to\infty} X_t = \infty \text{ and } \liminf_{t\to\infty} X_t = 0 \tag{16}$$

almost surely. This guarantees that we will eventually obtain overwhelming evidence against whichever of the hypotheses $\theta = 1/2$ and $\theta = 3/4$ that we want to reject. This may be called sampling to a foregone conclusion, but the foregone conclusion will be correct, since both $\theta = 1/2$ and $\theta = 3/4$ are wrong.

In order to obtain (16), we chose $x_1, \ldots, x_{10,000}$ from a probability distribution, $P_{\ln 2/\ln 3}$, that lies midway between $P_{1/2}$ and $P_{3/4}$ in the sense that it tends

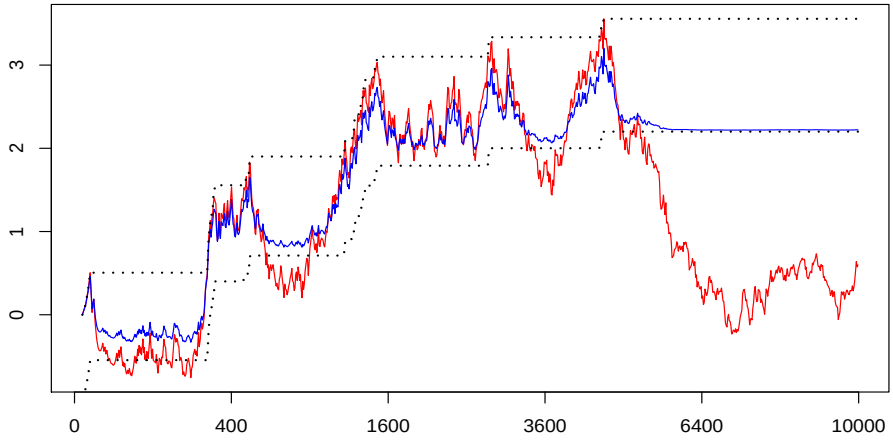


Figure 3: A realization over $10,000$ trials of the likelihood ratio for testing $\theta = 1/2$ against the probability distribution Q obtained by averaging P_θ with respect to the uniform distribution for θ. The vertical axis is again logarithmic. As in Figure 2, the oscillations would be unbounded if trials continued indefinitely.

to produce sequences that are as atypical with respect to the one measure as to the other. Had we chosen a sequence $x_1, \ldots, x_{10,000}$ less atypical with respect to $P_{3/4}$ than with respect to $P_{1/2}$, then we might have been able to sample to the foregone conclusion of rejecting $\theta = 1/2$, but not to the foregone conclusion of rejecting $\theta = 3/4$.

8.2 Testing $\theta = 1/2$ against a composite alternative

Retaining $\theta = 1/2$ as our null hypothesis, we now take as our alternative hypothesis the probability distribution Q obtained by averaging P_θ with respect to the uniform distribution for θ.

After we observe $x_1, \ldots, x_t$, the likelihood ratio for testing $P_{1/2}$ against Q is

$$X_t := \frac{Q(x_1, \ldots, x_t)}{P_{1/2}(x_1, \ldots, x_t)} = \frac{\int_0^1 \theta^{k_t}(1-\theta)^{t-k_t}\mathrm{d}\theta}{(1/2)^t} = \frac{k_t!(t-k_t)!2^t}{(t+1)!}. \qquad (17)$$

Figure 3 shows an example of this process and of the application of same martingale calibrator, (14), that we used in Figure 2. In this case, we generate the 0s and 1s in the sequence $x_1, \ldots, x_{10,000}$ independently but with a probability for $x_t = 1$ that slowly converges to $1/2$: $\frac{1}{2} + \frac{1}{4}\sqrt{\ln t/t}$. As we show in Appendix B, (16) again holds almost surely; if you wait long enough, you will have enough evidence to reject legitimately whichever of the two false hypotheses (independently and identically distributed with $\theta = 1/2$, or independently and identically distributed with $\theta \neq 1/2$) you want.

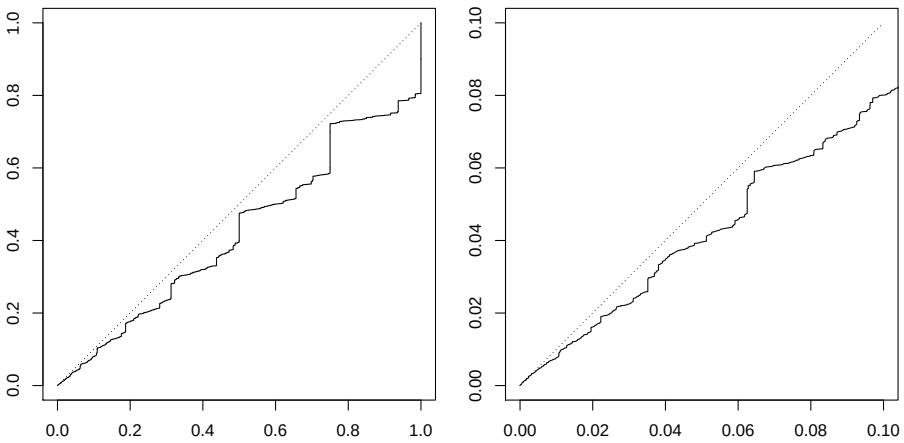


Figure 4: On the left we graph $\mathbf{P}\{p \leq \delta\}$ as a function of δ, where p is the function defined in (18). On the right, we magnify the lower left corner of this graph.

Here are the final values of the processes shown in Figure 3:

$$X_{10,000} = 3.5 \qquad X^*_{10,000} = 3,599$$
$$F_{10,000} = 159 \qquad Y_{10,000} = 166.$$

In this case, the evidence against $\theta = 1/2$ is very substantial but not overwhelming.

8.3 p-values for testing $\theta = 1/2$

By Theorem 2, $1/X^*_\infty$ is a p-test whenever (X_t) is a test martingale. Applying this to the test martingale (17) for testing $P_{1/2}$ against Q, we see that

$$p(x_1, x_2, \ldots) := \frac{1}{\sup_{1 \leq t < \infty} \frac{k_t!(t-k_t)!2^t}{(t+1)!}} = \inf_{1 \leq t < \infty} \frac{(t+1)!}{k_t!(t-k_t)!2^t} \tag{18}$$

is a p-test for testing $\theta = 1/2$ against $\theta \neq 1/2$. Figure 4 shows that it is only moderately conservative.

Any function of the observations that is bounded below by a p-test is also a p-test. So for any rule N for selecting a positive integer $N(x_1, x_2, \ldots)$ based on knowledge of some or all of the observations $x_1, x_2, \ldots$, the function

$$r_N(x_1, x_2, \ldots) := \frac{(N+1)!}{k_N!(N-k_N)!2^N} \tag{19}$$

is a p-test. It does not matter whether N qualifies as a stopping rule (i.e., whether $x_1, \ldots, x_n$ always determine whether $N(x_1, x_2, \ldots) \leq n$).

For each positive integer n, let

$$p_n := \frac{(n+1)!}{k_n!(n-k_n)!2^n}. \tag{20}$$

We can paraphrase the preceding paragraph by saying that p_n is a p-value (i.e., the value of a p-test) no matter what rule is used to select n. In particular, it is a p-value even if it was selected because it was the smallest number in the sequence $p_1, p_2, \ldots, p_n, \ldots, p_t$, where t is an integer much larger than n.

We must nevertheless be cautious if we do not know the rule N—if the experimenter who does the sampling reports to us p_n and perhaps some other information but not the rule N. We can consider the reported value of p_n a legitimate p-value whenever we know that the experimenter would have told us p_n for some n, even if we do not know what rule N he followed to choose n and even if he did not follow any clear rule. But we should not think of p_n as a p-value if it is possible that the experimenter would not have reported anything at all had he not found an n with a p_n to his liking. We are performing a p-test only if we learn the result no matter what it is.

Continuing to sample in search of evidence against $\theta = 1/2$ and stopping only when the p-value finally reaches 5% can be considered legitimate if instead of using conventional p-tests for fixed sample sizes we use the p-test (19) with N defined by

$$N(x_1, x_2, \ldots) := \inf\left\{ n \,\middle|\, \frac{(n+1)!}{k_n!(n-k_n)!2^n} \le 0.05 \right\}.$$

But we must bear in mind that $N(x_1, x_2, \ldots)$ may take the value ∞. If the experimenter stops only when the p-value dips down to the 5% level, he has a chance of at least 95%, under the null hypothesis, of never stopping. So it will be legitimate to interpret a reported p_n of 0.05 or less as a p-value (the observed value of a p-test) only if we were somehow also guaranteed to hear about the failure to stop.

8.4 Comparison with a standard p-test

If the number n of observations is known in advance, a standard sampling-theory procedure for testing the hypothesis $\theta = 1/2$ is to reject it if $|k_n - n/2| \ge c_{n,\delta}$, where $c_{n,\delta}$ is chosen so that $P_{1/2}\{|k_n - n/2| \ge c_{n,\delta}\}$ is equal (or less than but as close as possible) to a chosen significance level δ. To see how this compares with the p-value p_n given by (20) let us compare the conditions for non-rejection.

- If we use the standard procedure, the condition for not rejecting $\theta = 1/2$ at level δ is
$$|k_n - n/2| < c_{n,\delta}. \tag{21}$$
- If we use the p-value p_n, the condition for not rejecting $\theta = 1/2$ at level δ is $p_n > \delta$, or
$$\frac{(n+1)!}{k_n!(n-k_n)!2^n} > \delta. \tag{22}$$

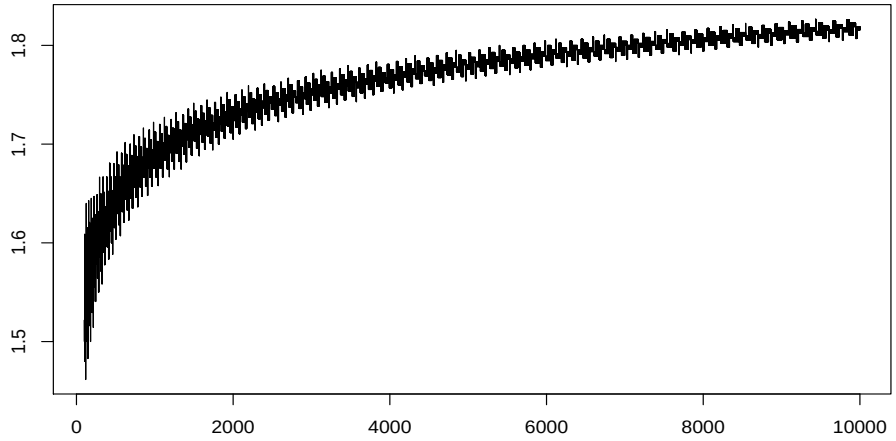


Figure 5: The ratio (23) as n ranges from 100 to 10,000. This is the factor by which not knowing n in advance widens the 99% prediction interval for k_n. Asymptotically, the ratio tends to infinity with n as $c\sqrt{\ln n}$ for some positive constant c.

In both cases, k_n satisfies the condition with probability at least $1 - \delta$ under the null hypothesis, and hence the condition defines a level $1 - \delta$ prediction interval for k_n. Because condition (21) requires the value of n to be known in advance and condition (22) does not, we can expect the prediction interval defined by (22) to be wider than the one determined by (22). How much wider?

Figure 5 answers this question for the case where $\delta = 0.01$ and $100 \le n \le 10,000$. It shows, for each value of n in this range, the ratio

$$\frac{\text{width of the 99\% prediction interval given by (22)}}{\text{width of the 99\% prediction interval given by (21)}}, \tag{23}$$

i.e., the factor by which not knowing n in advance widens the prediction interval. The factor is less than 2 over the whole range but increases steadily with n.

As n increases further, the factor by which the standard interval is multiplied increases without limit, but very slowly. To verify this, we first rewrite (22) as

$$|k_n - n/2| < (1 + \alpha_n)\sqrt{n}\sqrt{\frac{1}{2}\ln\frac{1}{\delta} + \frac{1}{4}\ln n}, \tag{24}$$

where α_n is a sequence such that $\alpha_n \to 0$ as $n \to \infty$. (For some α_n of order $o(1)$ the inequality (24) is stronger than $p_n > \delta$, whereas for others it is weaker; see Appendix B for details of calculations.) Then, using the Berry-Esseen theorem and letting z_ϵ stand for the upper ϵ-quantile of the standard Gaussian distribution, we rewrite (21) as

$$|k_n - n/2| < \frac{1}{2} z_{\delta/2+\alpha_n}\sqrt{n}, \tag{25}$$

where α_n is a sequence such that $|\alpha_n| \le (2\pi)^{-1/2} n^{-1/2}$ for all n. (See [17].) As $\delta \to 0$,

$$z_{\delta/2} \sim \sqrt{2 \ln \frac{2}{\delta}} \sim \sqrt{2 \ln \frac{1}{\delta}}.$$

So the main asymptotic difference between (24) and (25) is the presence of the term $\frac{1}{4} \ln n$ in (24).

The ratio (23) tends to infinity with n as $c\sqrt{\ln n}$ for a positive constant c (namely, for $c = 1/z_{\delta/2}$, where $\delta = 0.01$ is the chosen significance level). However, the expression on the right-hand side of (24) results from using the uniform probability measure on θ to average the probability measures P_θ. Averaging with respect to a different probability measure would give something different, but it is clear from the law of the iterated logarithm that the best we can get is a prediction interval whose ratio with the standard interval will grow like $\sqrt{\ln \ln n}$ instead of $\sqrt{\ln n}$. In fact, the method we just used to obtain (24) was used by Ville, with a more carefully chosen probability measure on θ, to prove the upper half of the law of the iterated logarithm ([39], Section V.3), and Ville's argument was rediscovered and simplified using the algorithmic theory of randomness in [40], Theorem 1.

8.5 Testing a composite hypothesis against a composite hypothesis

When Peter Armitage pointed out that even Bayesians can sample to a foregone conclusion, he used as example the Gaussian model with known variance and unknown mean [6]. We can adapt Armitage's idea to coin tossing by comparing two composite hypotheses: the null hypothesis $\theta \le 1/2$, represented by the uniform probability measure on $[0, 1/2]$, and the alternative hypothesis $\theta > 1/2$, represented by the uniform probability measure on $(1/2, 1]$. (These hypotheses are natural in the context of paired comparison: see, e.g., [23], Section 3.1.) The test martingale is

$$X_t = \frac{2 \int_{1/2}^1 \theta^{k_t} (1-\theta)^{t-k_t} \mathrm{d}\theta}{2 \int_0^{1/2} \theta^{k_t} (1-\theta)^{t-k_t} \mathrm{d}\theta} = \frac{\mathbf{P}\{B_{t+1} \le k_t\}}{\mathbf{P}\{B_{t+1} \ge k_t + 1\}}, \tag{26}$$

where B_n is the binomial random variable with parameters n and $1/2$; see Appendix B for details. If the sequence $x_1, x_2, \ldots$ turns out to be typical of $\theta = 1/2$, then by the law of the iterated logarithm, $(k_t - t/2)/\sqrt{t}$ will almost surely have ∞ as its upper limit and $-\infty$ as its lower limit; therefore, (16) will hold again. This confirms Armitage's intuition that arbitrarily strong evidence on both sides will emerge if we wait long enough, but the oscillation depends on increasingly extreme reversals of a random walk, and the lifetime of the universe may not be long enough for us to see any of them ($\sqrt{\ln \ln(5 \times 10^{23})} < 2$).

Figure 6 depicts one example, for which the final values are

$$X_{10,000} = 3.7 \qquad\qquad X^*_{10,000} = 272$$

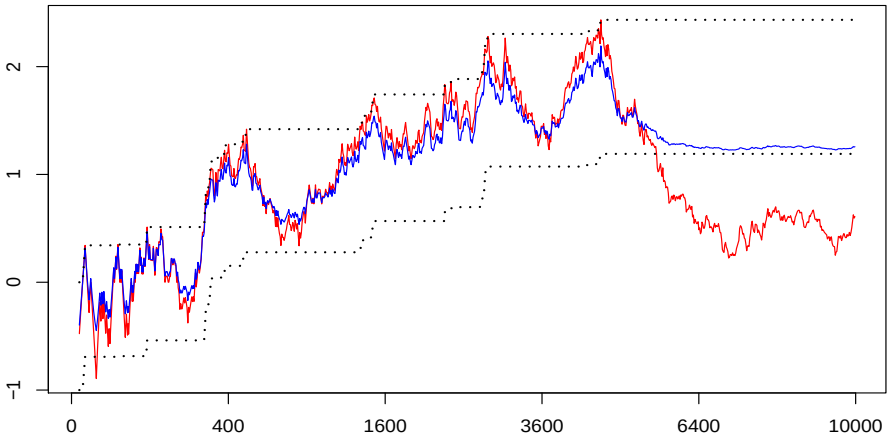


Figure 6: A realization over $10,000$ trials of the likelihood ratio for testing the probability distribution obtained by averaging P_θ with respect to the uniform probability measure on $[0, 1/2]$ against the probability distribution obtained by averaging P_θ with respect to the uniform probability measure on $(1/2, 1]$. As in the previous figures, the vertical axis is logarithmic, and the red line would be unbounded in both directions if observations continued indefinitely.

$$F_{10,000} = 15.5 \qquad\qquad Y_{10,000} = 17.9.$$

In this realization, the first $10,000$ observations provide modest evidence against $\theta \leq 1/2$ and none against $\theta > 1/2$. Figures 2 and 3 are reasonably typical for their setups, but in this setup it is unusual for the first $10,000$ observations to show even as much evidence against one of the hypotheses as we see in Figure 6.

8.6 A puzzle for Bayesians

From a Bayesian point of view, it may seem puzzling that we should want to shrink a likelihood ratio in order to avoid exaggerating the evidence against a null hypothesis. Observations affect Bayesian posterior odds only through the likelihood ratio, and we know that the likelihood ratio is not affected by the sampling plan. So why should we adjust it to take the sampling plan into account?

Suppose we assign equal prior probabilities of $1/2$ each to the two hypotheses $\theta = 1/2$ and $\theta = 3/4$ in our first coin-tossing example. Then if we stop at time t, the likelihood ratio X_t given by (13) is identical with the posterior odds in favor of $\theta = 3/4$. If we write $\mathbf{post}_t$ for the posterior probability measure at time t, then

$$X_t = \frac{\mathbf{post}_t\{\theta = 3/4\}}{\mathbf{post}_t\{\theta = 1/2\}} = \frac{1 - \mathbf{post}_t\{\theta = 1/2\}}{\mathbf{post}_t\{\theta = 1/2\}},$$

and

$$\mathbf{post}_t\{\theta = 1/2\} = \frac{1}{X_t + 1}. \tag{27}$$

This is our posterior probability given the evidence $x_1, \dots, x_t$ no matter why we decided to stop at time t. If we "calibrate" X_t and plug the calibrated value instead of the actual value into (27), we will get the posterior probability wrong.

It may help us escape from our puzzlement to acknowledge that if the model is wrong, then the observations may oscillate between providing overwhelming evidence against $\theta = 1/2$ and providing overwhelming evidence against $\theta = 3/4$, as in Figure 2. Only if we insist on retaining the model in spite of this very anomalous phenomenon will (27) continue to be our posterior probability for $\theta = 1/2$ at time t, and it is this stubbornness that opens the door to sampling to whichever foregone conclusion we want, $\theta = 1/2$ or $\theta = 3/4$.

The same issues arise when we test $\theta = 1/2$ against the composite hypothesis $\theta \neq 1/2$. A natural Bayesian method for doing this is to put half our probability on $\theta = 1/2$ and distribute the other half uniformly on $[0, 1]$ (which is a special case of a widely recommended procedure described in, e.g., [7], p. 391). This makes the likelihood ratio X_t given by (17) the posterior odds against $\theta = 1/2$. As we have seen, if the observations $x_1, x_2, \dots$ turn out to be typical for the distribution in which they are independent with the probability for $x_t = 1$ equal to $\frac{1}{2} + \frac{1}{4}\sqrt{\ln t/t}$, then if you wait long enough, you can observe values of X_t as small or as large as you like, and thus obtain a posterior probability for $\theta = 1/2$ as large or as small as you like.

Of course, it will not always happen that the actual observations are so equidistant from a simple null hypothesis and the probability distribution representing its negation that the likelihood ratio will oscillate wildly and you can sample to whichever side you want. More often, the likelihood ratio and hence the posterior probability will settle on one side or the other. But in the spirit of George Box's maxim that all models are wrong, we can interpret this not as confirmation of the side favored but only as confirmation that the other side should be rejected. The rejection will be legitimate from the Bayesian point of view, regardless of why we stopped sampling. It will also be legitimate from the sampling-theory point of view.

On this argument, it is legitimate to collect data until a point has been disproven but not legitimate to interpret this data as proof of an alternative hypothesis within the model. Only when we really know the model is correct can we prove one of its hypotheses by rejecting the others.

A Inadequacy of test martingales in continuous time

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 (Proposition 2). Fix a filtration $(\mathcal{F}_t)$ satisfying

the usual conditions; 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, [30], 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 (Proposition 2). For hypothesis testing we need "local martingales", a modification of the notion of martingales introduced by Itô and Watanabe [18] 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.

We will say 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 popular alternative definition requires that each $X_{t \wedge \tau_n} \mathbb{I}_{\{\tau_n > 0\}}$ should belong to $\mathcal{C}$.) A standard argument (see, e.g., [13], VI.29) shows that there is no difference between test supermartingales and local test supermartingales:

Proposition 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 $0 \le s < t$:

$$\begin{aligned} \mathbf{E}(X_t \mid \mathcal{F}_s) = \mathbf{E}\left(\lim_{n\to\infty} X_t^{\tau_n} \mid \mathcal{F}_s\right) &\le \liminf_{n\to\infty} \mathbf{E}\left(X_t^{\tau_n} \mid \mathcal{F}_s\right) \\ &\le \liminf_{n\to\infty} X_s^{\tau_n} = X_s \quad \text{a.s.} \end{aligned}$$

In particular, $\mathbf{E}(X_t) \le 1$. □

An adapted process (A_t) is called *increasing* if $A_0 = 0$ a.s. and its every path is right-continuous and increasing (as usual, not necessarily strictly increasing). According to the Doob-Meyer decomposition theorem ([13], 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 ([13], (VI.31.1)). In continuous time, however, the difference is essential. Suppose the filtration $(\mathcal{F}_t)$ admits a standard Brownian motion $(W_t, \mathcal{F}_t)$ in $\mathbb{R}^3$. A well-known example ([19]; see also [30], VI.21, and [13],

VI.26) of a local martingale which is not a martingale is $L_t := 1/\|W_t + e\|$, where 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., [29] (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 ([30], 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) belongs 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) ([30], 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 proposition.

Proposition 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 $Y_t \ge \delta X_t$ a.s. for all t.*

Proof. Let $X_t := L_t = 1/\|W_t + e\|$, and suppose there is a test martingale (Y_t) such that $Y_t \ge \delta X_t$ a.s. for all t. Let $\epsilon > 0$ be arbitrarily small. Since (Y_t) is in (DL) ([30], 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. □

B Details of calculations

In this appendix we will give details of some calculations omitted in Section 8. They will be based on Stirling's formula $n! = \sqrt{2\pi n}(n/e)^n e^{\lambda_n}$, where $\lambda_n = o(1)$ as $n \to \infty$.

B.1 Oscillating evidence when testing against a composite alternative

First we establish (16) for X_t defined by (17). Suppose we have made t observations and observed $k := k_t$ 1s so far. We start from finding bounds on k that

are implied by the law of the iterated logarithm. Using the simplest version of Euler's summation formula (as in [5], Theorem 1), we can find its expected value as

$$\begin{aligned}\mathbf{E}(k) &= \sum_{n=1}^{t}\left(\frac{1}{2}+\frac{1}{4}\sqrt{\frac{\ln n}{n}}\right) = \frac{t}{2}+\frac{1}{4}\sum_{n=2}^{t}\left(\frac{\ln n+1}{\sqrt{n\ln n}}\right)-\frac{1}{4}\sum_{n=2}^{t}\left(\frac{1}{\sqrt{n\ln n}}\right)\\ &= \frac{t}{2}+\frac{1}{4}\int_2^t\left(\frac{\ln u+1}{\sqrt{u\ln u}}\right)\mathrm{d}u+O(\sqrt{t}) = \frac{t}{2}+\frac{1}{2}\sqrt{t\ln t}+O(\sqrt{t}).\end{aligned}$$

Its variance is

$$\mathbf{var}(k) = \sum_{n=1}^{t}\left(\frac{1}{2}+\frac{1}{4}\sqrt{\frac{\ln n}{n}}\right)\left(\frac{1}{2}-\frac{1}{4}\sqrt{\frac{\ln n}{n}}\right) = \sum_{n=1}^{t}\left(\frac{1}{4}-\frac{1}{16}\frac{\ln n}{n}\right)\sim\frac{t}{4}.$$

Therefore, Kolmogorov's law of the iterated logarithm gives

$$\limsup_{t\to\infty}\frac{k-\frac{1}{2}\left(t+\sqrt{t\ln t}\right)}{\sqrt{\frac{1}{2}t\ln\ln t}} = 1 \text{ and } \liminf_{t\to\infty}\frac{k-\frac{1}{2}\left(t+\sqrt{t\ln t}\right)}{\sqrt{\frac{1}{2}t\ln\ln t}} = -1 \quad \text{a.s.} \tag{28}$$

Using the definition (17) and applying Stirling's formula, we obtain

$$\begin{aligned}\ln X_t &= t\ln 2+\ln\frac{k!(t-k)!}{t!}-\ln(t+1) \qquad (29)\\ &= t\ln 2-tH(k/t)+\ln\sqrt{2\pi\frac{k(t-k)}{t}}+\lambda_k+\lambda_{t-k}-\lambda_t-\ln(t+1)\\ &= t\left(\ln 2-H(k/t)\right)-\frac{1}{2}\ln t+O(1) = 2t\left(\frac{k}{t}-\frac{1}{2}\right)^2-\frac{1}{2}\ln t+O(1) \quad \text{a.s.},\end{aligned}$$

where $H(p) := -p\ln p-(1-p)\ln(1-p)$, $p\in[0,1]$, is the entropy function; the last equality in (29) uses $\ln 2-H(p) = 2(p-1/2)^2+O(|p-1/2|^3)$ as $p\to 1/2$. Combining (29) with (28), we further obtain

$$\limsup_{t\to\infty}\frac{\ln X_t}{\sqrt{2\ln t\ln\ln t}} = 1 \text{ and } \liminf_{t\to\infty}\frac{\ln X_t}{\sqrt{2\ln t\ln\ln t}} = -1 \quad \text{a.s.} \tag{30}$$

B.2 Prediction interval

Now we show that (22) can be rewritten as (24). For brevity, we write k for k_n. Similarly to (29), we can rewrite (22) as

$$\ln 2-H(k/n)+\frac{1}{n}\ln\sqrt{2\pi\frac{k(n-k)}{n}}+\frac{\lambda_k+\lambda_{n-k}-\lambda_n}{n}-\frac{1}{n}\ln(n+1) < \frac{\ln\frac{1}{\delta}}{n}. \tag{31}$$

Since $\ln 2-H(p)\sim 2(p-1/2)^2$ $(p\to 1/2)$, we have $k/n = 1/2+o(1)$ for k satisfying (31), as $n\to\infty$. Combining this with (31), we further obtain

$$2\left(\frac{k}{n}-\frac{1}{2}\right)^2 < (1+\alpha_n)\frac{\ln\frac{1}{\delta}-\ln\sqrt{n}+\ln(n+1)+\beta_n}{n},$$

for some $\alpha_n = o(1)$ and $\beta_n = O(1)$, which can be rewritten as (24) for a different sequence $\alpha_n = o(1)$.

B.3 Calculations for Armitage's example

Finally, we deduce (26). Using a well-known expression ([2], 6.6.4) for the regularized beta function $I_p(a,b) := B(p;a,b)/B(a,b)$ and writing k for k_t, we obtain

$$X_t = \frac{B(k+1,t-k+1) - B(1/2;k+1,t-k+1)}{B(1/2;k+1,t-k+1)}$$
$$= \frac{1}{I_{1/2}(k+1,t-k+1)} - 1 = \frac{1}{\mathbf{P}\{B_{t+1} \geq k+1\}} - 1 = \frac{\mathbf{P}\{B_{t+1} \leq k\}}{\mathbf{P}\{B_{t+1} \geq k+1\}}. \tag{32}$$

As a final remark, let us compare the sizes of oscillation of the log likelihood ratio $\ln X_t$ that we have obtained in Section 8 and in this appendix for our examples of the three kinds of Bayesian hypothesis testing. When testing a simple null hypothesis against a simple alternative, $\ln X_t$ oscillated between approximately $\pm 0.75\sqrt{t \ln \ln t}$ (as noticed in Subsection 8.1). When testing a simple null hypothesis against a composite alternative, $\ln X_t$ oscillated between $\pm\sqrt{2 \ln t \ln \ln t}$ (see (30)). And finally, when testing a composite null hypothesis against a composite alternative, we can deduce from (32) that

$$\limsup_{t\to\infty} \frac{\ln X_t}{\ln \ln t} = 1 \text{ and } \liminf_{t\to\infty} \frac{\ln X_t}{\ln \ln t} = -1 \quad \text{a.s.}$$

(details omitted); therefore, $\ln X_t$ oscillates between $\pm \ln \ln t$. Roughly, the size of oscillations of $\ln X_t$ goes down from $\sqrt{t}$ to $\sqrt{\ln t}$ to $\ln \ln t$. Of course, these sizes are only examples, but they illustrate a general tendency.

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. Comments by three reviewers have led to numerous corrections and improvements, including addition of Section 8. We are grateful to Irina Shevtsova for advising us on latest developments related to the Berry-Esseen theorem. In our computer simulations we have used the R language [33] and the GNU C++ compiler. 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.

References

[1] Odd Aalen, Per Kragh Andersen, Ørnulf Borgan, Richard Gill, and Niels Keiding. History of applications of martingales in survival analysis. *Elec-*

tronic Journal for History of Probability and Statistics, 5(1), June 2009. Available at `www.jehps.net`.

[2] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. US Government Printing Office, Washington, DC, 1964. Republished many times by Dover, New York, starting from 1965.

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

[4] Francis J. Anscombe. Fixed-sample-size analysis of sequential observations. *Biometrics*, 10:89–100, 1954.

[5] Tom M. Apostol. An elementary view of Euler's summation formula. *American Mathematical Monthly*, 106:409–418, 1999.

[6] Peter Armitage. Discussion of "Consistency in statistical inference and decision", by C. A. B. Smith. *Journal of the Royal Statistical Society B*, 23:30–31, 1961.

[7] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 2000.

[8] 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`.

[9] 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`.

[10] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, Cambridge, UK, 2006.

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

[12] A. Philip Dawid, Steven de Rooij, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Insuring against loss of evidence in game-theoretic probability. *Statistics and Probability Letters*, 81:157–162, 2011.

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

[14] A. P. Dempster. *Elements of Continuous Multivariate Analysis*. Addison Wesley, Reading, MA, 1969.

[15] Ward Edwards, Harold Lindman, and Leonard J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242, 1963.

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

[17] C. Hipp and L. Mattner. On the normal approximation to symmetric binomial distributions. *Теория вероятностей и ее применения*, 52:610–617, 2007.

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

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

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

[21] 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`.

[22] Pierre Simon Laplace. Mémoire sur la probabilité des causes par les évènemens. *Savants étranges*, 6:621–656, 1774. English translation (1986): Memoir on the probability of the causes of events. *Statistical Science* **1** 364–378.

[23] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Springer, New York, revised first edition, 2006.

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

[25] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, third edition, 2008.

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

[27] 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 `www.probabilityandfinance.com/misc/erlangen.pdf`.

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

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

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

[31] 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.

[32] 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.

[33] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2010.

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

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

[36] Glenn Shafer. From Cournot's principle to market efficiency. The Game-Theoretic Probability and Finance project, Working Paper 15, `probabilityandfinance.com`, March 2006.

[37] Stephen M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–363, 1986.

[38] Isaac Todhunter. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Macmillan, London, 1865.

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

[40] Vladimir Vovk. The law of the iterated logarithm for random Kolmogorov, or chaotic, sequences. *Theory of Probability and Its Applications*, 32(3):413–425, 1987. Russian original: Закон повторного логарифма для случайных по Колмогорову, или хаотических, последовательностей. *Теория вероятностей и ее применения* **32** 456–468.

[41] 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.

[42] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[43] Eric-Jan Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14:779–804, 2007. On-line appendix: Stopping rules and their irrelevance for Bayesian inference.